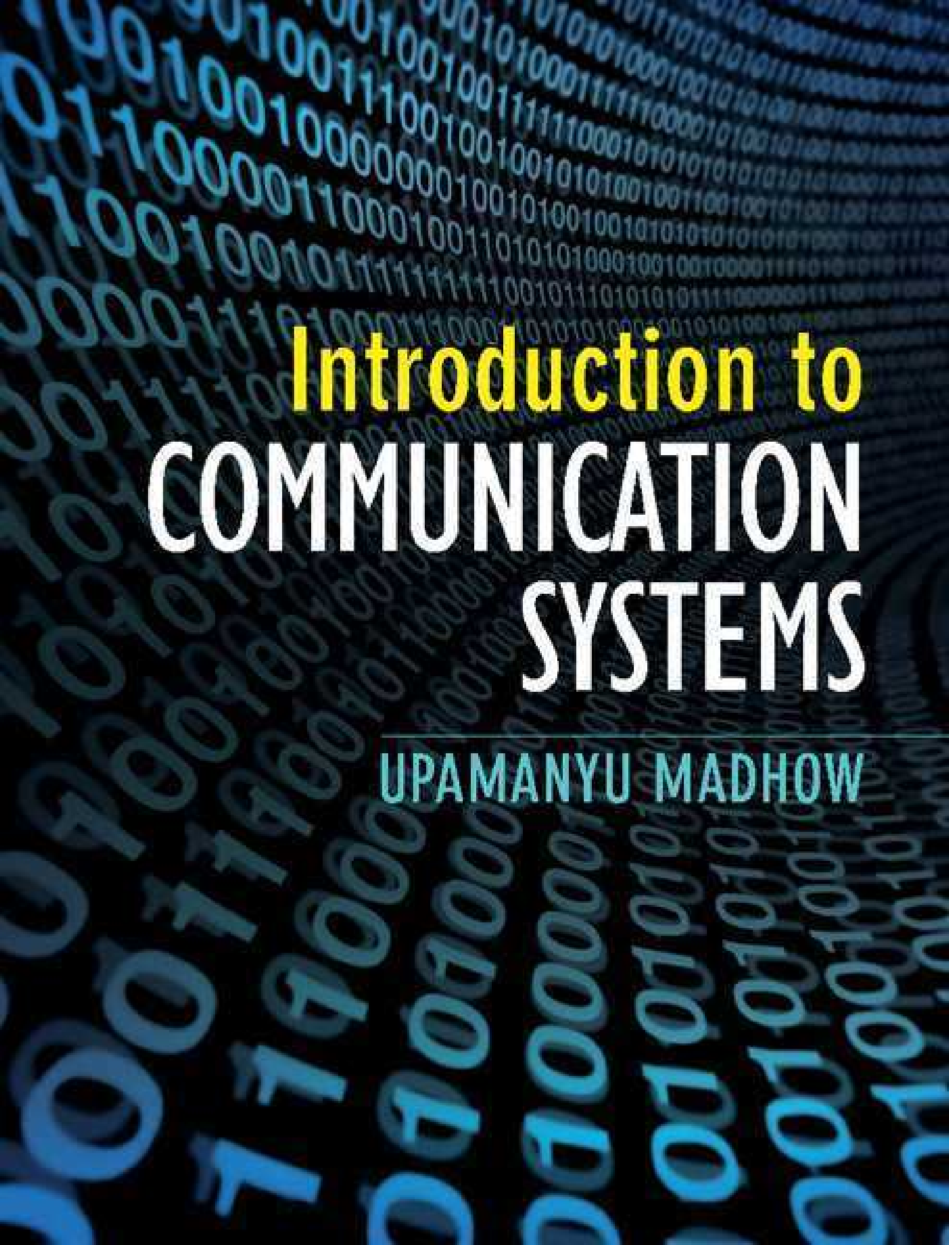

Introduction to COMMUNICATION SYSTEMS

UPAMANYU MADHOW

Introduction to Communication Systems

Showcasing the essential principles behind modern communication systems, this accessible undergraduate textbook provides a solid introduction to the foundations of communication theory.

- Carefully selected topics introduce students to the most important and fundamental concepts, giving them a focused, in-depth understanding of core material, and preparing them for more advanced study.
- Abstract concepts are introduced “just in time” and reinforced by nearly 200 end-of-chapter exercises, alongside numerous MATLAB code fragments, software problems, and practical lab exercises, firmly linking the underlying theory to real-world problems, and providing additional hands-on experience.
- An accessible lecture-style organisation makes it easy for students to navigate to key passages, and quickly identify the most relevant material.

Containing material suitable for a one- or two-semester course, and accompanied online by a password-protected solutions manual and supporting instructor resources, this is the perfect introductory textbook for undergraduate students studying electrical and computer engineering.

Upamanyu Madhow is Professor of Electrical and Computer Engineering at the University of California, Santa Barbara. His research interests broadly span communications, signal processing and networking, with current emphasis on next-generation wireless and bio-inspired approaches to networking and inference. He is a recipient of the NSF CAREER Award and the IEEE Marconi prize paper award in wireless communication, author of the graduate-level textbook *Fundamentals of Digital Communication* (2008), and a Fellow of the IEEE.

"Madhow does it again: *Introduction to Communication Systems* is an accessible yet rigorous new text that does for undergraduates what his digital communications book did for graduate students. It provides a superior treatment of not only the fundamentals of analog and digital communication, but also the theoretical underpinnings needed to understand them, including frequency domain analysis and probability. The book is unusual in that it also includes newer topics of pressing current relevance like multiple antenna communication and OFDM. I strongly recommend this book for faculty teaching senior level courses on communication systems."

Jeffrey G. Andrews
The University of Texas at Austin

"This is an excellent undergraduate text on analog and digital communications. It covers everything from classic analog techniques to recent wireless systems. Students will enjoy the inclusion of advanced topics such as channel coding and MIMO."

David Love
Purdue University

"*Introduction to Communication Systems* by Madhow is truly unique in the vast landscape of introductory books on communication systems. From the basics of signal processing, probability, and communications, to the advanced topics of coding, multipath mitigation, and multiple antenna systems, the book deftly interweaves abstract theory, design principles, and applications in a highly effective and insightful manner. This masterfully-written book will play a key role in teaching and inspiring the next generation of communication system engineers."

Andrea Goldsmith
Stanford University

"This is the textbook on communications I have wanted for a while. Crisply written, it forms the basis for an ideal two semester sequence. It nicely balances rigor, concepts and practice."

Saoura Dasgupta
University of Iowa

"This is a unique introduction to the basic principles of communication system design, with a remarkable combination of rigour and accessibility. The MATLAB exercises are expertly weaved together with theoretical principles making it an excellent textbook for training undergraduate communication systems engineers."

Suhas Diggavi
University of California, Los Angeles

"This is a valuable addition to the current set of textbooks on communication systems. It is comprehensive, and offers a modern perspective shaped by the author's research that has pushed the state of the art. The software labs enhance the practicality of the text and serve to illustrate more advanced material in an accessible way."

Michael Honig
Northwestern University

Introduction to Communication Systems

Upamanyu Madhow

University of California, Santa Barbara

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781107022775

First published 2014

Printed in the United States of America by Sheridan Books, Inc.

A catalog record for this publication is available from the British Library

Library of Congress Cataloging in Publication data

Madhow, Upamanyu.
Introduction to communication systems / Upamanyu Madhow, University of California, Santa Barbara.
pages cm
Includes bibliographical references.
ISBN 978-1-107-02277-5
1. Telecommunication. I. Title.
TK5101.M273 2014
621.382–dc23
2014004974

ISBN 978-1-107-02277-5 Hardback

Additional resources for this publication at www.cambridge.org/commnsystems

Cambridge University Press has no responsibility for the persistence or accuracy of URLs for external or third-party internet websites referred to in this publication, and does not guarantee that any content on such websites is, or will remain, accurate or appropriate.

To my family and students

Contents

Preface

Progress in telecommunications over the past two decades has been nothing short of revolutionary, with communications taken for granted in modern society to the same extent as electricity. There is therefore a persistent need for engineers who are well-versed in the principles of communication systems. These principles apply to communication between points in space, as well as communication between points in time (i.e., storage). Digital systems are fast replacing analog systems in both domains. This book has been written in response to the following core question: what is the basic material that an undergraduate student with an interest in communications should learn, in order to be well prepared for either industry or graduate school? For example, some institutions teach only digital communication, assuming that analog communication is dead or dying. Is that the right approach? From a purely pedagogical viewpoint, there are critical questions related to mathematical preparation: how much mathematics must a student learn to become well-versed in system design, what should be assumed as background, and at what point should the mathematics that is not in the background be introduced? Classically, students learn probability and random processes, and then tackle communication. This does not quite work today: students increasingly (and, I believe, rightly) question the applicability of the material they learn, and are less interested in abstraction for its own sake. On the other hand, I have found from my own teaching experience that students get truly excited about abstract concepts when they discover their power in applications, and it is possible to provide the means for such discovery using software packages such as MATLAB. Thus, we have the opportunity to get a new generation of students excited about this field: by covering abstractions “just in time” to shed light on engineering design, and by reinforcing concepts immediately using software experiments in addition to conventional pen-and-paper problem solving, we can remove the lag between learning and application, and ensure that the concepts stick.

This textbook represents my attempt to act upon the preceding observations, and is an outgrowth of my lectures for a two-course undergraduate elective sequence on communication at UCSB, which is often also taken by some beginning graduate students. Thus, it can be used as the basis for a two-course sequence in communication systems, or a single course on digital communication, at the undergraduate or beginning graduate level. The book also provides a review or introduction to communication systems for practitioners, easing the path to study of more advanced graduate texts and the research literature. The prerequisite is a course on signals and systems, together with an introductory course on probability. The required material on random processes is included in the text.

A student who masters the material here should be well-prepared for either graduate school or the telecommunications industry. The student should leave with an understanding of baseband and passband signals and channels, modulation formats appropriate for these channels, random processes and noise, a systematic framework for optimum demodulation based on signal-space concepts, performance analysis and power–bandwidth tradeoffs for common modulation schemes, a hint of the power of information theory and channel coding, and an introduction to communication techniques for dispersive channels and multiple antenna systems. Given the significant ongoing research and development activity in wireless communication, and the fact that an understanding of wireless link design provides a sound background for approaching other communication links, material enabling hands-on discovery of key concepts for wireless system design is distributed throughout the textbook.

I should add that I firmly believe that the utility of this material goes well beyond communications, important as that field is. Communications systems design merges concepts from signals and systems, probability and random processes, and statistical inference. Given the broad applicability of these concepts, a background in communications is of value in a large variety of areas requiring "systems thinking," as I discuss briefly at the end of Chapter 1.

The goal of the lecture-style exposition in this book is to clearly articulate a selection of concepts that I deem *fundamental* to communication system design, rather than to provide comprehensive coverage. "Just in time" coverage is provided by organizing and limiting the material so that we get to core concepts and applications as quickly as possible, and by sometimes asking the reader to operate with partial information (which is, of course, standard operating procedure in the real world of engineering design). However, the topics that we do cover are covered in sufficient detail to enable the student to solve nontrivial problems and to obtain hands-on involvement via software labs. Descriptive material that can easily be looked up online is omitted.

Organization

- Chapter 1 provides a perspective on communication systems, including a discussion of the transition from analog to digital communication and how it colors the selection of material in this text.
- Chapter 2 provides a review of signals and systems (biased towards communications applications), and then discusses the complex-baseband representation of passband signals and systems, emphasizing its critical role in modeling, design, and implementation. A software lab on modeling and undoing phase offsets in complex baseband, while providing a sneak preview of digital modulation, is included. Chapter 2 also includes a section on wireless-channel modeling in complex baseband using ray tracing, reinforced by a software lab that applies these ideas to simulate link time variations for a lamppost-based broadband wireless network.
- Chapter 3 covers analog communication techniques that remain relevant even as the world goes digital, including superheterodyne reception and phase-locked loops. Legacy analog modulation techniques are discussed to illustrate core concepts, as well as in

recognition of the fact that suboptimal analog techniques such as envelope detection and limiter–discriminator detection may have to be resurrected as we push the limits of digital communication in terms of speed and power consumption. Software labs reinforce and extend concepts in amplitude and angle modulation.

- Chapter 4 discusses digital modulation, including linear modulation using constellations such as pulse amplitude modulation (PAM), quadrature amplitude modulation (QAM), and phase-shift keying (PSK), and orthogonal modulation and its variants. The chapter includes discussion of the number of degrees of freedom available on a bandlimited channel, the Nyquist criterion for avoidance of intersymbol interference, and typical choices of Nyquist and square-root Nyquist signaling pulses. We also provide a sneak preview of power–bandwidth tradeoffs (with detailed discussion postponed until the effect of noise has been modeled in Chapters 5 and 6). A software lab providing a hands-on feel for Nyquist signaling is included in this chapter.

The material in Chapters 2 through 4 requires only a background in signals and systems.

