ERROR-CONTROL TECHNIQUES FOR DIGITAL COMMUNICATION

ARNOLD M. MICHELSON
ALLEN H. LEVESQUE

GTE Government Systems Corporation
Needham Heights, Massachusetts

A Wiley-Interscience Publication
JOHN WILEY & SONS
New York Chichester Brisbane Toronto Singapore

Library of Congress Cataloging in Publication Data:

Michelson, Arnold M.
Error-control techniques for digital communication.
"A Wiley-Interscience publication."
Bibliography: p.
Includes index.
1. Error-correcting codes (Information theory)
2. Digital communications. I. Levesque, Allen H.
II. Title.
TK5103.7.M53 1984 621.38 84-15327
ISBN 0-471-88074-4

Printed in the United States of America

10 9 8 7 6 5 4 3 2

CONTENTS

CHAPT

CHAPT

CHAPTER

CHAPTEI

CHAPTER ONE

Reliable Transmission of Digital Information

The purpose of this book is to provide the communication systems engineer with a basic understanding of error-control coding techniques and the role that coding plays in the design of efficient digital communication systems. Described in its simplest terms, error-control coding involves the addition of redundancy to transmitted data so as to provide the means for detecting and correcting errors that inevitably occur in any real communication process. Thus coding can be used to provide a desired level of accuracy in the digital data delivered to a user. There are, however, other ways to achieve accurate transmission of digital data, and this book is intended to aid the communication system designer in deciding when it makes sense to use coding and when it does not, in choosing a coding technique appropriate to the application and performance requirements at hand, and in evaluating the performance achievable with the chosen technique.

For example, in many communication systems, an alternative to the use of coding is simply to provide sufficient signal energy per unit of information to ensure that uncoded information is delivered with the required accuracy. The energy needed might be provided by setting signal power at a sufficiently high level or, if power limitations prevail, by using some form of diversity transmission and reception. However, in many cases, error-control coding can provide the required accuracy with less energy than uncoded operation and may be the economically preferred solution in spite of an increase in system complexity. Cost savings through the use of coding techniques can be dramatic when very high accuracy is needed and power is expensive. Furthermore, in some applications the savings in signal power are accompanied by important reductions in size and weight of the communication equipment.

The levels of performance that can ultimately be achieved with coded communication systems are given by the remarkable theorems of Claude Shannon, who in 1948 laid the foundation of the science of *information theory* in a famous paper entitled "A Mathematical Theory of Communication" [157]. The basic theorems of information theory not only mark the limits of efficiency in communication performance but also define the role that coding plays in achieving these limits. That is, digital codes are shown to be an efficient way of constructing the waveforms to be transmitted in order to achieve optimum communication performance for some applications.

Shannon's 1948 paper presented a statistical formulation of the communication problem, unifying earlier work by Hartley [61], Wiener [178], Rice [148], and Kotel'nikov [86]. Shannon's work sharply contradicted the long-standing intuitive but erroneous notion that noise places an inescapable limitation on the accuracy of communication. Shannon proved that the characteristics of a communication channel, namely the noise level, bandwidth, and signal power, can be used to derive a parameter C, called *channel capacity*, that gives the upper limit on the rate at which information can be transmitted through the channel and received reliably. Shannon's results showed that as long as the information transmission rate is below C, the probability of error in the information delivered can in principle be made arbitrarily low by using sufficiently long coded transmission signals. Thus noise limits the achievable information communication rate, but not the achievable accuracy. Much of the research in communication theory since the appearance of Shannon's early work has been concerned with extending and refining his basic results and with finding ways of approaching the full realization of these results in practical communication system designs. The development of error-control coding techniques has been a central element in this research.

In this book we present the most important of the error-control coding techniques that have been developed since Shannon's pioneering work. That is, we consider those techniques that have actually been used effectively in real communication systems. In this introductory chapter we begin with a description of the key elements of a modern digital communication system as well as the channel models that are used throughout the book. A heuristic discussion of information theory follows, concluding with a presentation of the key result, the channel coding theorem. It is not necessary to have a detailed understanding of information theory in order to make effective use of error-control coding techniques. However, a familiarity with the underlying principles and the meaning of the channel coding theorem is important. The fundamental limit on the efficiency of a digital communication system is given by the channel coding theorem, and this provides the gauge for measuring the overall efficiency of any given system design.

We then review the basic digital modulation and demodulation techniques. Performance curves are included that show that even the best of the practical signaling schemes fall far short of the performance limit given by the channel coding theorem. It will be seen in subsequent chapters that judicious choice of

modulation and coding techniques can, in many applications, provide significant improvements in communication efficiency. The chapter concludes with a discussion of the proper way to design a digital communication system. Specifically, an analytical approach is given that permits joint optimization of the key communication functions.

1.1. THE COMMUNICATION SYSTEM DESIGN PROBLEM

We can state the task of the digital communication system designer as that of providing a cost-effective system for transmitting information from a sender to a user at the rate and level of accuracy that the user requires. The key parameters of the design are transmission bandwidth, signal power, and the complexity of the implementation chosen. The information transmission rate and accuracy of the delivered information are typically determined by the user's requirements.

