CHAPTER

1 Introduction

We define communication as information transfer between different points in space or time, where the term *information* is loosely employed to cover standard formats that we are all familiar with, such as voice, audio, video, data files, web pages, etc. Examples of communication between two points in space include a telephone conversation, accessing an Internet website from our home or office computer, or tuning in to a TV or radio station. Examples of communication between two points in time include accessing a storage device, such as a record, CD, DVD, or hard drive. In the preceding examples, the information transferred is directly available for human consumption. However, there are many other communication systems, which we do not directly experience, but which form a crucial part of the infrastructure that we rely upon in our daily lives. Examples include high-speed packet transfer between routers on the Internet, inter- and intra-chip communication in integrated circuits, the connections between computers and computer peripherals (such as keyboards and printers), and control signals in communication networks.

In *digital* communication, the information being transferred is represented in digital form, most commonly as binary digits, or *bits*. This is in contrast to *analog* information, which takes on a continuum of values. Most communication systems used for transferring information today are either digital, or are being converted from analog to digital. Examples of some recent conversions that directly impact consumers include cellular telephony (from analog FM to several competing digital standards), music storage (from vinyl records to CDs), and video storage (from VHS or beta tapes to DVDs). However, we typically consume information in analog form; for example, reading a book or a computer screen, listening to a conversation or to music. Why, then, is the world going digital? We consider this issue after first discussing the components of a typical digital communication system.

CAMBRIDGE

Cambridge University Press
978-0-521-87414-4 - Fundamentals of Digital Communication
Upamanyu Madhow
Excerpt
More information

1.1 Components of a digital communication system

Consider the block diagram of a digital communication link depicted in Figure 1.1. Let us now briefly discuss the roles of the blocks shown in the figure.

Source encoder Information theory tells us that any information can be efficiently represented in digital form up to arbitrary precision, with the number of bits required for the representation depending on the required fidelity. The task of the source encoder is to accomplish this in a practical setting, reducing the redundancy in the original information in a manner that takes into account the end user's requirements. For example, voice can be intelligibly encoded into a 4 kbit/s bitstream for severely bandwidth constrained settings, or sent at 64 kbit/s for conventional wireline telephony. Similarly, audio encoding rates have a wide range – MP3 players for consumer applications may employ typical bit rates of 128 kbit/s, while high-end digital audio studio equipment may require around ten times higher bit rates. While the preceding examples refer to lossy source coding (in which a controlled amount of information is discarded), lossless compression of data files can also lead to substantial reductions in the amount of data to be transmitted.

Channel encoder and modulator While the source encoder eliminates unwanted redundancy in the information to be sent, the channel encoder introduces redundancy in a controlled fashion in order to combat errors that may arise from channel imperfections and noise. The output of the channel encoder is a codeword from a channel code, which is designed specifically for the anticipated channel characteristics and the requirements dictated by higher network layers. For example, for applications that are delay insensitive, the channel code may be optimized for error detection, followed by a request for retransmission. On the other hand, for real-time applications for which retransmissions are not possible, the channel code may be optimized for error correction. Often, a combination of error correction and detection may be employed. The modulator translates the discrete symbols output by the channel code into an analog waveform that can be transmitted over the

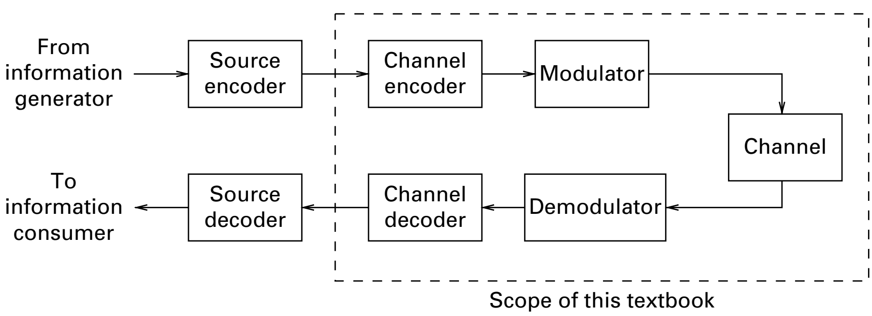

Figure 1.1 Block diagram of a digital communication link.