- Chapter 5 provides a review of basic probability and random variables, and then introduces random processes. This chapter provides detailed discussion of Gaussian random variables, vectors and processes; this is essential for modeling noise in communication systems. Examples giving a preview of receiver operations in communication systems, and computation of performance measures such as error probability and signal-to-noise ratio (SNR), are provided. A discussion of the circular symmetry of white noise, and noise analysis of analog modulation techniques, are placed in an appendix, since this is material that is often skipped in modern courses on communication systems.
- Chapter 6 covers classical material on optimum demodulation for M-ary signaling in the presence of additive white Gaussian noise (AWGN). The background on Gaussian random variables, vectors, and processes developed in Chapter 5 is applied to derive optimal receivers, and to analyze their performance. After discussing error probability computation as a function of SNR, we are able to combine the materials in Chapters 4 and 6 for a detailed discussion of power–bandwidth tradeoffs. Chapter 6 concludes with an introduction to link-budget analysis, which provides guidelines on the choice of physical link parameters such as transmit and receive antenna gains, and distance between transmitter and receiver, using what we know about the dependence of error probability as a function of SNR. This chapter includes a software lab that builds on the Nyquist signaling lab in Chapter 4 by investigating the effect of noise. It also includes another software lab simulating performance over a time-varying wireless channel, examining the effects of fading and diversity, and introduces the concept of differential demodulation for avoidance of explicit channel tracking.

Chapters 2 through 6 provide a systematic lecture-style exposition of what I consider core concepts in communication at an undergraduate level.

- Chapter 7 provides a glimpse of information theory and coding whose goal is to stimulate the reader to explore further using more advanced resources such as graduate courses and textbooks. It shows the critical role of channel coding, provides an initial exposure to information-theoretic performance benchmarks, and discusses belief propagation in detail, reinforcing the basic concepts through a software lab.

- Chapter 8 provides a first exposure to the more advanced topics of communication over dispersive channels, and to multiple antenna systems, often termed space–time communication, or multiple-input, multiple-output (MIMO) communication. These topics are grouped together because they use similar signal processing tools. We emphasize lab-style "discovery" in this chapter using three software labs, one on adaptive linear equalization for single-carrier modulation, one on basic orthogonal frequency-division multiplexing (OFDM) transceiver operations, and one on MIMO signal processing for space–time coding and spatial multiplexing. The goal is for students to acquire hands-on insight that should motivate them to undertake a deeper and more systematic investigation.
- Finally, the epilogue contains speculation on future directions in communications research and technology. The goal is to provide a high-level perspective on where mastery of the introductory material in this textbook could lead, and to argue that the innovations which this field has already seen set the stage for many exciting developments to come.

The role of software. Software problems and labs are integrated into the text, with "code fragments" implementing core functionalities provided in the text. While code can be provided online, separate from the text (and, indeed, sample code is made available online for instructors), code fragments are integrated into the text for two reasons. First, they enable readers to immediately see the software realization of a key concept as they read the text. Second, I feel that students learn more by putting in the work of writing their own code, building on these code fragments if they wish, rather than using code that is easily available online. The particular software that we use is MATLAB, because of its widespread availability, and because of its importance in design and performance evaluation both in academia and in industry. However, the code fragments can also be viewed as "pseudocode," and can be easily implemented using other software packages or languages. Block-based packages such as Simulink (which builds upon MATLAB) are avoided here, because the use of software here is pedagogical rather than aimed at, say, designing a complete system by putting together subsystems as one might do in industry.

Suggestions for using this book

I view Chapter 2 (complex baseband), Chapter 4 (digital modulation), and Chapter 6 (optimum demodulation) as core material that *must* be studied to understand the concepts underlying modern communication systems. Chapter 6 relies on the probability and random processes material in Chapter 5, especially the material on jointly Gaussian random variables and white Gaussian noise (WGN), but the remaining material in Chapter 5 can be skipped or covered selectively, depending on the students' background. Chapter 3 (analog communication techniques) is designed such that it can be completely skipped if one wishes to focus solely on digital communication. Finally, Chapter 7 and Chapter 8 contain glimpses of advanced material that can be sampled according to the instructor's discretion. The qualitative discussion in the epilogue is meant to provide the student with perspective, and is not intended for formal coverage in the classroom.

In my own teaching at UCSB, this material forms the basis for a two-course sequence, with Chapters 2–4 covered in the first course, and Chapters 5 and 6 covered in the second course, with the dispersive channels portion of Chapter 8 providing the basis for the labs in the second course. The content of these courses is constantly being revised, and it is expected that the material on channel coding and MIMO may displace some of the existing material in the future. UCSB is on a quarter system, hence the coverage is fast-paced, and many topics are omitted or skimmed. There is ample material here for a two-semester undergraduate course sequence. For a one-semester course, one possible organization is to cover Chapter 2 (focusing on the complex envelope), Chapter 4, a selection of Chapter 5, Chapter 6, and, if time permits, Chapter 7.

The slides accompanying the book are intended not to provide comprehensive coverage of the material, but rather to provide an example of selections from the material to be covered in the classroom. I must comment in particular on Chapter 5. While much of the book follows the format in which I lecture, Chapter 5 is structured as a reference on probability, random variables, and random processes that the instructor must pick and choose from, depending on the background of the students in the class. The particular choices I make in my own lectures on this material are reflected in the slides for this chapter.

Acknowledgements

This book is an outgrowth of lecture notes for an undergraduate elective course sequence in communications at UCSB, and I am grateful to the succession of students who have used, and provided encouraging comments on, the evolution of the course sequence and the notes. I would also like to acknowledge faculty in the communications area at UCSB who were kind enough to give me a “lock” on these courses over the past few years, as I was developing this textbook.

The first priority in a research university is to run a vibrant research program, hence I must acknowledge the extraordinarily capable graduate students in my research group over the years during which this textbook was developed. They have done superb research with minimal supervision from me, and the strength of their peer interactions and collaborations is what gave me the mental space, and time, needed to write this textbook. Current and former group members who have directly helped with aspects of this book include Andrew Irish, Babak Mamandipoor, Dinesh Ramasamy, Maryam Eslami Rasekh, Sumit Singh, Sriram Venkateswaran, and Aseem Wadhwa.

I gratefully acknowledge the funding agencies which have provided support for our research group in recent years, including the National Science Foundation (NSF), the Army Research Office (ARO), the Defense Advanced Research Projects Agency (DARPA), and Systems on Nanoscale Information Fabrics (SONIC), a center supported by DARPA and Microelectronics Advanced Research Corporation (MARCO). One of the primary advantages of a research university is that undergraduate education is influenced, and kept up to date, by cutting-edge research. This textbook embodies this paradigm both in its approach (an emphasis on what one can *do* with what one learns) and in its content (emphasis of concepts that are fundamental background for research in the area).

I thank Phil Meyler and his colleagues at Cambridge University Press for encouraging me to initiate this project, and for their blend of patience and persistence in getting me to see it through despite a host of other commitments. I also thank the anonymous reviewers of the book proposal and sample chapters sent to Cambridge several years back for their encouragement and constructive comments. I am also grateful to a number of faculty colleagues who have given encouragement, helpful suggestions, and pointers to alternative pedagogical approaches: Professor Soura Dasgupta (University of Iowa), Professor Jerry Gibson (UCSB), Professor Gerhard Kramer (Technische Universität München, Munich), Professor Phil Schniter (Ohio State University), and Professor Venu Veeravalli (University of Illinois at Urbana-Champaign).

Finally, I thank the anonymous reviewers of the almost-final manuscript for their helpful comments.

Finally, I would like to thank my wife and children for always being the most enjoyable and interesting people to spend time with. Recharging my batteries in their company, and that of our many pets, is what provides me with the energy needed for an active professional life.

1 Introduction

This textbook provides an introduction to the conceptual underpinnings of communication technologies. Most of us directly experience such technologies daily: browsing (and audio/video streaming from) the Internet, sending/receiving emails, watching television, or carrying out a phone conversation. Many of these experiences occur on mobile devices that we carry around with us, so that we are always connected to the cyberworld of modern communication systems. In addition, there is a huge amount of machine-to-machine communication that we do not directly experience, but which is indispensable for the operation of modern society. This includes, for example, signaling between routers on the Internet, or between processors and memories on any computing device.

We define *communication* as the process of *information transfer across space or time.* Communication across space is something we have an intuitive understanding of: for example, radio waves carry our phone conversation between our cell phone and the nearest base station, and coaxial cables (or optical fiber, or radio waves from a satellite) deliver television from a remote location to our home. However, a moment's thought shows that that communication across time, or storage of information, is also an everyday experience, given our use of storage media such as compact discs (CDs), digital video discs (DVDs), hard drives, and memory sticks. In all of these instances, the key steps in the operation of a communication link are as follows:

(a) insertion of information into a signal, termed the *transmitted signal,* compatible with the physical medium of interest;
(b) propagation of the signal through the physical medium (termed the *channel*) in space or time; and
(c) extraction of information from the signal (termed the *received signal*) obtained after propagation through the medium.

In this book, we study the fundamentals of modeling and design for these steps.

Chapter plan

In Section 1.1, we provide a high-level description of analog and digital communication systems, and discuss why digital communication is the inevitable design choice in modern systems. In Section 1.2, we briefly provide a technological perspective on recent developments in communication. We do not attempt to provide a comprehensive discussion of the fascinating history of communication: thanks to the advances in communication that brought us the Internet, it is easy to look it up online! A discussion of the scope of this textbook is provided in Section 1.3.

1.1 Analog or digital?

Even without defining information formally, we intuitively understand that speech, audio, and video signals contain information. We use the term *message signals* for such signals, since these are the messages we wish to convey over a communication system. In their original form – both during generation and consumption – these message signals are *analog*: they are continuous-time signals, with the signal values also lying in a continuum. When someone plays the violin, an analog acoustic signal is generated (often translated to an analog electrical signal using a microphone). Even when this music is recorded onto a digital storage medium such as a CD (using the digital communication framework outlined in Section 1.1.2), when we ultimately listen to the CD being played on an audio system, we hear an analog acoustic signal. The transmitted signals corresponding to physical communication media are also analog. For example, in both wireless and optical communication, we employ electromagnetic waves, which correspond to continuous-time electric and magnetic fields taking values in a continuum.

1.1.1 Analog communication

Given the analog nature of both the message signal and the communication medium, a natural design choice is to map the analog message signal (e.g., an audio signal, translated from the acoustic to the electrical domain using a microphone) to an analog transmitted signal (e.g., a radio wave carrying the audio signal) that is compatible with the physical medium over which we wish to communicate (e.g., broadcasting audio over the air from an FM radio station). This approach to communication system design, depicted in Figure 1.1, is termed *analog communication.* Early communication systems were all analog: examples include AM (amplitude modulation) and FM (frequency modulation) radio, analog television, first-generation cellular-phone technology (based on FM), vinyl records, audio cassettes, and VHS or Betamax videocassettes

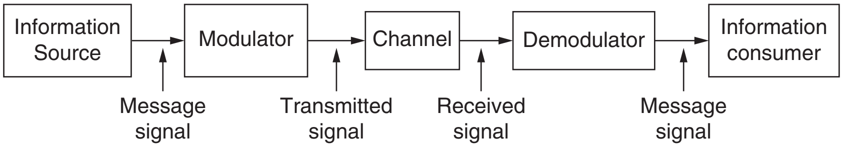

Figure 1.1 A block diagram for an analog communication system. The modulator transforms the message signal into the transmitted signal. The channel distorts and adds noise to the transmitted signal. The demodulator extracts an estimate of the message signal from the received signal arriving from the channel.

While analog communication might seem like the most natural option, it is in fact obsolete. Cellular-phone technologies from the second generation onwards are digital; vinyl records and audio cassettes have been supplanted by CDs, and videocassettes by DVDs. Broadcast technologies such as radio and television are often slower to upgrade because of economic and political factors, but digital broadcast radio and television technologies are either replacing or sidestepping (e.g., via satellite) analog FM/AM radio and television broadcast. Let us now define what we mean by digital communication, before discussing the reasons for the inexorable trend away from analog and towards digital communication.

1.1.2 Digital communication

The conceptual basis for digital communication was established in 1948 by Claude Shannon, when he founded the field of information theory. There are two main threads to this theory.

- **Source coding and compression.** Any information-bearing signal can be represented efficiently, to within a desired accuracy of reproduction, by a digital signal (i.e., a discrete-time signal taking values from a discrete set), which in its simplest form is just a sequence of binary digits (zeros or ones), or *bits.* This is true irrespective of whether the information source is text, speech, audio, or video. Techniques for performing the mapping from the original source signal to a bit sequence are generically termed *source coding.* They often involve *compression,* or removal of redundancy, in a manner that exploits the properties of the source signal (e.g., the heavy spatial correlation among adjacent pixels in an image can be exploited to represent it more efficiently than a pixel-by-pixel representation).
- **Digital information transfer.** Once the source encoding has been done, our communication task reduces to reliably transferring the bit sequence at the output of the source encoder across space or time, without worrying about the original source and the sophisticated tricks that have been used to encode it. The performance of any communication system depends on the relative strengths of the signal and noise or interference, and the distortions imposed by the channel. Shannon showed that, once we have fixed these operational parameters for any communication channel, there exists a maximum possible rate of reliable communication, termed the *channel capacity.* Thus, given the information bits at the output of the source encoder, in principle, we can transmit them reliably over a given link as long as the information rate is smaller than the channel capacity, and we cannot transmit them reliably if the information rate is larger than the channel capacity. This sharp transition between reliable and unreliable communication differs fundamentally from analog communication, where the quality of the reproduced source signal typically degrades gradually as the channel conditions get worse.