The transmission bandwidth is often constrained by factors specific to the particular transmission medium used. For example, telephone circuits are separated into nominal 3-kHz bandwidth segments by longstanding engineering practice in the telephone industry. Similarly, there are standard bandwidths for individual channels on terrestrial radio circuits and satellite links due to established government regulations on spectrum utilization. In other cases, however, bandwidth constraints are not a critical issue. Examples include links to and from vehicles in deep space, where wide transmission bandwidths for a few individual links can be chosen freely without concern about possible interference with other users of the spectrum.

Finally, signal power and implementation complexity are system characteristics that are usually very much under the designer's control, and possible trade-offs between power and complexity are central issues in the design task. Both characteristics represent cost factors for the designer to consider. For example, in most systems a desired level of accuracy in the information delivered can be achieved by supplying enough power in the transmitted signal to overcome channel disturbances that produce errors. An alternative to increasing signal power is to add systematic redundancy to the transmitted information in the form of error-control coding. However, the use of coding adds complexity to the system, particularly for the implementation of the decoding operations. Since the addition of redundancy also implies the need to increase transmission bandwidth, the design trade-off must include considerations of bandwidth. In fact, in applications where bandwidth is strictly limited or very costly and when we are not permitted to lengthen message transmission time, it is difficult to use coding effectively as a means of improving information accuracy, and increasing signal power may be the only means available. These issues will be discussed in detail later in this chapter, when the fundamental results in information theory are reviewed and the implications for the design of efficient communication systems are presented.

It is clear from the outset that with any real communication system we cannot expect to receive exactly what is transmitted. At the very least, we can expect noise to be added to the transmission, causing random errors. As was stated earlier, the work of Shannon showed that channel noise limits the rate at which we can communicate reliably but not the achievable level of accuracy. The purpose of this introductory chapter is to present and illuminate this important result. It is necessary first to review the key elements of a digital communication system and the limitations imposed by the physical world. Then the role that error-control coding plays in the design of an energy-efficient communication system can be discussed in detail.

1.2. ELEMENTS OF A DIGITAL COMMUNICATION SYSTEM

The basic elements of a one-way communication system are illustrated with the general block diagram shown in Fig. 1.1. We now examine each of these elements in detail.

1.2.1. Information Source

The source of information to be transmitted may be either analog or discrete. An *analog source* produces time-continuous signals, while a *discrete source* produces sequences of discrete symbols. Examples of signals from an analog source are speech signals, radar outputs, and photographic scan data. Typical discrete sources are computer data files and messages generated at teleprinter terminals.

For analog sources, techniques are needed to efficiently represent the analog signals by sequences of digital symbols. Ordinarily this is done simply with a sampler and analog-to-digital converter. Ideally, we would like to represent the source information by the smallest number of digital samples to make the most efficient use of the digital communication system. In other words, we would like to remove as much redundancy as possible from the source information prior to transmission. Techniques for converting arbitrary information sources into digital messages efficiently are described broadly as *source coding*

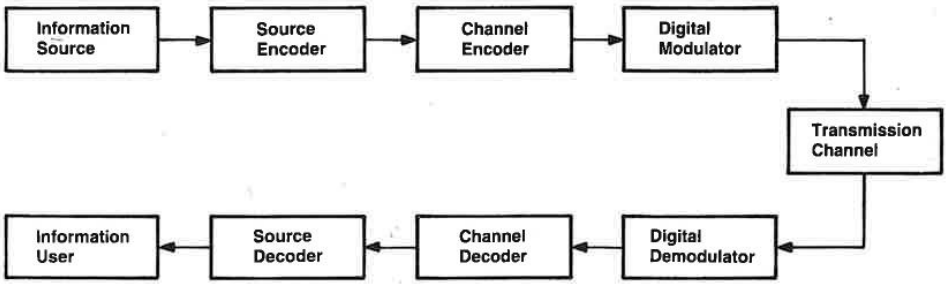

FIGURE 1.1. Block diagram of a digital communication system.

techniques. There is now a considerable body of literature on the subject of source coding, dealing with the theoretical limits on performance as well as with some very effective techniques for coding specific types of sources. To treat this subject in depth is outside the scope of this book, and we refer the interested reader to the literature on source coding [6, 175].

For our purposes in this book, we assume we are given a digital information source and our job is to provide an effective and efficient system for communicating the information it produces from one place to another. Further, we assume that the source generates a stream of statistically independent, equally likely discrete symbols. In particular, we assume that it produces binary digits, which we call *bits*, each bit occurring with equally likely binary values independent of all other bits generated. We shall say that the source produces data at a constant rate of R_s bits per second, which is defined as the *information rate of the source*. The purpose of the communication system is to deliver the source data to a user at the rate R_s and at some required level of accuracy, which is typically stated as an upper limit on acceptable *probability of bit error* or *bit-error rate* in the delivered information. It will be seen in a later discussion that the overall efficiency of the communication system is directly related to the amount of signal energy needed to deliver each information bit with the required accuracy. We denote this amount of energy as E_b and call it the required *energy per information bit*. The required signal power S is therefore given by $S = E_b R_s$.