physical channel. The physical channel for an 802.11b based wireless local area network link is, for example, a band of 20 MHz width at a frequency of approximately 2.4 GHz. For this example, the modulator translates a bitstream of rate 1, 2, 5.5, or 11 Mbit/s (the rate varies, depending on the channel conditions) into a waveform that fits within the specified 20 MHz frequency band.

Channel The physical characteristics of communication channels can vary widely, and good channel models are critical to the design of efficient communication systems. While receiver thermal noise is an impairment common to most communication systems, the channel distorts the transmitted waveform in a manner that may differ significantly in different settings. For wireline communication, the channel is well modeled as a linear time-invariant system, and the transfer function in the band used by the modulator can often be assumed to be known at the transmitter, based on feedback obtained from the receiver at the link set-up phase. For example, in high-speed digital subscriber line (DSL) systems over twisted pairs, such channel feedback is exploited to send more information at frequencies at which the channel gain is larger. On the other hand, for wireless mobile communication, the channel may vary because of relative mobility between the transmitter and receiver, which affects both transmitter design (accurate channel feedback is typically not available) and receiver design (the channel must either be estimated, or methods that do not require accurate channel estimates must be used). Further, since wireless is a broadcast medium, multiple-access interference due to simultaneous transmissions must be avoided either by appropriate resource sharing mechanisms, or by designing signaling waveforms and receivers to provide robust performance in the presence of interference.

Demodulator and channel decoder The demodulator processes the analog received waveform, which is a distorted and noisy version of the transmitted waveform. One of its key tasks is synchronization: the demodulator must account for the fact that the channel can produce phase, frequency, and time shifts, and that the clocks and oscillators at the transmitter and receiver are not synchronized a priori. Another task may be channel equalization, or compensation of the intersymbol interference induced by a dispersive channel. The ultimate goal of the demodulator is to produce tentative decisions on the transmitted symbols to be fed to the channel decoder. These decisions may be "hard" (e.g., the demodulator guesses that a particular bit is 0 or 1), or "soft" (e.g., the demodulator estimates the likelihood of a particular bit being 0 or 1). The channel decoder then exploits the redundancy in the channel to code to improve upon the estimates from the demodulator, with its final goal being to produce an estimate of the sequence of information symbols that were the input to the channel encoder. While the demodulator and decoder operate independently in traditional receiver designs, recent advances in coding and communication theory show that iterative information exchange between the demodulator and the decoder can dramatically improve performance.

Source decoder The source decoder converts the estimated information bits produced by the channel decoder into a format that can be used by the end user. This may or may not be the same as the original format that was the input to the source encoder. For example, the original source encoder could have translated speech into text, and then encoded it into bits, and the source decoder may then display the text to the end user, rather than trying to reproduce the original speech.

We are now ready to consider why the world is going digital. The two key advantages of the digital communication approach to the design of transmission and storage media are as follows:

Source-independent design Once information is transformed into bits by the source encoder, it can be stored or transmitted without interpretation: as long as the bits are recovered, the information they represent can be reconstructed with the same degree of precision as originally encoded. This means that the storage or communication medium can be independent of the source characteristics, so that a variety of information sources can share the same communication medium. This leads to significant economies of scale in the design of individual communication links as well as communication networks comprising many links, such as the Internet. Indeed, when information has to traverse multiple communication links in a network, the source encoding and decoding in Figure 1.1 would typically be done at the end points alone, with the network transporting the information bits put out by the source encoder without interpretation.