A block diagram for a typical digital communication system based on these two threads is shown in Figure 1.2. We now briefly describe the role of each component, together with simplified examples of its function.

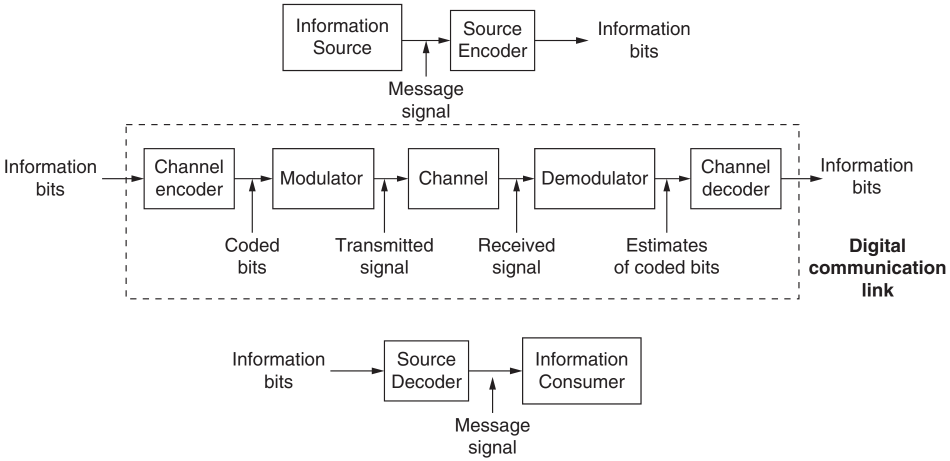

Figure 1.2 Components of a digital communication system.

Source encoder As already discussed, the source encoder converts the message signal into a sequence of information bits. The information bit rate depends on the nature of the message signal (e.g., speech, audio, video) and the application requirements. Even when we fix the class of message signals, the choice of source encoder is heavily dependent on the setting. For example, video signals are heavily compressed when they are sent over a cellular link to a mobile device, but are lightly compressed when sent to a high-definition television (HDTV) set. A cellular link can support a much smaller bit rate than, say, the cable connecting a DVD player to an HDTV set, and a smaller mobile display device requires lower resolution than a large HDTV screen. In general, the source encoder must be chosen such that the bit rate it generates can be supported by the digital communication link we wish to transfer information over. Other than this, source coding can be decoupled entirely from link design (we comment further on this a bit later).

Example. A laptop display may have resolution 1024 × 768 pixels. For a grayscale digital image, the intensity for each pixel might be represented by 8 bits. Multiplying by the number of pixels gives us about 6.3 million bits, or about 0.8 Mbyte (a byte equals 8 bits). However, for a typical image, the intensities for neighboring pixels are heavily correlated, which can be exploited for significantly reducing the number of bits required to represent the image, without noticeably distorting it. For example, one could take a two-dimensional Fourier transform, which concentrates most of the information in the image at lower frequencies, and then discard many of the high-frequency coefficients. There are other possible transforms one could use, and also several more processing stages, but the bottom line is that, for natural images, state-of-the-art image-compression algorithms can provide 10× compression (i.e., reduction in the number of bits relative to the original uncompressed digital image) with hardly any perceptual degradation. Far more aggressive compression ratios are possible if we are willing to tolerate more distortion. For video, in addition to the spatial correlation exploited for image compression, we can also exploit temporal correlation across successive frames.

Channel encoder The channel encoder adds redundancy to the information bits obtained from the source encoder, in order to facilitate error recovery after transmission over the channel. It might appear that we are putting in too much work, adding redundancy just after the source encoder has removed it. However, the redundancy added by the channel encoder is tailored to the channel over which information transfer is to occur, whereas the redundancy in the original message signal is beyond our control, so that it would be inefficient to keep it when we transmit the signal over the channel.

Example. The noise and distortion introduced by the channel can cause errors in the bits we send over it. Consider the following abstraction for a channel: we can send a string of bits (zeros or ones) over it, and the channel randomly flips each bit with probability 0.01 (i.e., the channel has a 1% error rate). If we cannot tolerate this error rate, we could repeat each bit that we wish to send three times, and use a majority rule to decide on its value. Now, we only make an error if two or more of the three bits are flipped by the channel. It is left as an exercise to calculate that an error now happens with probability approximately 0.0003 (i.e., the error rate has gone down to 0.03%). That is, we have improved performance by introducing redundancy. Of course, there are far more sophisticated and efficient techniques for introducing redundancy than the simple repetition strategy just described; see Chapter 7.

Modulator The modulator maps the coded bits at the output of the channel encoder to a transmitted signal to be sent over the channel. For example, we may insist that the transmitted signal fit within a given frequency band and adhere to stringent power constraints in a wireless system, where interference between users and between co-existing systems is a major concern. Unlicensed WiFi transmissions typically occupy 20–40 MHz of bandwidth in the 2.4- or 5-GHz bands. Transmissions in fourth-generation cellular systems may often occupy bandwidths ranging from 1 to 20 MHz at frequencies ranging from 700 MHz to 3 GHz. While these signal bandwidths are being increased in an effort to increase data rates (e.g., up to 160 GHz for emerging WiFi standards, and up to 100 MHz for emerging cellular standards), and new frequency bands are being actively explored (see the epilogue for more discussion), the transmitted signal still needs to be shaped to fit within certain spectral constraints.

Example. Suppose that we send bit value 0 by transmitting the signal $s(t)$, and bit value 1 by transmitting $-s(t)$. Even for this simple example, we must design the signal $s(t)$ so it fits within spectral constraints (e.g., two different users may use two different segments of spectrum to avoid interfering with each other), and we must figure out how to prevent successive bits of the same user from interfering with each other. For wireless communication, these signals are voltages generated by circuits coupled to antennas, and are ultimately emitted as electromagnetic waves from the antennas.

The channel encoder and modulator are typically jointly designed, keeping in mind the anticipated channel conditions, and the result is termed a *coded modulator.*

Channel The channel distorts and adds noise, and possibly interference, to the transmitted signal. Much of our success in developing communication technologies has resulted from being able to optimize communication strategies based on accurate mathematical models for the channel. Such models are typically statistical, and are developed with significant effort using a combination of measurement and computation. The physical characteristics of the communication medium vary widely, and hence so do the channel models. Wireline channels are typically well modeled as linear and time-invariant, while optical-fiber channels exhibit nonlinearities. Wireless mobile channels are particularly challenging because of the time variations caused by mobility, and due to the potential for interference due to the broadcast nature of the medium. The link design also depends on system-level characteristics, such as whether or not the transmitter has feedback regarding the channel, and what strategy is used to manage interference.

Example. Consider communication between a cellular base station and a mobile device. The electromagnetic waves emitted by the base station can reach the mobile's antennas through multiple paths, including bounces off streets and building surfaces. The received signal at the mobile can be modeled as multiple copies of the transmitted signal with different gains and delays. These gains and delays change due to mobility, but the rate of change is often slow compared with the data rate, hence, over short intervals, we can get away with modeling the channel as a linear time-invariant system that the transmitted signal goes through before arriving at the receiver.

Demodulator The demodulator processes the signal received from the channel to produce bit estimates to be fed to the channel decoder. It typically performs a number of signal-processing tasks, such as synchronization of phase, frequency, and timing, and compensating for distortions induced by the channel.

Example. Consider the simplest possible channel model, where the channel just adds noise to the transmitted signal. In our earlier example of sending $\pm s(t)$ to send 0 or 1, the demodulator must guess, based on the noisy received signal, which of these two options is true. It might make a hard decision (e.g., guess that 0 was sent), or hedge its bets, and make a soft decision, saying, for example, that it is 80% sure that the transmitted bit is a zero. There are many other aspects of demodulation that we have swept under the rug: for example, before making any decisions, the demodulator has to perform functions such as synchronization (making sure that the receiver's notion of time and frequency is consistent with the transmitter's) and equalization (compensating for the distortions due to the channel).

Channel decoder The channel decoder processes the imperfect bit estimates provided by the demodulator, and exploits the controlled redundancy introduced by the channel encoder to estimate the information bits.

Example. The channel decoder takes the guesses from the demodulator and uses the redundancies in the channel code to clean up the decisions. In our simple example of repeating every bit three times, it might use a majority rule to make its final decision if the demodulator is putting out hard decisions. For soft decisions, it might use more sophisticated combining rules with improved performance.

While we have described the demodulator and decoder as operating separately and in sequence for simplicity, there can be significant benefits from iterative information exchange between the two. In addition, for certain coded modulation strategies in which channel coding and modulation are tightly coupled, the demodulator and channel decoder may be integrated into a single entity.

Source decoder The source decoder processes the estimated information bits at the output of the channel decoder to obtain an estimate of the message. The message format may, but need not, be the same as that of the original message input to the source encoder: for example, the source encoder may translate speech to text before encoding into bits, and the source decoder may output a text message to the end user.

Example. For the example of a digital image considered earlier, the compressed image can be translated back to a pixel-by-pixel representation by taking the inverse spatial Fourier transform of the coefficients that survived the compression.

We are now ready to compare analog and digital communication, and discuss why the trend towards digital is inevitable.

1.1.3 Why digital?

On comparing the block diagrams for analog and digital communication in Figures 1.1 and 1.2, respectively, we see that the digital communication system involves far more processing. However, this is not an obstacle for modern transceiver design, due to the exponential increase in the computational power of low-cost silicon integrated circuits. Digital communication has the following key advantages.

Optimality For a point-to-point link, it is optimal to separately optimize source coding and channel coding, as long as we do not mind the delay and processing incurred in doing so. Owing to this *source–channel separation principle,* we can leverage the best available source codes and the best available channel codes in designing a digital communication system, independently of each other. Efficient source encoders must be highly specialized. For example, state-of-the-art speech encoders, video-compression algorithms, and text-compression algorithms are very different from each other, and are each the result of significant effort over many years by a large community of researchers. However, once source encoding has been performed, the coded modulation scheme used over the communication link can be engineered to transmit the information bits reliably, regardless of what kind of source they correspond to, with the bit rate limited only by the channel and transceiver characteristics. Thus, the design of a digital communication link is *source-independent* and *channel-optimized.* In contrast, the waveform transmitted in an analog communication system depends on the message signal, which is beyond the control of the link designer, hence we do not have the freedom to optimize link performance over all possible communication schemes. This is not just a theoretical observation: in practice, huge performance gains are obtained from switching from analog to digital communication.

Scalability While Figure 1.2 shows a single digital communication link between source encoder and decoder, under the source–channel-separation principle, there is nothing preventing us from inserting additional links, putting the source encoder and decoder at the end points. This is because digital communication allows *ideal regeneration* of the information bits, hence every time we add a link, we can focus on communicating reliably over that particular link. (Of course, information bits do not always get through reliably, hence we typically add error-recovery mechanisms such as retransmission, at the level of an individual link or "end-to-end" over a sequence of links between the information source and sink.) Another consequence of the source–channel-separation principle is that, since information bits are transported without interpretation, the same link can be used to carry multiple kinds of messages. A particularly useful approach is to chop the information bits up into discrete chunks, or *packets*, which can then be processed independently on each link. These properties of digital communication are critical for enabling massively scalable, general-purpose, *communication networks* such as the Internet. Such networks can have large numbers of digital communication links, possibly with different characteristics, independently engineered to provide "bit pipes" that can support data rates. Messages of various kinds, after source encoding, are reduced to packets, and these packets are switched along different paths along the network, depending on the identities of the source and destination nodes, and the loads on different links in the network. None of this would be possible with analog communication: link performance in an analog communication system depends on message properties, and successive links incur noise accumulation, which limits the number of links which can be cascaded.

The preceding makes it clear that source–channel separation, and the associated bit-pipe abstraction, is crucial in the formation and growth of modern communication networks. However, there are some important caveats that are worth noting. Joint source–channel design can provide better performance in some settings, especially when there are constraints on delay or complexity, or if multiple users are being supported simultaneously on a given communication medium. In practice, this means that "local" violations of the separation principle (e.g., over a wireless last hop in a communication network) may be a useful design trick. Similarly, the bit-pipe abstraction used by network designers is too simplistic for the design of wireless networks at the edge of the Internet: physical properties of the wireless channel such as interference, multipath propagation, and mobility must be taken into account in network engineering.

1.1.4 Why analog design remains important

While we are interested in transporting bits in digital communication, the physical link over which these bits are sent is analog. Thus, analog and mixed-signal (digital/analog) design continue to play a crucial role in modern digital communication systems. Analog design of digital-to-analog converters, mixers, amplifiers, and antennas is required in order to translate bits to physical waveforms to be emitted by the transmitter. At the receiver, analog design of antennas, amplifiers, mixers, and analog-to-digital converters is required in order to translate the physical received waveforms to digital (discrete-valued, discrete-time) signals that are amenable to the digital signal processing that is at the core of modern transceivers. Analog circuit design for communications is therefore a thriving field in its own right, which this textbook makes no attempt to cover. However, the material in Chapter 3 on analog communication techniques is intended to introduce digital communication system designers to some of the high-level issues addressed by analog circuit designers.