1.2.2. Channel Encoder

The channel encoder performs all the digital operations needed to prepare the source data for modulation. We define the encoder here in a general way to encompass a variety of possible operations. In the simplest case, where no redundancy is to be added and the transmission in the physical channel is to use a binary signaling alphabet, the encoder has no function. If no redundancy is to be used but the transmission alphabet is to be nonbinary, the encoder performs the necessary binary-to-nonbinary symbol conversion. For example, if an 8-ary signaling alphabet is to be used, the encoder accepts source bits in successive blocks of three bits each and produces 8-ary symbols at a rate that is one-third of R_s.

If binary error-control coding is to be used, the encoder accepts information bits at the rate R_s and adds redundancy, producing encoded data at a higher rate R_c. For encoding with a *block code*, the encoder accepts information in successive k-bit blocks and for each k bits generates a block of n bits, where $n \geq k$. The n-bit block is called a *code block* or a *codeword*. Thus the encoder releases bits at a rate $R_c = R_s(n/k)$. We define the dimensionless ratio $R = k/n$ as the rate of the code, or simply the *code rate*. Note that we let n have values $n \geq k$ rather than simply $n > k$, to include uncoded operation as a special case. If the error-control coding is nonbinary, say M-ary where $M = 2^m$ and m is an integer greater than 1, the encoder accepts information bits in

blocks of km bits each and produces encoded blocks of n M-ary symbols each. Again, the code rate is $R = k/n$.

For encoding with a *convolutional code*, the encoder accepts information bits as a continuous stream and, for a binary code, generates a continuous stream of encoded bits at a higher rate. The information stream is fed to the encoder b bits at a time, where b is typically in the range 1 to 6. The encoder operates on the current b-bit input and some number of immediately preceding b-bit inputs to produce V output bits, with $V > b$. Thus the code rate is $R = b/V$. The number of successive b-bit groups of information bits over which each encoding step operates is called the *constraint length* of the code, which we also denote by k. The encoder for a convolutional code might be thought of as a form of digital filter with memory extending $k - 1$ symbols into the past. A typical binary convolutional code will have $b = 1$, $V = 2$ or 3, and k in the range 4 to 7, although in special applications constraint lengths in the range 30 to 70 might be used.

To use a convolutional code with nonbinary transmission, each b-bit input to the encoder results in the generation of V M-ary coded symbols, where usually $M = 2^m$, and $mV > b$. A typical rate-1/2 encoder for a 16-ary ($m = 4$) transmission alphabet might have $b = 4$, $V = 2$, and $k = 2$ (four-bit symbols).

We shall not delve any further into the details of code design at this point, our immediate purpose having been served by the introduction of the concepts of code rate, block length, and constraint length. As we shall see in later discussions, these are the key parameters of a code design, since the reciprocal of the code rate gives us a measure of the required bandwidth expansion and the code block length or constraint length and rate provide a measure of the complexity of the required encoding and (more important) decoding operations. In Section 1.4 we shall see that much can be said about the communication performance achievable with well-designed codes by dealing with only these design parameters.

1.2.3. Digital Modulator

The function of the modulator is to match the encoder output to the transmission channel. The modulator accepts binary or M-ary encoded symbols and produces waveforms appropriate to the physical transmission medium, which is always analog. In many systems where coding is to be applied, the modulation and demodulation techniques and equipment are difficult or impossible to modify or replace. In other cases, the modulation technique is fixed, but changes in the method of demodulation are feasible. In yet other applications, it is possible to design the modulation and demodulation system along with the coding technique, and greatly increased latitude is provided for overall optimization of the design.

It has been conventional in much of the communication literature to define "the channel" as representing that portion of the communication system that the designer is unable or unwilling to change. In following this convention, if

the modulation and demodulation equipment (usually shortened to *modem* for convenience) is not available for modification, those functions would be incorporated into the definition of the channel. However, we prefer to depart from this convention to the extent that while we shall treat the modem as part of the channel for purposes of analysis, we ask the reader to bear in mind that the design of an efficient system is best done by designing the modem functions in conjunction with the encoding and decoding functions. Section 1.6 addresses this point in more detail.