Channel-optimized design For each communication link, the channel encoder or decoder and modulator or demodulator can be optimized for the specific channel characteristics. Since the bits being transported are regenerated at each link, there is no "noise accumulation."

The preceding framework is based on a separation of source coding and channel coding. Not only does this *separation principle* yield practical advantages as mentioned above, but we are also reassured by the source–channel separation theorem of information theory that it is theoretically optimal for point-to-point links (under mild conditions). While the separation approach is critical to obtaining the economies of scale driving the growth of digital communication systems, we note in passing that joint source and channel coding can yield superior performance, both in theory and practice, in certain settings (e.g., multiple-access and broadcast channels, or applications with delay or complexity constraints).

The scope of this textbook is indicated in Figure 1.1: we consider modulation and demodulation, channel encoding and decoding, and channel modeling.

Source encoding and decoding are not covered. Thus, we implicitly restrict attention to communication systems based on the separation principle.

1.2 Text outline

The objective of this text is to convey an understanding of the principles underlying the design of a modern digital communication link. An introduction to modulation techniques (i.e., how to convert bits into a form that can be sent over a channel) is provided in Chapter 2. We emphasize the important role played by the complex baseband representation for passband signals in both transmitter and receiver design, describe some common modulation formats, and discuss how to determine how much bandwidth is required to support a given modulation format. An introduction to demodulation (i.e., how to estimate the transmitted bits from a noisy received signal) for the classical additive white Gaussian noise (AWGN) channel is provided in Chapter 3. Our starting point is the theory of hypothesis testing. We emphasize the geometric view of demodulation first popularized by the classic text of Wozencraft and Jacobs, introduce the concept of soft decisions, and provide a brief exposure to link budget analysis (which is used by system designers for determining parameters such as antenna gains and transmit powers). Mastery of Chapters 2 and 3 is a prerequisite for the remainder of this book. The remaining chapters essentially stand on their own. Chapter 4 contains a framework for estimation of parameters such as delay and phase, starting from the derivation of the likelihood ratio of a signal in AWGN. Optimal noncoherent receivers are derived based on this framework. Chapter 5 describes the key ideas used in channel equalization, including maximum likelihood sequence estimation (MLSE) using the Viterbi algorithm, linear equalization, and decision feedback equalization. Chapter 6 contains a brief treatment of information theory, focused on the *computation* of performance benchmarks. This is increasingly important for the communication system designer, now that turbo-like codes provide a framework for approaching information-theoretic limits for virtually any channel model. Chapter 7 introduces error-correction coding. It includes convolutional codes, serial and parallel concatenated turbo codes, and low density parity check (LDPC) codes. It also provides a very brief discussion of how algebraic codes (which are covered in depth in coding theory texts) fit within modern communication link design, with an emphasis on Reed–Solomon codes. Finally, Chapter 8 contains an introduction to wireless communication, including channel modeling, the effect of fading, and a discussion of some modulation formats commonly used over the wireless channel that are not covered in the introductory treatment in Chapter 2. The latter include orthogonal frequency division multiplexing (OFDM), spread spectrum communication, continuous phase modulation, and space–time (or multiple antenna) communication.

1.3 Further reading

Useful resources for getting a quick exposure to many topics on communication systems are *The Communications Handbook* [1] and *The Mobile Communications Handbook* [2], both edited by Gibson. Standards for communication systems are typically available online from organizations such as the Institute for Electrical and Electronics Engineers (IEEE). Recently published graduate-level textbooks on digital communication include Proakis [3], Benedetto and Biglieri [4], and Barry, Lee, and Messerschmitt [5]. Undergraduate texts on communications include Haykin [6], Proakis and Salehi [7], Pursley [8], and Ziemer and Tranter [9]. Classical texts of enduring value include Wozencraft and Jacobs [10], which was perhaps the first textbook to introduce signal space design techniques, Viterbi [11], which provides detailed performance analysis of demodulation and synchronization techniques, Viterbi and Omura [12], which provides a rigorous treatment of modulation and coding, and Blahut [13], which provides an excellent perspective on the concepts underlying digital communication systems.