The goal is to establish enough of a common language to facilitate interaction between system and circuit designers. While much of digital communication system design can be carried out by abstracting out the intervening analog design (as done in Chapters 4 through 8), closer interaction between system and circuit designers becomes increasingly important as we push the limits of communication systems, as briefly indicated in the epilogue.

1.2 A technology perspective

We now discuss some technology trends and concepts that have driven the astonishing growth in communication systems in the past two decades, and that are expected to impact future developments in this area. Our discussion is structured in terms of big technology "stories."

Technology story 1: the Internet Some of the key ingredients that contributed to its growth and the essential role it plays in our lives are as follows.

- Any kind of message can be chopped up into packets and routed across the network, using an Internet Protocol (IP) that is simple to implement in software.
- Advances in optical-fiber communication and high-speed digital hardware enable a super-fast "core" of routers connected by very high-speed, long-range links, that enable worldwide coverage;
- The World Wide Web, or web, makes it easy to organize information into interlinked hypertext documents, which can be browsed from anywhere in the world.
- The digitization of content (audio, video, books) means that ultimately "all" information is expected to be available on the web.
- Search engines enable us to efficiently search for this information.
- Connectivity applications such as email, teleconferencing, videoconferencing and online social networks have become indispensable in our daily lives.

Technology story 2: wireless Cellular mobile networks are everywhere, and are based on the breakthrough concept that ubiquitous tetherless connectivity can be provided by breaking the world into cells, with "spatial reuse" of precious spectrum resources in cells that are "far enough" apart. Base stations serve mobiles in their cells, and hand them off to adjacent base stations when the mobile moves to another cell. While cellular networks were invented to support voice calls for mobile users, today's mobile devices (e.g., "smart phones" and tablet computers) are actually powerful computers with displays large enough for users to consume video on the go. Thus, cellular networks must now support seamless access to the Internet. The billions of mobile devices in use easily outnumber desktop and laptop computers, so that the most important parts of the Internet today are arguably the cellular networks at its edge. Mobile service providers are having great difficulty keeping up with the increase in demand resulting from this convergence of cellular and Internet; by some estimates, the capacity of cellular networks must be scaled up by several orders of magnitude, at least in densely populated urban areas! As discussed in the epilogue, a major challenge for the communication researcher and technologist, therefore, is to come up with the breakthroughs required to deliver such capacity gains.

Another major success in wireless is WiFi, a catchy term for a class of standardized wireless local-area network (WLAN) technologies based on the IEEE 802.11 family of standards. Currently, WiFi networks use unlicensed spectrum in the 2.4- and 5-GHz bands, and have come into widespread use in both residential and commercial environments. WiFi transceivers are now incorporated into almost every computer and mobile device. One way of alleviating the cellular capacity crunch that was just mentioned is to offload Internet access to the nearest WiFi network. Of course, since different WiFi networks are often controlled by different entities, seamless switching between cellular and WiFi is not always possible.

It is instructive to devote some thought to the contrast between cellular and WiFi technologies. Cellular transceivers and networks are far more tightly engineered. They employ spectrum that mobile operators pay a great deal of money to license, hence it is critical to use this spectrum efficiently. Furthermore, cellular networks must provide robust wide-area coverage in the face of rapid mobility (e.g., automobiles at highway speeds). In contrast, WiFi uses unlicensed (i.e., free!) spectrum, must provide only local coverage, and typically handles much slower mobility (e.g., pedestrian motion through a home or building). As a result, WiFi can be more loosely engineered than cellular. It is interesting to note that, despite the deployment of many uncoordinated WiFi networks in an unlicensed setting, WiFi typically provides acceptable performance, partly because the relatively large amount of unlicensed spectrum (especially in the 5-GHz band) allows channel switching on encountering excessive interference, and partly because of naturally occurring spatial reuse (WiFi networks that are “far enough” from each other do not interfere with each other). Of course, in densely populated urban environments with many independently deployed WiFi networks, the performance can deteriorate significantly, a phenomenon sometimes referred to as a tragedy of the commons (individually selfish behavior leading to poor utilization of a shared resource). As we briefly discuss in the epilogue, both the cellular and the WiFi design paradigms need to evolve to meet our future needs.

Technology story 3: Moore’s law Moore’s “law” is actually an empirical observation attributed to Gordon Moore, one of the founders of Intel Corporation. It can be paraphrased as saying that the density of transistors in an integrated circuit, and hence the amount of computation per unit cost, can be expected to increase exponentially over time. This observation has become a self-fulfilling prophecy, because it has been taken up by the semiconductor industry as a growth benchmark driving their technology roadmap. While the growth in density implied by Moore’s law may be slowing down somewhat, it has already had a spectacular impact on the communications industry by drastically lowering the cost and increasing the speed of digital computation. By converting analog signals to the digital domain as soon as possible, advanced transceiver algorithms can be implemented in digital signal processing (DSP) using low-cost integrated circuits, so that research breakthroughs in coding and modulation can be quickly transitioned into products. This leads to economies of scale that have been critical to the growth of mass-market products in both wireless (e.g., cellular and WiFi) and wireline (e.g., cable modems and DSL) communication.

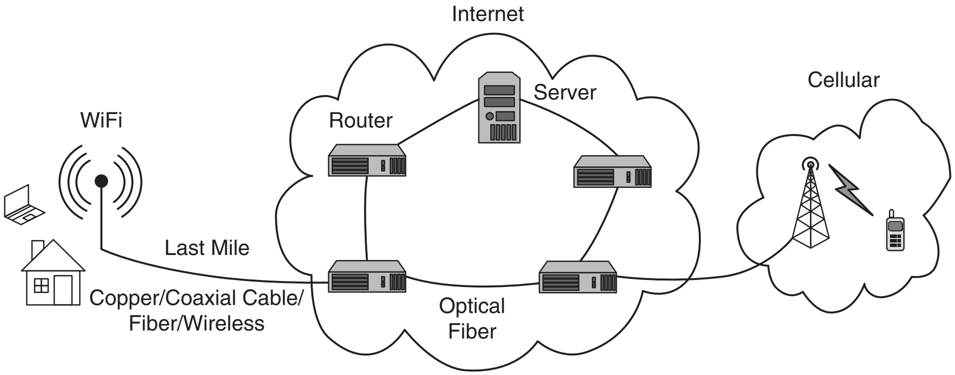

Figure 1.3 The Internet has a core of routers and servers connected by high-speed fiber links, with wireless networks hanging off the edge (figure courtesy of Aseem Wadhwa).

How do these stories come together? The sketch in Figure 1.3 highlights key building blocks of the Internet today. The *core* of the network consists of powerful routers that direct packets of data from an incoming edge to an outgoing edge, and *servers* (often housed in large *data centers*) that serve up content requested by *clients* such as personal computers and mobile devices. The elements in the core network are connected by high-speed optical fiber. Wireless can be viewed as hanging off the edge of the Internet. Wide-area cellular networks may have worldwide coverage, but each base station is typically connected by a high-speed link to the wired Internet. WiFi networks are wireless local-area networks, typically deployed indoors (but potentially also providing outdoor coverage for low-mobility scenarios) in homes and office buildings, connected to the Internet via *last-mile* links, which might run over copper wires (a legacy of wired telephony, with transceivers typically upgraded to support broadband Internet access) or coaxial cable (originally deployed to deliver cable television, but now also providing broadband Internet access). Some areas have been upgraded to optical fiber to the curb or even to the home, while some others might be remote enough to require wireless last-mile solutions.

Zooming in now on cellular networks, Figure 1.4 shows three adjacent cells in a cellular network with hexagonal cells. A working definition of a cell is that it is the area around a base station where the signal strength is higher than that from other base stations. Of course, under realistic propagation conditions, cells are never hexagonal, but the concept of spatial reuse still holds: the interference between distant cells can be neglected, hence they can use the same communication resources. For example, in Figure 1.4, we might decide to use three different frequency bands in the three cells shown, but might then reuse these bands in other cells. Figure 1.4 also shows that a user may be simultaneously in range of multiple base stations when near cell boundaries. Crossing these boundaries may result in a *handoff* from one base station to another. In addition, near cell boundaries, a mobile device may be in communication with multiple base stations simultaneously, a concept known as *soft handoff*.

It is useful for a communication system designer to be aware of the preceding "big picture" of technology trends and network architectures in order to understand how to

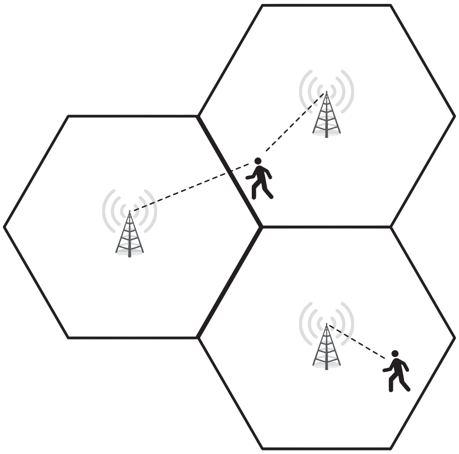

Figure 1.4 A segment of a cellular network with idealized hexagonal shapes (figure courtesy of Aseem Wadhwa).

direct his or her talents as these systems continue to evolve (the epilogue contains more detailed speculation regarding this evolution). However, the first order of business is to acquire the *fundamentals* required to get going in this field. These are quite simply stated: a communication system designer must be comfortable with mathematical modeling (in order to understand the state of the art, as well as to devise new models as required), and with devising and evaluating signal-processing algorithms based on these models. The goal of this textbook is to provide a first exposure to such a technical background.

1.3 The scope of this textbook

Referring to the block diagram of a digital communication system in Figure 1.2, our focus in this textbook is to provide an introduction to the design of a digital communication link as shown inside the dashed box. While we are primarily interested in digital communication, circuit designers implementing such systems must deal with analog waveforms, hence we believe that a rudimentary background in analog communication techniques, as provided in this book, is useful for the communication system designer. We do not discuss source encoding and decoding in this book; these topics are highly specialized and technical, and doing them justice requires an entire textbook of its own at the graduate level. A detailed outline of the book is provided in the preface, hence we restrict ourselves here to summarizing the roles of the various chapters:

Chapter 2 introduces the signal-processing background required for DSP-centric implementations of communication transceivers;

Chapter 3 provides just enough background on analog communication techniques (this can be skipped to focus exclusively on digital communication);

Chapter 4 discusses digital modulation techniques;

Chapter 5 provides the probability background required for receiver design, including noise modeling;

Chapter 6 discusses design and performance analysis of demodulators in digital communication systems for idealized link models;

Chapter 7 provides an initial exposure to channel coding techniques and benchmarks;

Chapter 8 provides an introduction to approaches for handling channel dispersion, and to multiple antenna communication; and the

Epilogue discusses emerging trends shaping research and development in communications.

Chapters 2, 4, and 6 are core material that must be mastered (much of Chapter 5 is also core material, but some readers may already have enough probability background that they can skip, or skim, it). Chapter 3 is highly recommended for communication system designers with an interest in radio-frequency circuit design, since it highlights, at a high level, some of the ideas and issues that come up there. Chapters 7 and 8 are independent of each other, and contain more advanced material that might not always fit within an undergraduate curriculum. They contain "hands-on" introductions to these topics via code fragments and software labs that should encourage the reader to explore further.

1.4 Why study communication systems?

Before launching into our formal study, it makes sense to ask why the material in this textbook is worth studying. There are several obvious answers to this question. The indispensable role of communications in modern life, and the success of the communications industry, implies that a solid understanding of this material constitutes a valuable skill set. The vibrant future of communications (see the epilogue) ensures the continuing value of this skill set for many decades to come. However, there is also an indirect, and perhaps more fundamental, answer to this question. The design of communication systems today represents a triumph of mathematical modeling and statistical signal processing. Detailed, hands-on experience building confidence in such techniques is therefore excellent preparation for tackling more complex systems for which complete mathematical models might not be available, as the author has discovered in his own research. Examples of such systems include the Internet itself, online social networks running on the Internet, financial systems exhibiting a complex web of interdependences, as well as signal processing, inference, and machine-learning techniques for the huge volumes of data ("big data") being generated in a host of other applications.

1.5 Concept summary

The goal of this chapter is to provide an intellectual framework and motivation for the rest of this textbook. Some of the key concepts are as follows.

- Communication refers to information transfer across either space or time, where the latter refers to storage media.

- Signals carrying information and signals that can be sent over a communication medium are both inherently analog (i.e., continuous-time, continuous-valued).
- Analog communication corresponds to transforming an analog message waveform directly into an analog transmitted waveform at the transmitter, and undoing this transformation at the receiver.
- Digital communication corresponds to first reducing message waveforms to information bits, and then transporting these bits over the communication channel.
- Digital communication requires the following steps: source encoding and decoding, modulation and demodulation, channel encoding and decoding.
- While digital communication requires more processing steps than analog communication, it has the advantages of optimality and scalability, hence there is an unstoppable trend from analog to digital.
- The growth in communication has been driven by major technology stories including the Internet, wireless, and Moore's law.
- Key components of the communication system designer's toolbox are mathematical modeling and signal processing.