For *binary modulation*, the modulator simply converts a binary digit, 0 or 1, to a waveform, say $s_0(t)$ or $s_1(t)$, respectively, of equal duration T_s. For *M-ary modulation*, the M possible encoded symbols are converted to a corresponding set of M waveforms $s_0(t), s_1(t), \ldots, s_{M-1}(t)$. It is assumed that the reader is familiar with the common forms of digital modulation. For binary signaling, conventional modulation types include phase shift keying (PSK), differentially encoded PSK (DPSK), and frequency shift keying (FSK). Nonbinary forms of these basic modulation types are M-ary PSK (MPSK), M-ary DPSK (MDPSK), and M-ary FSK (MFSK). With the conventional forms of these modulation types, the nominal bandwidth of each waveform $s_i(t), i = 0, 1, 2, \ldots, M-1$, is approximately $1/T_s$. However, for *spread spectrum* signaling, as is implied by the name, the bandwidth of each waveform can be much wider than $1/T_s$, perhaps by as much as several orders of magnitude. For example, a spread spectrum version of binary PSK might utilize waveforms $s_0(t)$ and $s_1(t)$ in which $s_0(t)$ is a sequence of much shorter binary PSK pulses, usually called *chips*, and $s_1(t)$ is the complement of the chip sequence in $s_0(t)$. Spread spectrum signaling is used as a *multiple-access* technique and also as a means of protecting a communication system against *jamming*. For further discussion of spread spectrum systems, the reader can refer to a special *IEEE Transactions* issue [70] as well as books on the subject [31, 68].

We shall return to the modulation and demodulation functions in Section 1.5.

1.2.4. Transmission Channel

We include in the term *transmission channel* all the operations required to prepare the baseband (low-pass) modulated waveforms for transmission in the physical channel, the transmission medium itself, and the receiving operations required to bring the signals to the point just prior to demodulation. In this way, we incorporate into the transmission channel any practical limitations or impairments in the equipment. As a practical matter we are primarily concerned here with power and bandwidth limitations, which are reflected in the design of the transmitting and receiving equipment.

Transmitted signal power provides the obvious means in many systems for providing a required level of accuracy in received information. However, signal power cannot be increased arbitrarily. In telephone networks, for example, signal levels are fixed by established industry standards. In radio

communication systems, there is more freedom in selecting power levels, but practical physical and economic limitations apply. An increase in power level invariably implies increases in size, weight, and cost of transmitting equipment. Even if the added cost is judged acceptable, there are some applications where mobility is very important, and strict limitations on the size and weight of the equipment must be complied with. In yet other applications, particularly in very low regions of the radio spectrum, enormous amounts of energy are required to radiate usable signals, and therefore transmitted signal power is a dominant factor in the cost of the system.

Bandwidth is the other critical design parameter governing achievable performance, since it limits the rate at which we can modulate waveforms in the channel. Restrictions on the choice of bandwidth are more or less severe depending upon the transmission medium. In telephone systems and many radio systems, where strict channelization standards are in place, bandwidth can be provided only in fixed increments, such as 3 or 6 kHz. Therefore, providing increased bandwidth in turn implies leasing additional wireline channels or acquiring added radio channel allocations and transmission equipment, the latter typically designed with standard bandwidths. In some parts of the radio spectrum, crowding is a serious problem, and it is difficult to use more bandwidth without encountering significant levels of interference with other communication signals.

Noise in received signals constitutes the most prevalent factor limiting the performance of a communication system, since noise limits the ability of the demodulator to reliably distinguish one modulated waveform from another, thereby producing errors in the demodulator output. Thermal noise is always present in electrical circuitry, for example, in the front end of the receiving equipment. However, receiver noise is not necessarily the primary concern, since in many parts of the radio spectrum other sources of noise are also significant. For example, atmospheric impulse noise due to lightning discharges can be of a sufficiently high level to be the dominant factor limiting performance. Atmospheric noise can also affect wireline communication systems, since considerable energy can couple into transmission lines during thunderstorms. Additional forms of impulsive noise affect telephone networks, arising from transients in switching equipment as well as from accidental circuit interruptions during maintenance. What distinguishes impulsive noise in its various forms from the ever-present thermal noise is a distinct difference in temporal characteristics as manifested in the ways digital errors occur. Thermal noise is broadband, is essentially steady in its power level, and has Gaussian amplitude statistics. Therefore, errors tend to occur independently from one signaling interval to the next, the rate of occurrence being derivable in a straightforward way from knowledge of the Gaussian distribution. Impulsive noise, on the other hand, is characterized by relatively long quiet intervals punctuated by short periods of intense noise. This characteristic in turn results in long error-free intervals interspersed with short *bursts* of errors.

If one takes detailed account of the different characteristics of Gaussian noise channels and burst-error channels, one is led to quite different coding techniques for each. However, what is usually done is to select the modulation and coding schemes with a view toward the limitations imposed by the Gaussian background noise and then to adapt certain features of the coding implementation to the particular characteristics of the burst-error phenomena if they are of concern for the application at hand. We shall say more on this point in Chapter 11.

Although we shall be concerned primarily with additive noise, many real communication channels exhibit other phenomena that severely limit communication performance. For example, many channels contain a sufficient amount of time dispersion that the received symbols flow into one another. This effect is commonly called *intersymbol interference*. In wireline channels, the time dispersion arises largely from significant nonlinearity of the phase characteristic within the channel bandwidth.