We do not cover source coding in this text. An information-theoretic treatment of source coding is provided in Cover and Thomas [14], while a more detailed description of compression algorithms is found in Sayood [15].

Finally, while this text deals with the design of individual communication links, the true value of these links comes from connecting them together to form communication networks, such as the Internet, the wireline phone network, and the wireless cellular communication network. Two useful texts on communication networks are Bertsekas and Gallager [16] and Walrand and Varaiya [17]. On a less technical note, Friedman [18] provides an interesting discussion on the immense impact of advances in communication networking on the global economy.

CHAPTER 2 Modulation

Modulation refers to the representation of digital information in terms of analog waveforms that can be transmitted over physical channels. A simple example is depicted in Figure 2.1, where a sequence of bits is translated into a waveform. The original information may be in the form of bits taking the values 0 and 1. These bits are translated into symbols using a bit-to-symbol map, which in this case could be as simple as mapping the bit 0 to the symbol $+1$, and the bit 1 to the symbol -1. These symbols are then mapped to an analog waveform by multiplying with translates of a transmit waveform (a rectangular pulse in the example shown): this is an example of *linear modulation*, to be discussed in detail in Section 2.5. For the bit-to-symbol map just described, the bitstream encoded into the analog waveform shown in Figure 2.1 is 01100010100.

While a rectangular timelimited transmit waveform is shown in the example of Figure 2.1, in practice, the analog waveforms employed for modulation are often constrained in the frequency domain. Such constraints arise either from the physical characteristics of the communication medium, or from external factors such as government regulation of spectrum usage. Thus, we typically classify channels, and the signals transmitted over them, in terms of the frequency bands they occupy. In this chapter, we discuss some important modulation techniques, after first reviewing some basic concepts regarding frequency domain characterization of signals and systems. The material in this chapter is often covered in detail in introductory digital communication texts,

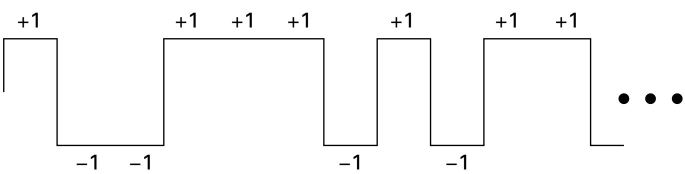

Figure 2.1 A simple example of binary modulation.

but we emphasize some specific points in somewhat more detail than usual. One of these is the complex baseband representation of passband signals, which is a crucial tool both for understanding and implementing modern communication systems. Thus, the reader who is familiar with this material is still encouraged to skim through this chapter.

Map of this chapter In Section 2.1, we review basic notions such as the frequency domain representation of signals, inner products between signals, and the concept of baseband and passband signals. While currents and voltages in a circuit are always real-valued, both baseband and passband signals can be treated under a unified framework by allowing baseband signals to take on complex values. This complex baseband representation of passband signals is developed in Section 2.2, where we point out that manipulation of complex baseband signals is an essential component of modern transceivers. While the preceding development is for deterministic, finite energy signals, modeling of signals and noise in digital communication relies heavily on finite power, random processes. We therefore discuss frequency domain description of random processes in Section 2.3. This completes the background needed to discuss the main theme of this chapter: modulation. Section 2.4 briefly discusses the degrees of freedom available for modulation, and introduces the concept of bandwidth efficiency. Section 2.5 covers linear modulation using two-dimensional constellations, which, in principle, can utilize all available degrees of freedom in a bandlimited channel. The Nyquist criterion for avoidance of intersymbol interference (ISI) is discussed, in order to establish guidelines relating bandwidth to bit rate. Section 2.6 discusses orthogonal and biorthogonal modulation, which are nonlinear modulation formats optimized for power efficiency. Finally, Section 2.7 discusses differential modulation as a means of combating phase uncertainty. This concludes our introduction to modulation. Several other modulation formats are discussed in Chapter 8, where we describe some modulation techniques commonly employed in wireless communication.