1.6 Notes

There are many textbooks on communication systems at both undergraduate and graduate level. Undergraduate texts include Haykin [1], Proakis and Salehi [2], Pursley [3], and Ziemer and Tranter [4]. Graduate texts, which typically focus on digital communication, include Barry, Lee, and Messerschmitt [5], Benedetto and Biglieri [6], Madhow [7], and Proakis and Salehi [8]. The first coherent exposition of the modern theory of communication receiver design is in the classical (graduate level) textbook by Wozencraft and Jacobs [9]. Other important classical graduate-level texts are Viterbi and Omura [10] and Blahut [11]. More specialized references (e.g., on signal processing, information theory, channel coding, wireless communication) are mentioned in later chapters. In addition to these textbooks, an overview of many important topics can be found in the recently updated mobile communications handbook [12] edited by Gibson.

This book is intended to be accessible to readers who have never been exposed to communication systems before. It has some overlap with more advanced graduate texts (e.g., Chapters 2, 4, 5, and 6 here overlap heavily with Chapters 2 and 3 in the author's own graduate text [7]), and provides the technical background and motivation required to easily access these more advanced texts. Of course, the best way to continue building expertise in the field is by actually working in it. Research and development in this field requires study of the research literature, of more specialized texts (e.g., on information theory, channel coding, synchronization), and of commercial standards. The Institute for Electrical and Electronics Engineers (IEEE) is responsible for publication of many conference proceedings and journals in communications: major conferences include that IEEE Global Telecommunications Conference (Globecom) and the IEEE International Communications Conference (ICC), major journals and magazines include *IEEE Communications*

Magazine, the *IEEE Transactions on Communications*, and the *IEEE Journal on Selected Areas in Communications*. Closely related fields such as information theory and signal processing have their own conferences, journals, and magazines. Major conferences include the IEEE International Symposium on Information Theory (ISIT) and IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); journals include the *IEEE Transactions on Information Theory* and the *IEEE Transactions on Signal Processing*. The IEEE also publishes a number of standards online, such as the IEEE 802 family of standards for local-area networks.

A useful resource for learning source coding and data compression, which are not discussed in this text, is the textbook by Sayood [13]. Textbooks on core concepts in communication networks include Bertsekas and Gallager [14], Kumar, Manjunath, and Kuri [15], and Walrand and Varaiya [16].

2 Signals and systems

A communication link involves several stages of signal manipulation: the transmitter transforms the message into a signal that can be sent over a communication channel; the channel distorts the signal and adds noise to it; and the receiver processes the noisy received signal to extract the message. Thus, communication systems design must be based on a sound understanding of signals, and the systems that shape them. In this chapter, we discuss concepts and terminology from signals and systems, with a focus on how we plan to apply them in our discussion of communication systems. Much of this chapter is a review of concepts with which the reader might already be familiar from prior exposure to signals and systems. However, special attention should be paid to the discussion of baseband and passband signals and systems (Sections 2.7 and 2.8). This material, which is crucial for our purpose, is typically not emphasized in a first course on signals and systems. Additional material on the geometric relationship between signals is covered in later chapters, when we discuss digital communication.

Chapter plan

After a review of complex numbers and complex arithmetic in Section 2.1, we provide some examples of useful signals in Section 2.2. We then discuss LTI systems and convolution in Section 2.3. This is followed by Fourier series (Section 2.4) and the Fourier transform (Section 2.5). These sections (Sections 2.1 through Section 2.5) correspond to a review of material that is part of the assumed background for the core content of this textbook. However, even readers familiar with the material are encouraged to skim through it quickly in order to gain familiarity with the notation. This gets us to the point where we can classify signals and systems based on the frequency band they occupy. Specifically, we discuss baseband and passband signals and systems in Sections 2.7 and 2.8. Messages are typically baseband, while signals sent over channels (especially radio channels) are typically passband. We discuss methods for going from baseband to passband and back. We specifically emphasize the fact that a real-valued passband signal is equivalent (in a mathematically convenient and physically meaningful sense) to a complex-valued baseband signal, called the *complex-baseband representation*, or *complex envelope*, of the passband signal. We note that the information carried by a passband signal resides in its complex envelope, so that modulation (or the process of encoding messages in waveforms that can be sent over physical channels) consists of mapping information into a complex envelope, and then converting this complex envelope into a passband signal. We discuss the physical significance of the rectangular form of the complex envelope, which corresponds to the *in-phase (I)* and *quadrature (Q)* components of the passband signal, and that of the polar form of the complex envelope, which corresponds to the *envelope* and *phase* of the passband signal. We conclude by discussing the role of complex baseband in transceiver implementations, and by illustrating its use for wireless channel modeling.

Software

The software labs in this chapter introduce the use of MATLAB for signal processing. They provide practice in writing MATLAB code from scratch (i.e., without using prepackaged routines or Simulink) for simple computations. Software Lab 2.1 is an introduction to the use of MATLAB for typical operations of interest to us, and illustrates how we approximate continuous-time operations in discrete time. Software Lab 2.2 shows how to model and undo the effects of carrier-phase offsets in complex baseband. Software Lab 2.3 develops complex-baseband models for wireless multipath channels, and explores the phenomenon of signal fading due to constructive and destructive interference between the paths.

2.1 Complex numbers

A complex number z can be written as $z = x + jy$, where x and y are real numbers, and $j = \sqrt{-1}$. We say that $x = \mathrm{Re}(z)$ is the real part of z and $y = \mathrm{Im}(z)$ is the imaginary part of z. As depicted in Figure 2.1, it is often advantageous to interpret the complex number z as a two-dimensional real vector, which can be represented in rectangular form as $(x, y) = (\mathrm{Re}(z), \mathrm{Im}(z))$, or in polar form (r, θ) as

$$\begin{aligned} r &= |z| = \sqrt{x^2 + y^2} \\ \theta &= \underline{/z} = \tan^{-1}(y/x) \end{aligned} \tag{2.1}$$

We can go back from polar form to rectangular form as follows:

$$x = r\cos\theta, \quad y = r\sin\theta \tag{2.2}$$

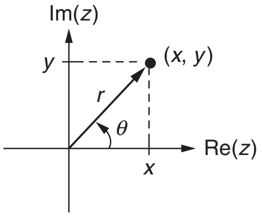

Figure 2.1 A complex number z represented in the two-dimensional real plane.

Complex conjugation For a complex number $z = x + jy = re^{j\theta}$, its complex conjugate

$$z^* = x - jy = re^{-j\theta} \tag{2.3}$$

That is,

$$\begin{gathered} \mathrm{Re}(z^*) = \mathrm{Re}(z), \quad \mathrm{Im}(z^*) = -\mathrm{Im}(z) \\ |z^*| = |z|, \quad \underline{/z^*} = -\underline{/z} \end{gathered} \tag{2.4}$$

The real and imaginary parts of a complex number z can be written in terms of z and z^* as follows:

$$\mathrm{Re}(z) = \frac{z + z^*}{2}, \quad \mathrm{Im}(z) = \frac{z - z^*}{2j} \tag{2.5}$$

Euler's formula This formula is of fundamental importance in complex analysis, and relates the rectangular and polar forms of a complex number:

$$e^{j\theta} = \cos\theta + j\sin\theta \tag{2.6}$$

The complex conjugate of $e^{j\theta}$ is given by

$$e^{-j\theta} = \left(e^{j\theta}\right)^* = \cos\theta - j\sin\theta$$

We can express cosines and sines in terms of $e^{j\theta}$ and its complex conjugate as follows:

$$\mathrm{Re}(e^{j\theta}) = \frac{e^{j\theta} + e^{-j\theta}}{2} = \cos\theta, \quad \mathrm{Im}(e^{j\theta}) = \frac{e^{j\theta} - e^{-j\theta}}{2j} = \sin\theta \tag{2.7}$$

On applying Euler's formula to (2.1), we can write

$$z = x + jy = r\cos\theta + jr\sin\theta = re^{j\theta} \tag{2.8}$$

Being able to go back and forth between the rectangular and polar forms of a complex number is useful. For example, it is easier to add in the rectangular form, but it is easier to multiply in the polar form.

Complex addition For two complex numbers $z_1 = x_1 + jy_1$ and $z_2 = x_2 + jy_2$,

$$z_1 + z_2 = (x_1 + x_2) + j(y_1 + y_2) \tag{2.9}$$

That is,

$$\mathrm{Re}(z_1 + z_2) = \mathrm{Re}(z_1) + \mathrm{Re}(z_2), \quad \mathrm{Im}(z_1 + z_2) = \mathrm{Im}(z_1) + \mathrm{Im}(z_2) \tag{2.10}$$

Complex multiplication (rectangular form) For two complex numbers $z_1 = x_1 + jy_1$ and $z_2 = x_2 + jy_2$,

$$z_1 z_2 = (x_1 x_2 - y_1 y_2) + j(y_1 x_2 + x_1 y_2) \tag{2.11}$$

This follows simply by multiplying out, and setting $j^2 = -1$. We have

$$\mathrm{Re}(z_1 z_2) = \mathrm{Re}(z_1)\mathrm{Re}(z_2) - \mathrm{Im}(z_1)\mathrm{Im}(z_2), \quad \mathrm{Im}(z_1 z_2) = \mathrm{Im}(z_1)\mathrm{Re}(z_2) + \mathrm{Re}(z_1)\mathrm{Im}(z_2) \tag{2.12}$$

Note that, using the rectangular form, a single complex multiplication requires four real multiplications.

Complex multiplication (polar form) Complex multiplication is easier when the numbers are expressed in polar form. For $z_1 = r_1 e^{j\theta_1}$ and $z_2 = r_2 e^{j\theta_2}$, we have

$$z_1 z_2 = r_1 r_2 e^{j(\theta_1 + \theta_2)} \tag{2.13}$$

That is,

$$|z_1 z_2| = |z_1||z_2|, \quad \underline{\angle z_1 z_2} = \underline{\angle z_1} + \underline{\angle z_2} \tag{2.14}$$

Division For two complex numbers $z_1 = x_1 + jy_1 = r_1 e^{j\theta_1}$ and $z_2 = x_2 + jy_2 = r_2 e^{j\theta_2}$ (with $z_2 \neq 0$, i.e., $r_2 > 0$), it is easiest to express the result of division in polar form:

$$z_1/z_2 = (r_1/r_2) e^{j(\theta_1 - \theta_2)} \tag{2.15}$$

That is,

$$|z_1/z_2| = |z_1|/|z_2|, \quad \underline{\angle z_1/z_2} = \underline{\angle z_1} - \underline{\angle z_2} \tag{2.16}$$

In order to divide using rectangular form, it is convenient to multiply the numerator and the denominator by z_2^*, which gives

$$z_1/z_2 = z_1 z_2^*/(z_2 z_2^*) = z_1 z_2^*/|z_2|^2 = \frac{(x_1 + jy_1)(x_2 - jy_2)}{x_2^2 + y_2^2}$$

On multiplying out as usual, we get

$$z_1/z_2 = \frac{(x_1 x_2 + y_1 y_2) + j(-x_1 y_2 + y_1 x_2)}{x_2^2 + y_2^2} \tag{2.17}$$

Example 2.1.1 (Computations with complex numbers) Consider the complex numbers $z_1 = 1+j$ and $z_2 = 2e^{-j\pi/6}$. Find $z_1 + z_2$, $z_1 z_2$, and z_1/z_2. Also specify z_1^* and z_2^*.

For complex addition, it is convenient to express both numbers in rectangular form. Thus,

$$z_2 = 2(\cos(-\pi/6) + j\sin(-\pi/6)) = \sqrt{3} - j$$

and

$$z_1 + z_2 = (1 + j) + \left(\sqrt{3} - j\right) = \sqrt{3} + 1$$

For complex multiplication and division, it is convenient to express both numbers in polar form. We obtain $z_1 = \sqrt{2} e^{j\pi/4}$ by applying (2.1). Now, from (2.11), we have

$$z_1 z_2 = \sqrt{2} e^{j\pi/4} 2 e^{-j\pi/6} = 2\sqrt{2} e^{j(\pi/4 - \pi/6)} = 2\sqrt{2} e^{j\pi/12}$$

Similarly,

$$z_1/z_2 = \frac{\sqrt{2} e^{j\pi/4}}{2 e^{-j\pi/6}} = \frac{1}{\sqrt{2}} e^{j(\pi/4 + \pi/6)} = \frac{1}{\sqrt{2}} e^{j5\pi/12}$$

Multiplication using the rectangular forms of the complex numbers yields the following:

$$z_1 z_2 = (1 + j)\left(\sqrt{3} - j\right) = \sqrt{3} - j + \sqrt{3}j + 1 = \left(\sqrt{3} + 1\right) + j\left(\sqrt{3} - 1\right)$$

Note that $z_1^* = 1 - j = \sqrt{2}e^{-j\pi/4}$ and $z_2^* = 2e^{j\pi/6} = \sqrt{3} + j$. Division using rectangular forms gives

$$z_1/z_2 = z_1 z_2^*/|z_2|^2 = (1+j)\left(\sqrt{3}+j\right)\Big/ 2^2 = \frac{\sqrt{3}-1}{4} + j\frac{\sqrt{3}+1}{4}$$

No need to memorize trigonometric identities any more Once we can do computations using complex numbers, we can use Euler's formula to quickly derive well-known trigonometric identities involving sines and cosines. For example,

$$\cos(\theta_1 + \theta_2) = \mathrm{Re}(e^{j(\theta_1+\theta_2)})$$

But

$$\begin{aligned} e^{j(\theta_1+\theta_2)} &= e^{j\theta_1} e^{j\theta_2} = (\cos\theta_1 + j\sin\theta_1)(\cos\theta_2 + j\sin\theta_2) \\ &= (\cos\theta_1\cos\theta_2 - \sin\theta_1\sin\theta_2) + j(\cos\theta_1\sin\theta_2 + \sin\theta_1\cos\theta_2) \end{aligned}$$

Taking the real part, we can read off the identity

$$\cos(\theta_1 + \theta_2) = \cos\theta_1\cos\theta_2 - \sin\theta_1\sin\theta_2 \tag{2.18}$$

Moreover, taking the imaginary part, we can read off

$$\sin(\theta_1 + \theta_2) = \cos\theta_1\sin\theta_2 + \sin\theta_1\cos\theta_2 \tag{2.19}$$

2.2 Signals

Signal A signal $s(t)$ is a function of time (or some other independent variable, such as frequency, or spatial coordinates) that has an interesting physical interpretation. For example, it is generated by a transmitter, or processed by a receiver. While physically realizable signals such as those sent over a wire or over the air must take real values, we shall see that it is extremely useful (and physically meaningful) to consider *a pair* of real-valued signals, interpreted as the real and imaginary parts of a complex-valued signal. Thus, in general, we allow signals to take complex values.