Intersymbol interference from any cause is a form of "self-noise" in a digital communication system, which cannot be overcome by increasing signal power. For this reason, channels affected in this way exhibit an irreducible error rate at high levels of signal-to-noise ratio (SNR), which cannot be avoided except by implementing some scheme for dealing directly with the time dispersion. Error-control coding usually does not provide an attractive approach, since intersymbol interference comes about by trying to signal as rapidly as possible within a given transmission bandwidth. It is implicit that the added bandwidth that coding would require cannot be readily provided. For this reason, intersymbol interference is treated by other techniques, for example, on telephone channels by *adaptive equalization* [98, 139].

On many radio channels, time dispersion is due directly to the *multipath* nature of signal propagation at extended distances. For example, in the high frequency (HF) band, 3 to 30 MHz, communication beyond the horizon is accomplished by refraction of signals at various layers of the ionosphere. The structure of the ionosphere causes a transmitted HF signal to arrive at a distant receiving site by a multiplicity of *propagation modes*, the modes in general having different path delays.

A direct consequence of multipath on radio channels is the phenomenon known as signal fading or simply *fading*. This comes about in ionospheric propagation due to the fact that the state of the ionosphere is dynamic. Ions within each layer are constantly in motion, and the layers also move relative to one another. As a result, the summation of several fluctuating modal components in a received multipath signal, with component signal phases randomly sliding in and out of alignment, produces the random fluctuations in received signal amplitude called fading. Associated with the amplitude variations in fading are fluctuations in the instantaneous phase of the received signal. Thus the multipath structure directly accounts for time dispersion, and the time-varying nature of the multipath accounts for fading. Radio channels

that behave in this way are often called *fading multipath channels*. In addition to the HF frequency band already mentioned, examples include the VHF and UHF frequency bands (30 to 300 MHz and 300 to 3000 MHz), when used for beyond-the-horizon terrestrial communication by ionospheric and tropospheric scatter propagation, and the SHF band (3000 to 30,000 MHz), which is used for satellite communication.

There has been some success in applying adaptive equalization techniques to radio channels [3, 114]. However, the fading process greatly complicates the use of equalization, since, unlike having to adjust to a static phase characteristic, the equalizer has to accurately keep up with rapid continuous changes in signal amplitude and phase. For this reason, rather complex algorithms are used in applying these techniques to radio channels, and much of this work must still be regarded as developmental.

Where fading cannot be treated effectively by adaptive equalization, the resulting error characteristics can be dealt with by the use of efficient error-control coding techniques. On fading channels, the received errors tend to be clustered in bursts that occur in the intervals when the signal attenuation is large, that is, when the channel is going through deep fades. Since the propagation variations that produce fading are random, the durations of the error bursts and the intervening intervals of relatively error-free data are themselves random. However, the statistical parameters of the error clustering behavior can often be predicted within broad limits, given the operating frequency of a radio system and certain details of the transmission path. Some attention is given in Chapter 11 to the problem of applying error-control coding to burst-error channels.

1.2.5. Digital Demodulator

At the receiving end of the communication link, the demodulator provides the interface between the transmission channel and the functions that compute and deliver estimates of the transmitted data to the user. We include the rf receiving equipment in our definition of the transmission channel. The demodulator operates on the waveform received in each separate transmission symbol interval and produces a number or a set of numbers that represent an estimate of a transmitted binary or M-ary symbol. In some applications, the designer may choose the level of precision of this estimate.

In the simplest cases, the demodulator is designed to make a definite decision for each received symbol, that is, 0 or 1 for binary transmission or one of $0, 1, \ldots, M-1$ for M-ary transmission. It is convenient to refer to such cases as *hard-decision demodulation*. Since the transmitted waveforms have been corrupted by the various nonidealities of the transmission channel, the symbol decisions are subject to error, and the average rate of occurrence of symbol errors, taken as a fraction of the total number of symbols received over a long period of time, is called the *symbol-error rate* or *probability of symbol error*. For binary transmission, this is the *bit-error rate* or *probability of bit*

error. It is also conventional to apply the same bit-error terminology after conversion of M-ary symbol decisions to their binary representations. In a system where no error-control coding is used, these error rates are the error rates for the data delivered to the user. In a coded system the error rate at this point in the system is often called the *raw-channel error rate* or the *uncoded error rate* to make a distinction from the error statistics measured after decoding is performed.

All real transmission channels are, of course, analog channels that deliver waveforms that can in principle vary continuously over some range limited only by nonlinearities in the transmission medium and the receiving equipment. Thus the demodulator can be viewed as a form of waveform filtering followed by quantization, say to Q levels. The case of hard-decision binary demodulation thus requires quantization to $Q = 2$ levels. If the output of the binary demodulator is quantized to $Q > 2$ levels, we refer to this as *soft-decision demodulation*. In the limiting case of $Q = \infty$, the demodulator output is unquantized, corresponding to an analog matched filter output being delivered directly. For M-ary transmission, $Q > M$ constitutes soft-decision demodulation. Quantization incurs a loss of information, and thus soft-decision demodulation preserves information that can, we shall see in later discussions, be utilized profitably with an appropriate error-control decoding technique.