2.1 Preliminaries

This section contains a description of just enough material on signals and systems for our purpose in this text, including the definitions of inner product, norm and energy for signals, convolution, Fourier transform, and baseband and passband signals.

Complex numbers A complex number z can be written as $z = x + \mathrm{j}y$, where x and y are real numbers, and $\mathrm{j} = \sqrt{-1}$. 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 2.2, it is often advantageous to interpret the complex number z as

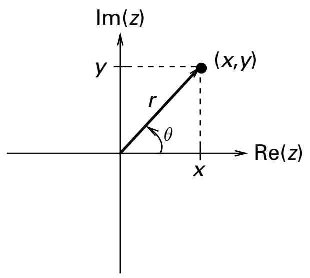

Figure 2.2 A complex number z represented in the two-dimensional real plane.

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 as

$$r = |z| = \sqrt{x^2 + y^2},$$
$$\theta = \arg(z) = \tan^{-1} \frac{y}{x}.$$

Euler's identity We routinely employ this to decompose a complex exponential into real-valued sinusoids as follows:

$$\mathrm{e}^{\mathrm{j}\theta} = \cos\theta + \mathrm{j}\sin\theta. \tag{2.1}$$

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 $\mathbf{s} = (s[1], \ldots, s[m])^T$ and $\mathbf{r} = (r[1], \ldots, r[m])^T$ is given by

$$\langle \mathbf{s}, \mathbf{r} \rangle = \sum_{i=1}^{m} s[i] r^*[i] = \mathbf{r}^H \mathbf{s}. \tag{2.2}$$

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

$$\langle s, r \rangle = \int_{-\infty}^{\infty} s(t) r^*(t)\, \mathrm{d}t. \tag{2.3}$$

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, a_2 are complex-valued constants, and s, s_1, s_2, r, r_1, 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 remain aware of when computing inner products for complex signals.

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

$$E_s = ||s||^2 = \langle s, s \rangle = \int_{-\infty}^{\infty} |s(t)|^2 \mathrm{d}t, \tag{2.4}$$

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.

Cauchy–Schwartz inequality The inner product obeys the *Cauchy–Schwartz* inequality, stated as follows:

$$|\langle s, r \rangle| \le ||s|| \, ||r||, \tag{2.5}$$

with equality if and only if, for some complex constant a, $s(t) = ar(t)$ or $r(t) = as(t)$ almost everywhere. That is, equality occurs if and only if one signal is a scalar multiple of the other. The proof of this inequality is given in Problem 2.4.

Convolution The convolution of two signals s and r gives the signal

$$q(t) = (s * r)(t) = \int_{-\infty}^{\infty} s(u) r(t-u) \mathrm{d}u.$$

Here, the convolution is evaluated at time t, while u is a "dummy" variable that is integrated out. However, it is sometimes convenient to abuse notation and use $q(t) = s(t) * r(t)$ to denote the convolution between s and r. For example, this enables us to state compactly 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.

Delta function The delta function $\delta(t)$ is defined via the following "sifting" property: for any finite energy signal $s(t)$, we have

$$\int_{-\infty}^{\infty} \delta(t-t_0) s(t) \mathrm{d}t = s(t_0). \tag{2.6}$$

In particular, this implies that convolution of a signal with a shifted version of the delta function gives a shifted version of the signal:

$$\delta(t-t_0) * s(t) = s(t-t_0). \tag{2.7}$$

Equation (2.6) can be shown to imply that $\delta(0) = \infty$ and $\delta(t) = 0$ for $t \neq 0$. Thus, thinking of the delta function as a signal is a convenient abstraction, since it is not physically realizable.