Discrete versus continuous time We generically use the notation $x(t)$ to denote continuous-time signals (t taking real values), and $x[n]$ to denote discrete-time signals (n taking integer values). A continuous-time signal $x(t)$ sampled at rate T_s produces discrete time samples $x(nT_s + t_0)$ (t_0 is an arbitrary offset), which we often denote as a discrete-time signal $x[n]$. While signals sent over a physical communication channel are inherently continuous-time, implementations both at the transmitter and at the receiver make heavy use of discrete-time implementations on digitized samples corresponding to the analog continuous-time waveforms of interest.

We now introduce some signals that recur often in this text.

Sinusoid This is a periodic function of time of the form

$$s(t) = A\cos(2\pi f_0 t + \theta) \tag{2.20}$$

where $A > 0$ is the amplitude, f_0 is the frequency, and $\theta \in [0, 2\pi]$ is the phase. By setting $\theta = 0$, we obtain a pure cosine $A\cos(2\pi f_c t)$, and by setting $\theta = -\pi/2$, we obtain a pure sine $A\sin(2\pi f_c t)$. In general, using (2.18), we can rewrite (2.20) as

$$s(t) = A_c \cos(2\pi f_0 t) - A_s \sin(2\pi f_0 t) \tag{2.21}$$

where $A_c = A\cos\theta$ and $A_s = A\sin\theta$ are real numbers. Using Euler's formula, we can write

$$Ae^{j\theta} = A_c + jA_s \tag{2.22}$$

Thus, the parameters of a sinusoid at frequency f_0 can be represented by the complex number in (2.22), with (2.20) using the polar form, and (2.21) the rectangular form, of this number. Note that $A = \sqrt{A_c^2 + A_s^2}$ and $\theta = \tan^{-1}(A_s/A_c)$.

Clearly, sinusoids with known amplitude, phase, and frequency are perfectly predictable, and hence cannot carry any information. As we shall see, information can be transmitted by making the complex number $Ae^{j\theta} = A_c + jA_s$ associated with the parameters of the sinusoid vary in a way that depends on the message to be conveyed. Of course, once this has been done, the resulting signal will no longer be a pure sinusoid, and part of the work of the communication system designer is to decide what shape such a signal should take in the frequency domain.

We now define complex exponentials, which play a key role in understanding signals and systems in the frequency domain.

Complex exponential A complex exponential at a frequency f_0 is defined as

$$s(t) = Ae^{j(2\pi f_0 t + \theta)} = \alpha e^{j2\pi f_0 t} \tag{2.23}$$

where $A > 0$ is the amplitude, f_0 is the frequency, $\theta \in [0, 2\pi]$ is the phase, and $\alpha = Ae^{j\theta}$ is a complex number that contains both the amplitude and the phase information. Let us now make three observations. First, note the ease with which we handle amplitude and phase for complex exponentials: they simply combine into a complex number that factors out of the complex exponential. Second, by Euler's formula,

$$\text{Re}(Ae^{j(2\pi f_0 t + \theta)}) = A\cos(2\pi f_0 t + \theta)$$

so that real-valued sinusoids are "contained in" complex exponentials. Third, as we shall soon see, the set of complex exponentials $\{e^{j2\pi ft}\}$, where f takes values in $(-\infty, \infty)$, forms a "basis" for a large class of signals (basically, for all signals that are of interest to us), and the Fourier transform of a signal is simply its expansion with respect to this basis. Such observations are key to why complex exponentials play such an important role in signals and systems in general, and in communication systems in particular.

The delta, or impulse, function Another signal that plays a crucial role in signals and systems is the delta function, or the unit impulse, which we denote by $\delta(t)$. Physically, we can think of it as a narrow, tall pulse with unit area: examples are shown in Figure 2.2. Mathematically, we can think of it as a limit of such pulses as the pulse width shrinks (and hence the

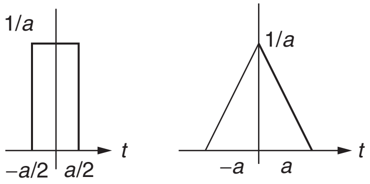

Figure 2.2 The impulse function may be viewed as a limit of tall thin pulses ($a \rightarrow 0$ in the examples shown in the figure).

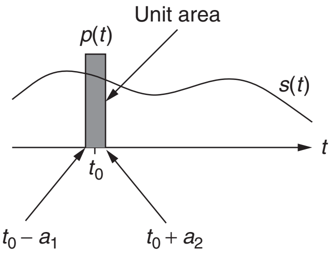

Figure 2.3 Multiplying a signal by a tall thin pulse to select its value at t_0.

pulse height goes to infinity). Such a limit is not physically realizable, but it serves a very useful purpose in terms of understanding the structure of physically realizable signals. To see this, consider a signal $s(t)$ that varies smoothly, and multiply it by a tall, thin pulse of unit area, centered at time t_0, as shown in Figure 2.3. If we now integrate the product, we obtain

$$\int_{-\infty}^{\infty} s(t)p(t)dt = \int_{t_0-a_1}^{t_0+a_2} s(t)p(t)dt \approx s(t_0) \int_{t_0-a_1}^{t_0+a_1} p(t)dt = s(t_0)$$

That is, the preceding operation "selects" the value of the signal at time t_0. On taking the limit of the tall thin pulse as its width $a_1 + a_2 \rightarrow 0$, we get a translated version of the delta function, namely $\delta(t - t_0)$. Note that the exact shape of the pulse does not matter in the preceding argument. The delta function is therefore *defined* by means of the following sifting property: for any "smooth" function $s(t)$, we have

$$\int_{-\infty}^{\infty} s(t)\delta(t - t_0)dt = s(t_0) \quad \textbf{sifting property of the impulse} \qquad (2.24)$$

Thus, the delta function is defined mathematically by the way it acts on other signals, rather than as a signal by itself. However, it is also important to keep in mind its intuitive interpretation as (the limit of) a tall, thin, pulse of unit area.

The following function is useful for expressing signals compactly.

Indicator function We use I_A to denote the indicator function of a set A, defined as

$$I_A(x) = \begin{cases} 1, & x \in A \\ 0, & \text{otherwise} \end{cases}$$

The indicator function of an interval is a rectangular pulse, as shown in Figure 2.4.

The indicator function can also be used to compactly express more complex signals, as shown in the examples in Figure 2.5.

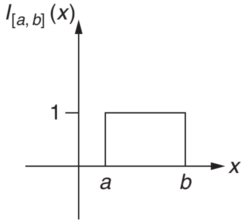

Figure 2.4 The indicator function of an interval is a rectangular pulse.

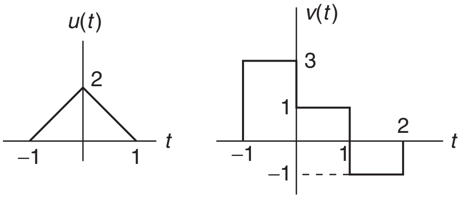

Figure 2.5 The functions $u(t) = 2(1 - |t|)I_{[-1,1]}(t)$ and $v(t) = 3I_{[-1,0]}(t) + I_{[0,1]}(t) - I_{[1,2]}(t)$ can be written compactly in terms of indicator functions.

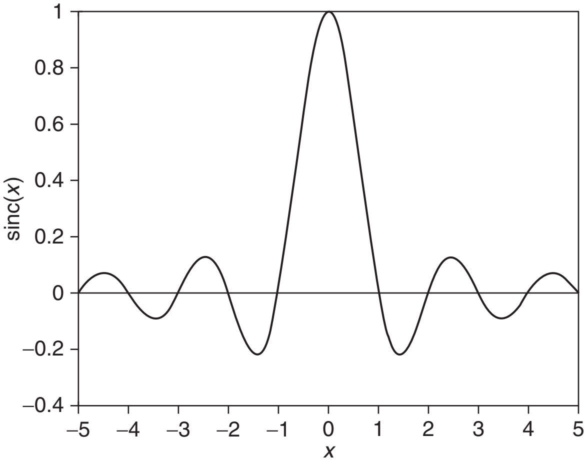

Figure 2.6 The sinc function.

Sinc function The sinc function, plotted in Figure 2.6, is defined as

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

where the value at $x = 0$ is defined in the limit as $x \to 0$ to be $\text{sinc}(0) = 1$. Since $|\sin(\pi x)| \le 1$, we have that $|\text{sinc}(x)| \le 1/(\pi |x|)$, with equality if and only if x is an odd multiple of $1/2$. That is, the sinc function exhibits a sinusoidal variation, with an envelope that decays as $1/|x|$.

The analogy between signals and vectors Even though signals can be complicated functions of time that live in an infinite-dimensional space, the mathematics for manipulating them is very similar to that for manipulating finite-dimensional vectors, with sums replaced by integrals. A key building block of communication theory is the relative geometry of the signals used, which is governed by the inner products between signals. Inner products for continuous-time signals can be defined in a manner exactly analogous to the corresponding definitions in finite-dimensional vector space.

Inner product The inner product for two $m \times 1$ complex vectors $\mathbf{s} = (s[1], \ldots, s[m])^{\mathrm{T}}$ and $\mathbf{r} = (r[1], \ldots, r[m])^{\mathrm{T}}$ is given by

$$\langle \mathbf{s}, \mathbf{r} \rangle = \sum_{i=1}^{m} s[i] r^*[i] = \mathbf{r}^{\mathrm{H}} \mathbf{s} \tag{2.25}$$

Similarly, we define the inner product of two (possibly complex-valued) signals $s(t)$ and $r(t)$ as follows:

$$\langle s, r \rangle = \int_{-\infty}^{\infty} s(t) r^*(t) dt \tag{2.26}$$

The inner product obeys the following linearity properties:

$$\langle a_1 s_1 + a_2 s_2, r \rangle = a_1 \langle s_1, r \rangle + a_2 \langle s_2, r \rangle$$
$$\langle s, a_1 r_1 + a_2 r_2 \rangle = a_1^* \langle s, r_1 \rangle + a_2^* \langle s, r_2 \rangle$$

where a_1 and a_2 are complex-valued constants, and s, s_1, s_2, r, r_1, and r_2 are signals (or vectors). The complex conjugation when we pull out constants from the second argument of the inner product is something that we need to maintain awareness of when computing inner products for complex-valued signals.

Energy and norm The *energy* E_s of a signal s is defined as its inner product with itself:

$$E_s = ||s||^2 = \langle s, s \rangle = \int_{-\infty}^{\infty} |s(t)|^2 \, dt \tag{2.27}$$

where $||s||$ denotes the *norm* of s. If the energy of s is zero, then s must be zero "almost everywhere" (e.g., $s(t)$ cannot be nonzero over any interval, no matter how small its length). For continuous-time signals, we take this to be equivalent to being zero everywhere. With this understanding, $||s|| = 0$ implies that s is zero, which is a property that is true for norms in finite-dimensional vector spaces.

Example 2.2.1 (Energy computations) Consider $s(t) = 2I_{[0,T]} + jI_{[T/2,2T]}$. On writing it out in more detail, we have

$$s(t) = \begin{cases} 2, & 0 \le t < T/2 \\ 2 + j, & T/2 \le t < T \\ j, & T \le t < 2T \end{cases}$$

so that its energy is given by

$$||s||^2 = \int_0^{T/2} 2^2 \, dt + \int_{T/2}^{T} |2 + j|^2 \, dt + \int_T^{2T} |j|^2 \, dt = 4(T/2) + 5(T/2) + T = 11T/2$$

As another example, consider $s(t) = e^{-3|t|+j2\pi t}$, for which the energy is given by

$$||s||^2 = \int_{-\infty}^{\infty} |e^{-3|t|+j2\pi t}|^2 \, dt = \int_{-\infty}^{\infty} e^{-6|t|} \, dt = 2\int_0^{\infty} e^{-6t} \, dt = 1/3$$

Note that the complex phase term $j2\pi t$ does not affect the energy, since it goes away when we take the magnitude.