1.2.6. Channel Decoder

In a system using error-control coding, the decoder accomplishes the conversion of demodulator outputs into symbol decisions that reproduce, as accurately as possible, the data that was encoded by the channel encoder. For block coding, the decoder accepts consecutive blocks of n demodulator outputs and produces k decoded symbols for each block. With convolutional coding, the decoder accepts a steady stream of demodulator outputs and operates over the current received symbols and some number of previous symbols, producing b decoded outputs for each group of V received symbols. For both block and convolutional codes, the decoder attempts to make definite symbol decisions, binary or M-ary in accordance with the code design. However, the inputs need not be definite symbols.

Decoding techniques that operate on hard-decision demodulator outputs are commonly termed *hard-decision decoding* techniques. As will be seen in later chapters, they are essentially algebraic equation-solving algorithms. On the other hand, there are a number of decoding techniques, for both block and convolutional codes, that operate on soft-decision demodulator outputs and are collectively termed *soft-decision decoding* techniques. These techniques more nearly resemble signal correlation or matched-filtering operations than equation-solving routines. For many codes, practical soft-decision decoding algorithms are available that outperform the best hard-decision decoding algorithm for the same code. The range of potential performance advantage will depend to a great degree on characteristics of the transmission channel, the

performance margin generally being smallest on steady-signal Gaussian noise channels.

Error probability at the output of the decoder provides an important measure of the overall performance of the communication system. In fact, for convolutional codes, performance is usually stated in terms of post-decoding bit-error rate or M-ary symbol-error rate. However, there are other measures of communication performance that are meaningful. As an example, for block-coded systems, performance can conveniently be given in terms of the probability of correctly decoded blocks. It is also possible to state the performance of block-coded systems in terms of average post-decoding bit-error rate, where the error events are averaged over all decoded k-symbol blocks. For either convolutional or block codes, other measures of performance are sometimes used as well. For example, in a system designed to transmit and receive messages of a given length, if a message decoded with one or more errors is judged to be unacceptable, the appropriate measure of performance might well be the probability of receiving an error-free message.

Thus we see that in systems designed with error-control coding there is some flexibility in the way that performance is measured. It is important to note that there is no single figure of merit that can be used to realistically compare various coded and uncoded system alternatives for an arbitrary application, since requirements vary widely. The performance afforded by many error-control coding techniques is considered in detail in Chapters 7 and 9 through 11 for several measures of communication performance.

1.2.7. Source Decoder

The final stage of processing indicated in Fig. 1.1 is source decoding. The *source decoder* accepts the sequence of symbols from the channel decoder and, in accordance with the encoding method used, attempts to reproduce the information originally generated by the analog source. Generally speaking, the output of the source decoder is an approximation to the original source output, with discrepancies due to errors in channel decoding as well as loss of detail suffered in source encoding and perhaps in decoding as well. For some analog sources, the fidelity in reproduction of the source information can be measured by a simple statistic, for example, the mean-squared error between corresponding samples of the original source output and the output of the source decoder. For some sources, however, it is very difficult to find a reasonable mathematical measure of fidelity. A good example is that of speech signals, for which statistical characterization is known to be very difficult.

We shall not pursue source coding and decoding in great detail, since we wish to concern ourselves primarily with the problem of transmitting and accurately reproducing symbols generated by a digital information source. However, we do make one final point here on the subject of source coding. In a system that uses both source coding and error-control coding, the source coding operation can be viewed as one that removes the natural, perhaps

inefficient, redundancy from the source so that it can be replaced with highly structured and more efficient redundancy in the form of a well-chosen error-control code. This, in fact, is exactly what information theory requires us to do if we wish to achieve accurate transmission of the source information with the least expenditure of signal energy. These points will be made clearer in the remaining sections of this chapter.

1.3. IMPORTANT CHANNEL MODELS

A *model* of a communication system is a mathematical representation defined to realistically describe the way signals are constructed and processed and the way they are affected by the real-world communication environment. In Fig. 1.2 we show a simplified model of a digital communication system, in which the information to be transmitted is assumed to be generated by a discrete or digital information source. Using this model we next address the question of how to efficiently convey information to the user.

1.3.1. The Discrete-Time Channel

It is conventional to define a *channel model* to include the modulator, the demodulator, and all the intervening transmission equipment and media. This model, enclosed within dashed lines in Fig. 1.2, is compactly defined by the set of modulator inputs, the set of demodulator outputs, and the statistics that relate the possible outputs to each possible input. This is commonly called a *discrete-time channel model* or simply a *discrete channel*. The input-to-output statistics represent the ways in which the modulated signals are affected by amplitude and phase fluctuations, noise, interference, and equipment nonidealities and impairments. In most cases it is very difficult to define a

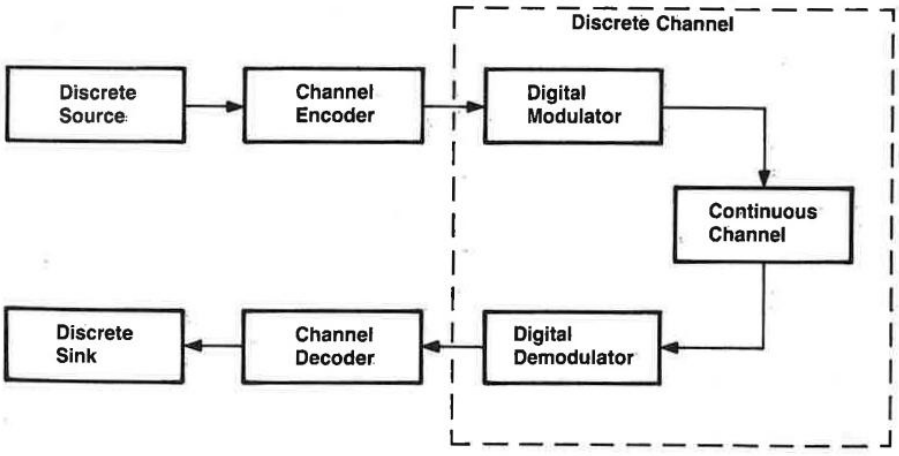

FIGURE 1.2. Model of a digital communication system.

model that thoroughly accounts for all the disturbances affecting the signals, and one must resort to reasonable approximations. However, experience has shown that even reasonably simple channel models can provide a sufficient degree of realism to enable proper design of efficient systems. Furthermore, simplified models often yield insights into underlying principles, which can be obscured by a myriad of details in more elaborate, though more accurate, models. We shall consider several channel models commonly used to analyze and design digital communication systems.

Most of the remaining discussion in this chapter will consider various forms of a channel model called the *discrete memoryless channel* (DMC), which is defined by an M-ary set of input symbols $\{x_i\}$, a Q-ary set of output symbols $\{y_j\}$, and a set of conditional probabilities, called *transition probabilities*, which we can write as

$$P(y = y_j | x = x_i) = P(y_j | x_i)$$

where $i = 0, 1, \ldots, M-1$, and $j = 0, 1, \ldots, Q-1$. The description of the channel as memoryless refers to the assumption that the output symbol at any instant of time depends statistically only on the input symbol at that time. The application of the DMC model to any real independent-error channel simply requires determination of the transition probabilities from definitions of the transmitted waveforms, signal power levels, and transmission channel characteristics, as well as the description of the demodulator. Examples of types of channels to which the DMC model does not apply are channels affected by atmospheric impulse noise or intersymbol interference.

1.3.2. The Binary Symmetric Channel

We now describe an important example of a DMC model. Suppose that binary modulation is used and that hard-decision demodulation is performed at the receiving end of the link. We let the modulator input x have value 0 or 1, and the demodulator output y have value 0 or 1. Let us now suppose that for either input value of x, and regardless of the transmitted or received values of any earlier bits, x is received in error with probability p or received correctly with probability $1 - p$. Using standard notation for conditional probabilities, we write this as

$$P(y = 1 | x = 0) = P(y = 0 | x = 1) = p$$

$$P(y = 0 | x = 0) = P(y = 1 | x = 1) = 1 - p$$

With these definitions we have combined the modulator, the transmission channel, and the demodulator into a compact binary-input, binary-output model depicted by the transition diagram in Fig. 1.3*a*. This simple channel model is known as the *binary symmetric channel*, usually abbreviated as BSC.

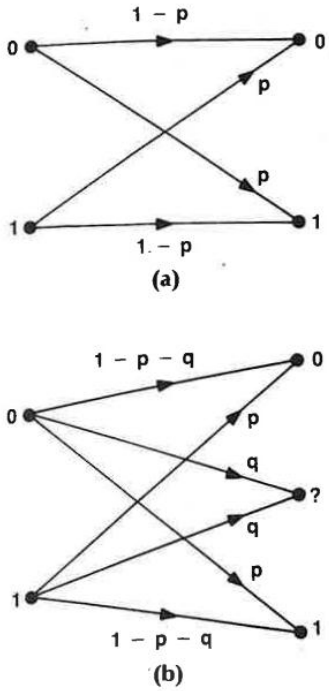

FIGURE 1.3. Discrete channel models. (*a*) Binary symmetric channel, (*b*) binary symmetric erasure channel.

1.3.3. The Binary Symmetric Erasure Channel

Another special case of the DMC is a binary-input, ternary-output channel called the *binary symmetric erasure channel* (BSEC). This channel model, depicted in Fig. 1.3*b*, includes a symmetric transition from either input symbol to an output symbol labeled ? to denote ambiguity. This model corresponds to a demodulation rule in which certain outputs are judged not to give a sufficiently reliable indication (e.g., due to a weak received signal) of which binary symbol was sent, and those outputs are erased as they leave the demodulator. A demodulation rule producing the three outputs 0, ?, and 1 is the simplest example of *soft-decision demodulation*.