Power The power of a signal $s(t)$ is defined as the time average of its energy computed over a large time interval:

$$P_s = \lim_{T_o \to \infty} \frac{1}{T_o} \int_{-T_o/2}^{T_o/2} |s(t)|^2 \, dt \tag{2.28}$$

Finite-energy signals, of course, have zero power.

We see from (2.28) that power is defined as a time average. It is useful to introduce a compact notation for time averages.

Time average For a function $g(t)$, define the time average as

$$\overline{g} = \lim_{T_o \to \infty} \frac{1}{T_o} \int_{-T_o/2}^{T_o/2} g(t) dt \tag{2.29}$$

That is, we compute the time average over an observation interval of length T_o, and then let the observation interval get large. We can now rewrite the power computation in (2.28) in this notation as follows.

Power The power of a signal $s(t)$ is defined as

$$P_s = \overline{|s(t)|^2} \tag{2.30}$$

Another time average of interest is the DC value of a signal.

DC value The DC value of $s(t)$ is defined as $\overline{s(t)}$.

Let us compute these quantities for the simple example of a complex exponential, $s(t) = Ae^{j(2\pi f_0 t+\theta)}$, where $A > 0$ is the amplitude, $\theta \in [0, 2\pi]$ is the phase, and f_0 is a real-valued frequency. Since $|s(t)|^2 \equiv A^2$ for all t, we get the same value when we average it. Thus, the power is given by $P_s = \overline{s^2(t)} = A^2$. For nonzero frequency f_0, it is intuitively clear that all the power in s is concentrated away from DC, since $s(t) = Ae^{j(2\pi f_0 t+\theta)} \leftrightarrow S(f) = Ae^{j\theta}\delta(f - f_0)$. We therefore see that the DC value is zero. While this is a convincing intuitive argument, it is instructive to prove this starting from the definition (2.29).

Proving that a complex exponential has zero DC value For $s(t) = Ae^{j(2\pi f_0 t+\theta)}$, the integral over its period (of length $1/f_0$) is zero. As shown in Figure 2.7, the length L of any interval I can be written as $L = K/f_0 + \ell$, where K is a nonnegative integer and $0 \le \ell < 1/f_0$ is the length of the remaining interval I_r. Since the integral over an integer number of periods is zero, we have

$$\int_I s(t)dt = \int_{I_r} s(t)dt$$

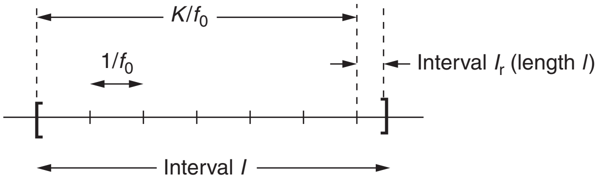

Figure 2.7 The interval I for computing the time average of a periodic function with period $1/f_0$ can be decomposed into an integer number K of periods, with the remaining interval I_r of length $\ell < 1/f_0$.

Thus,

$$\left|\int_I s(t)dt\right| = \left|\int_{I_r} s(t)dt\right| \le \ell \max_t |s(t)| = A\ell < \frac{A}{f_0}$$

since $|s(t)| = A$. We therefore obtain

$$\left|\int_{-T_o/2}^{T_o/2} s(t)dt\right| \le A/f_0$$

which yields that the DC value $\bar{s} = 0$, since

$$|\bar{s}| = \left|\lim_{T_o \to \infty} \frac{1}{T_o} \int_{-T_o/2}^{T_o/2} s(t)dt\right| \le \lim_{T_o \to \infty} \frac{A}{f_0 T_o} = 0$$

Essentially the same argument implies that, in general, the time average of a periodic signal equals the average over a single period. We use this fact without further comment henceforth.

Power and DC value of a sinusoid For a real-valued sinusoid $s(t) = A\cos(2\pi f_0 t + \theta)$, we can use the results derived for complex exponentials above. Using Euler's identity, a real-valued sinusoid at f_0 is a sum of complex exponentials at $\pm f_0$:

$$s(t) = \frac{A}{2} e^{j(2\pi f_0 t + \theta)} + \frac{A}{2} e^{-j(2\pi f_0 t + \theta)}$$

Since each complex exponential has zero DC value, we obtain

$$\bar{s} = 0$$

That is, the DC value of any real-valued sinusoid is zero. Then

$$P_s = \overline{s^2(t)} = \overline{A^2 \cos^2(2\pi f_0 t + \theta)} = \overline{\frac{A^2}{2} + \frac{A^2}{2}\cos(4\pi f_0 t + 2\theta)} = \frac{A^2}{2}$$

since the DC value of the sinusoid at $2f_0$ is zero.

2.3 Linear time-invariant systems

System A system takes as input one or more signals, and produces as output one or more signals. A system is specified once we characterize its input–output relationship; that is, if we can determine the output, or response, $y(t)$, corresponding to any possible input $x(t)$ in a given class of signals of interest.

Our primary focus here is on *linear time-invariant (LTI)* systems, which provide good models for filters at the transmitter and receiver, as well as for the distortion induced by a variety of channels. We shall see that the input–output relationship is particularly easy to characterize for such systems.

Linear system Let $x_1(t)$ and $x_2(t)$ denote arbitrary input signals, and let $y_1(t)$ and $y_2(t)$ denote the corresponding system outputs, respectively. Then, for arbitrary scalars a_1 and a_2, the response of the system to input $a_1x_1(t) + a_2x_2(t)$ is $a_1y_1(t) + a_2y_2(t)$.

Time-invariant system Let $y(t)$ denote the system response to an input $x(t)$. Then the system response to a time-shifted version of the input, $x_1(t) = x(t - t_0)$, is $y_1(t) = y(t - t_0)$. That is, a time shift in the input causes an identical time shift in the output.

Example 2.3.1 (Examples of linear systems) It can (and should) be checked that the following systems are linear. These examples show that linear systems may, but need not, be time-invariant:

$$
\begin{aligned}
y(t) &= 2x(t-1) - jx(t-2) \quad \textit{time-invariant}\\
y(t) &= (3-2j)x(1-t) \quad \textit{time-varying}\\
y(t) &= x(t)\cos(100\pi t) - x(t-1)\sin(100\pi t) \quad \textit{time-varying}\\
y(t) &= \int_{t-1}^{t+1} x(\tau)d\tau \quad \textit{time-invariant}
\end{aligned}
$$

Example 2.3.2 (Examples of time-invariant systems) It can (and should) be checked that the following systems are time-invariant. These examples show that time-invariant systems may, but need not, be linear:

$$
\begin{aligned}
y(t) &= e^{2x(t-1)} \quad \textit{nonlinear}\\
y(t) &= \int_{-\infty}^{t} x(\tau)e^{-(t-\tau)}\,d\tau \quad \textit{linear}\\
y(t) &= \int_{t-1}^{t+1} x^2(\tau)d\tau \quad \textit{nonlinear}
\end{aligned}
$$

Linear time-invariant system A linear time-invariant (LTI) system is (unsurprisingly) defined to be a system that is both linear and time-invariant. What is surprising, however, is how powerful the LTI property is in terms of dictating what the input–output relationship must look like. Specifically, if we know the *impulse response* of an LTI system (i.e., the output signal when the input signal is the delta function), then we can compute the system response for *any* input signal. Before deriving and stating this result, we illustrate the LTI property using an example; see Figure 2.8. Suppose that the response of an LTI system to the rectangular pulse $p_1(t) = I_{[-1/2,1/2]}(t)$ is given by the trapezoidal waveform $h_1(t)$.

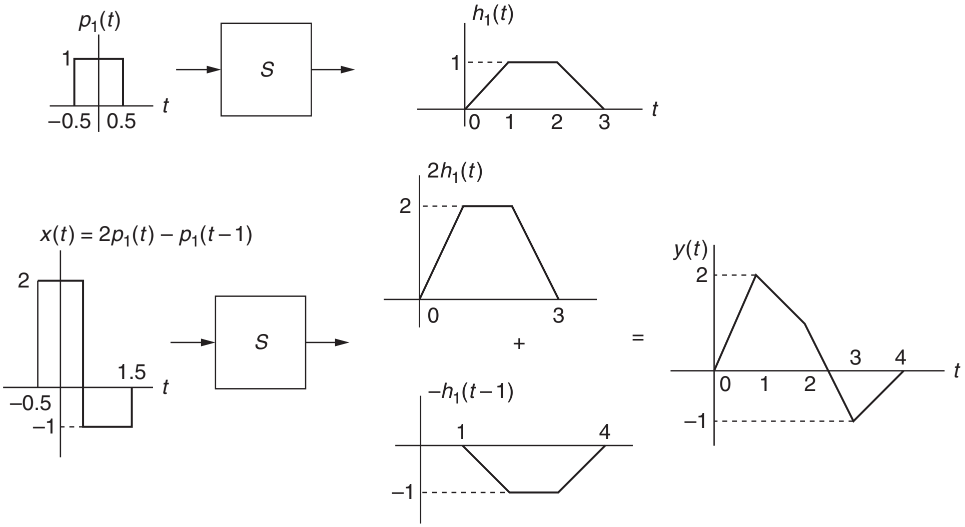

Figure 2.8 Given that the response of an LTI system $\mathcal{S}$ to the pulse $p_1(t)$ is $h_1(t)$, we can use the LTI property to infer that the response to $x(t) = 2p_1(t) - p_1(t-1)$ is $y(t) = 2h_1(t) - h_1(t-1)$.

We can now compute the system response to any linear combination of time shifts of the pulse $p(t)$, as illustrated by the example in Figure 2.8. More generally, using the LTI property, we infer that the response to an input signal of the form $x(t) = \sum_i a_i p_1(t - t_i)$ is $y(t) = \sum_i a_i h_1(t - t_i)$.

Can we extend the preceding idea to compute the system response to arbitrary input signals? The answer is yes: if we know the system response to thinner and thinner pulses, then we can approximate arbitrary signals better and better using linear combinations of shifts of these pulses. Consider $p_\Delta(t) = (1/\Delta)I_{[-\Delta/2,\Delta/2]}(t)$, where $\Delta > 0$ is getting smaller and smaller. Note that we have normalized the area of the pulse to unity, so that the limit of $p_\Delta(t)$ as $\Delta \to 0$ is the delta function. Figure 2.9 shows how to approximate a smooth input signal as a linear combination of shifts of $p_\Delta(t)$. That is, for Δ small, we have

$$x(t) \approx x_\Delta(t) = \sum_{k=-\infty}^{\infty} x(k\Delta)\Delta p_\Delta(t - k\Delta) \tag{2.31}$$

If the system response to $p_\Delta(t)$ is $h_\Delta(t)$, then we can use the LTI property to compute the response $y_\Delta(t)$ to $x_\Delta(t)$, and use this to approximate the response $y(t)$ to the input $x(t)$, as follows:

$$y(t) \approx y_\Delta(t) = \sum_{k=-\infty}^{\infty} x(k\Delta)\Delta h_\Delta(t - k\Delta) \tag{2.32}$$

As $\Delta \to 0$, the sums above tend to integrals, and the pulse $p_\Delta(t)$ tends to the delta function $\delta(t)$. The approximation to the input signal in Equation (2.31) becomes exact, with the sum tending to an integral:

$$\lim_{\Delta \to 0} x_\Delta(t) = x(t) = \int_{-\infty}^{\infty} x(\tau)\delta(t - \tau)d\tau$$

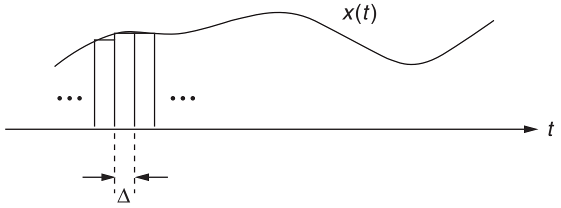

Figure 2.9 A smooth signal can be approximated as a linear combination of shifts of tall thin pulses.

replacing the discrete time shifts $k\Delta$ by the continuous variable τ, the discrete increment Δ by the infinitesimal $d\tau$, and the sum by an integral. This is just a restatement of the sifting property of the impulse. That is, an arbitrary input signal can be expressed as a linear combination of time-shifted versions of the delta function, where we now consider a continuum of time shifts.

In similar fashion, the approximation to the output signal in (2.32) becomes exact, with the sum reducing to the following *convolution* integral:

$$\lim_{\Delta\to 0} y_\Delta(t) = y(t) = \int_{-\infty}^{\infty} x(\tau)h(t-\tau)d\tau \tag{2.33}$$

where $h(t)$ denotes the *impulse response* of the LTI system.

Convolution and its computation The convolution $v(t)$ of two signals $u_1(t)$ and $u_2(t)$ is given by

$$v(t) = (u_1 * u_2)(t) = \int_{-\infty}^{\infty} u_1(\tau)u_2(t-\tau)d\tau = \int_{-\infty}^{\infty} u_1(t-\tau)u_2(\tau)d\tau \tag{2.34}$$

Note that τ is a dummy variable that is integrated out in order to determine the value of the signal $v(t)$ at each possible time t. The roles of u_1 and u_2 in the integral can be exchanged. This can be proved using a change of variables, replacing $t-\tau$ by τ. We often drop the time variable, and write $v = u_1 * u_2 = u_2 * u_1$.

An LTI system is completely characterized by its impulse response As derived in (2.33), the output y of an LTI system is the convolution of the input signal u and the system impulse response h. That is, $y = u * h$. From (2.34), we realize that the roles of the signal and the system can be exchanged: that is, we would get the same output y if a signal h were sent through a system with impulse response u.

Flip and slide Consider the expression for the convolution in (2.34):

$$v(t) = \int_{-\infty}^{\infty} u_1(\tau)u_2(t-\tau)d\tau$$

Fix a value of time t at which we wish to evaluate v. In order to compute $v(t)$, we must multiply two functions of a "dummy variable" τ and then integrate over τ. In particular, $s_2(\tau) = u_2(-\tau)$ is the signal $u_2(\tau)$ flipped around the origin, so that $u_2(t-\tau) = u_2(-(\tau-t)) = s_2(\tau-t)$ is $s_2(\tau)$ translated to the right by t (if $t < 0$, translation to the right by

t actually corresponds to translation to the left by $|t|$). In short, the mechanics of computing the convolution involves flipping and sliding one of the signals, multiplying by the other signal, and integrating. Pictures are extremely helpful when doing such computations by hand, as illustrated by the following example.

Example 2.3.3 (Convolving rectangular pulses) Consider the rectangular pulses $u_1(t) = I_{[5,11]}(t)$ and $u_2(t) = I_{[1,3]}(t)$. We wish to compute the convolution

$$v(t) = (u_1 * u_2)(t) = \int_{-\infty}^{\infty} u_1(\tau)u_2(t-\tau)d\tau$$

We now draw pictures of the signals involved in these "flip and slide" computations in order to figure out the limits of integration for different ranges of t. Figure 2.10 shows that there are five different ranges of interest, and yields the following results.

(a) For $t < 6$, $u_1(\tau)u_2(t-\tau) \equiv 0$, so that $v(t) = 0$.

(b) For $6 < t < 8$, $u_1(\tau)u_2(t-\tau) = 1$ for $5 < \tau < t-1$, so that

$$v(t) = \int_5^{t-1} d\tau = t-6$$

(c) For $8 < t < 12$, $u_1(\tau)u_2(t-\tau) = 1$ for $t-3 < \tau < t-1$, so that

$$v(t) = \int_{t-3}^{t-1} d\tau = 2$$

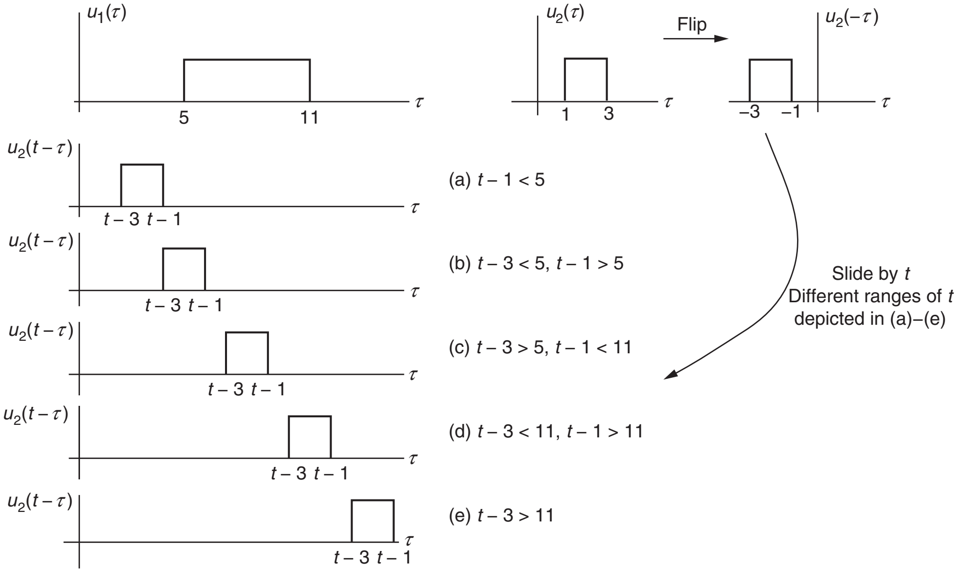

Figure 2.10 Illustrating the "flip and slide" operation for the convolution of two rectangular pulses.

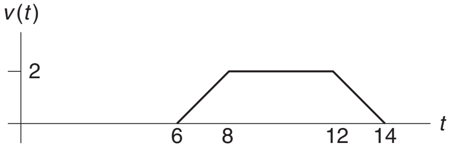

Figure 2.11 The convolution of the two rectangular pulses in Example 2.3.3 results in a trapezoidal pulse.

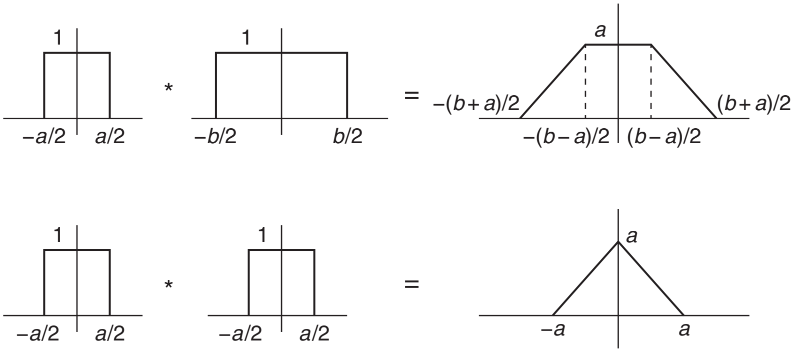

Figure 2.12 Convolution of two rectangular pulses as a function of the pulse durations. The trapezoidal pulse reduces to a triangular pulse for equal pulse durations.

(d) For $12 < t < 14$, $u_1(\tau)u_2(t-\tau) = 1$ for $t-3 < \tau < 11$, so that

$$v(t) = \int_{t-3}^{11} d\tau = 11 - (t-3) = 14 - t$$

(e) For $t > 14$, $u_1(\tau)u_2(t-\tau) \equiv 0$, so that $v(t) = 0$.

The result of the convolution is the trapezoidal pulse sketched in Figure 2.11.

It is useful to record the general form of the convolution between two rectangular pulses of the form $I_{[-a/2,a/2]}(t)$ and $I_{[-b/2,b/2]}(t)$, where we take $a \leq b$ without loss of generality. The result is a trapezoidal pulse, which reduces to a triangular pulse for $a = b$, as shown in Figure 2.12. Once we know this, using the LTI property, we can infer the convolution of any signals that can be expressed as linear combinations of shifts of rectangular pulses.

Occasional notational sloppiness can be useful As the preceding example shows, a convolution computation as in (2.34) requires a careful distinction between the variable t at which the convolution is being evaluated and the dummy variable τ. This is why we make sure that the dummy variable does not appear in our notation $(s * r)(t)$ for the convolution between signals $s(t)$ and $r(t)$. However, it is sometimes convenient to abuse notation and use the notation $s(t) * r(t)$ instead, as long we remain aware of what we are doing. For example, this enables us to compactly state the following linear time-invariance (LTI) property:

$$(a_1 s_1(t-t_1) + a_2 s_2(t-t_2)) * r(t) = a_1(s_1 * r)(t-t_1) + a_2(s_2 * r)(t-t_2)$$

for any complex gains a_1 and a_2, and any time offsets t_1 and t_2.

Example 2.3.4 (Modeling a multipath channel) We can get a delayed version of a signal by convolving it with a delayed impulse as follows:

$$y_1(t) = u(t) * \delta(t - t_1) = u(t - t_1) \tag{2.35}$$

To see this, compute

$$y_1(t) = \int u(\tau)\delta(t - \tau - t_1)d\tau = \int u(\tau)\delta(\tau - (t - t_1))d\tau = u(t - t_1)$$

where we first use the fact that the delta function is even, and then use its sifting property.

Equation (2.35) immediately tells us how to model *multipath channels,* in which multiply scattered versions of a transmitted signal $u(t)$ combine to give a received signal $y(t)$ that is a superposition of delayed versions of the transmitted signal, as illustrated in Figure 2.13:

$$y(t) = \alpha_1 u(t - \tau_1) + \cdots + \alpha_{\mathrm{m}} u(t - \tau_{\mathrm{m}})$$

(plus noise, which we have not talked about yet). From (2.35), we see that we can write

$$\begin{aligned} y(t) &= \alpha_1 u(t) * \delta(t - \tau_1) + \cdots + \alpha_{\mathrm{m}} u(t) * \delta(t - \tau_{\mathrm{m}}) \\ &= u(t) * (\alpha_1 \delta(t - \tau_1) + \cdots + \alpha_{\mathrm{m}} \delta(t - \tau_{\mathrm{m}})) \end{aligned}$$

That is, we can model the received signal as a convolution of the transmitted signal with a channel impulse response that is a linear combination of time-shifted impulses:

$$h(t) = \alpha_1 \delta(t - \tau_1) + \cdots + \alpha_{\mathrm{m}} \delta(t - \tau_{\mathrm{m}}) \tag{2.36}$$

Figure 2.14 illustrates how a rectangular pulse spreads as it goes through a multipath channel with impulse response $h(t) = \delta(t - 1) - 0.5\delta(t - 1.5) + 0.5\delta(t - 3.5)$. While the gains $\{\alpha_k\}$ in this example are real-valued, as we shall soon see (in Section 2.8), we need to allow both the signal $u(t)$ and the gains $\{\alpha_k\}$ to take complex values in order to model, for example, signals carrying information over radio channels.

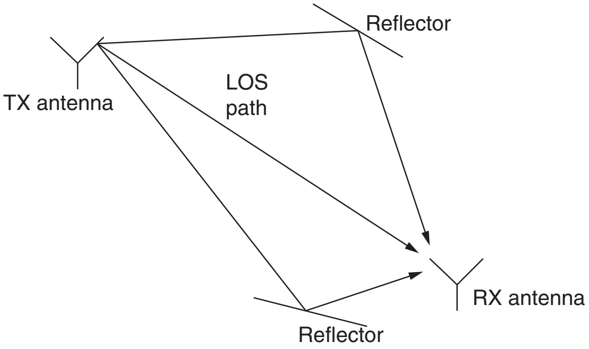

Figure 2.13 Multipath channels typical of wireless communication can include line-of-sight (LOS) and reflected paths.

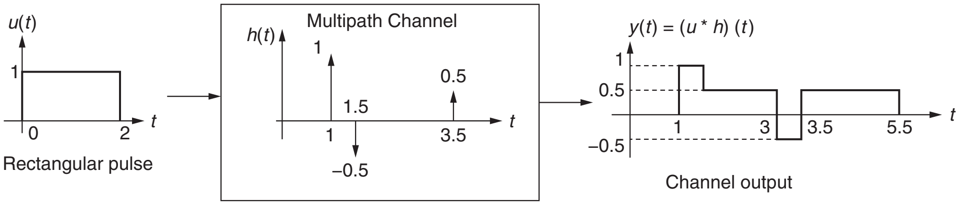

Figure 2.14 A rectangular pulse through a multipath channel.

Complex exponential through an LTI system In order to understand LTI systems in the frequency domain, let us consider what happens to a complex exponential $u(t) = e^{j2\pi f_0 t}$ when it goes through an LTI system with impulse response $h(t)$. The output is given by

$$y(t) = (u * h)(t) = \int_{-\infty}^{\infty} h(\tau)e^{j2\pi f_0(t-\tau)}\, d\tau$$
$$= e^{j2\pi f_0 t} \int_{-\infty}^{\infty} h(\tau)e^{-j2\pi f_0 \tau}\, d\tau = H(f_0)e^{j2\pi f_0 t} \tag{2.37}$$

where

$$H(f_0) = \int_{-\infty}^{\infty} h(\tau)e^{-j2\pi f_0 \tau}\, d\tau$$

is the Fourier transform of h evaluated at the frequency f_0. We shall discuss the Fourier transform and its properties in more detail shortly.

Complex exponentials are eigenfunctions of LTI systems Recall that an eigenvector of a matrix $\mathbf{H}$ is any vector $\mathbf{x}$ that satisfies $\mathbf{Hx} = \lambda\mathbf{x}$. That is, the matrix leaves its eigenvectors unchanged except for a scale factor λ, which is the eigenvalue associated with that eigenvector. In an entirely analogous fashion, we see that the complex exponential signal $e^{j2\pi f_0 t}$ is an *eigenfunction* of the LTI system with impulse response h, with eigenvalue $H(f_0)$. See Figure 2.15. Since we have not constrained h, we conclude that complex exponentials are eigenfunctions of *any* LTI system. We shall soon see, when we discuss Fourier transforms, that this eigenfunction property allows us to characterize LTI systems in the frequency-domain, which in turn enables powerful frequency domain design and analysis tools.

2.3.1 Discrete-time convolution

DSP-based implementations of convolutions are inherently discrete-time operations. For two discrete-time sequences $\{u_1[n]\}$ and $\{u_2[n]\}$, their convolution $y = u_1 * u_2$ is defined analogously to continuous-time convolution, replacing integration by summation:

$$y[n] = \sum_k u_1[k]u_2[n-k] \tag{2.38}$$

MATLAB implements this using the “conv” function. This can be interpreted as u_1 being the input to a system with impulse response u_2, where a discrete time impulse is simply a one, followed by all zeros.