1.3.4. The Additive White Gaussian Noise Channel

A channel of the DMC type having great theoretical and practical importance is one in which the output is simply the input plus broadband Gaussian noise. It is conventional to represent broadband Gaussian noise with the white Gaussian noise model. *White Gaussian noise* is defined to be a random process, each sample of which is a zero-mean Gaussian random variable and whose power spectral density is flat over the entire frequency range $-\infty \leq f \leq \infty$, with a level $N_0/2$ watts per hertz. Equivalently, the one-sided *noise spectral density* is N_0, so that, for example, a filter with a rectangular passband of width W hertz will pass N_0W watts of noise power.

The *additive white Gaussian noise* (AWGN) channel can now be described simply in terms of the input x and the output y, which are related by

$$y = x + n_G$$

where n_G is a zero-mean *Gaussian random variable* with variance σ^2 and the input x can have any one of M discrete values, where $M \geq 2$. That is, the conditional probability density function of the output y, given an input x_i, is given by

$$p(y|x = x_i) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-(y-x_i)^2/2\sigma^2}$$

The AWGN channel is an accurate model for many communication links, such as satellite and deep-space links, in which the dominant effect limiting communication performance is additive thermal or galactic noise.

1.4. INFORMATION THEORY AND CHANNEL CAPACITY

In this section we review some of the principal results of information theory and use them to provide insight into the role that coding plays, in combination with the choice of modulation technique, to achieve reliable communication with an efficient expenditure of signal energy. A heuristic explanation of Shannon's theorems on channel capacity is outlined. More complete presentations of the principles of information theory can be found in several excellent textbooks, such as Gallagher [49] or Viterbi and Omura [175].

1.4.1. Logarithmic Measures of Information

The beginnings of information theory lie in early papers by Nyquist [117] and Hartley [61], who were concerned with the achievable transmission capabilities of telegraph circuits and, more generally, with a mathematical characterization of the ultimate limitations on the amount of data that can be transmitted reliably over any given physical channel. The formulation of such problems requires an explicit mathematical measure of information. Hartley considered this question in relation to an information source producing messages, where each message is drawn with equal likelihood from a discrete set. Hartley suggested that the most natural measure of information is the logarithmic function. That is, the information content of one message drawn from a set of M equally likely messages is $\log M$, where the logarithm base is arbitrary and depends on the basic unit of information. (Note that all logarithmic measures can be related directly to one another, since $\log_a x = (\log_a b)(\log_b x)$ for any x.) Conventionally, the logarithm base is chosen to be 2, and the unit of information is called a *bit*, a contraction of "binary digit."

Note that earlier in the chapter we used the term "bit" in a more general way to mean any binary digit; both usages are common in the literature.

The logarithmic measure of information is logical and intuitively appealing, since if we equate a unit of information to a unit of storage capacity, the quantity $\log M$ corresponds exactly to the amount of storage needed to hold a representation of each of the possible messages. With binary storage, for example, we can hold a representation of one of 128 messages with seven bits, since $2^7 = 128$. Given this convention, we can define the *rate of an information source*. We assume that the source produces equally likely M-ary symbols, with the output symbols being independent from one symbol interval to the next. The rate of the source is then $\log_2 M$ bits per symbol. Throughout most of this book we assume a source that produces equally likely information bits or symbols. However, in the discussion at hand, a more general measure of information is needed for sources in which messages or symbols are not necessarily generated with equal probability.

A measure of information for general sources was provided by Shannon with the application of the concept of *entropy* to an information source. The concept of entropy has long been used in the field of physics as a measure of the randomness of a physical system whose states can be described only in statistical terms. Entropy provides an appropriate measure of the a priori uncertainty of any symbol or message to be produced by a discrete information source. Let us say that a source produces any one of M symbols, where the probabilities of occurrence are $p_1, p_2, \ldots, p_M$, with $p_1 + p_2 + \cdots + p_M = 1$. The entropy of the source is defined by

$$H = -\sum_{i=1}^{M} p_i \log p_i \tag{1.1}$$

This definition in effect averages Hartley's logarithmic information measure with respect to the set of probabilities of the individual source symbols. It is readily seen that for a set of equally likely symbols, the entropy of the source is simply $H = -\log p_i = \log M$. The entropy function H provides a measure of the average amount of information "produced" per symbol by the source. Equivalently, from the point of view of the ultimate recipient of the information, it represents the average "prior uncertainty" of the information in each symbol generated by the source. The notion of prior uncertainty will be useful when we discuss measures of information transfer through a channel. But first we want to point out certain important properties of the entropy function.

Using Eq. (1.1) we readily obtain the entropy function for a binary information source as

$$H = -p \log p - (1 - p)\log(1 - p)$$

where p is the probability of occurrence of either of the symbols. The reader should calculate and sketch a few values of H and note that the entropy is

maximized when $p = 0.5$, which yields one bit of information per symbol. For an M-ary information source, it would be a simple matter to show that the entropy is maximized by having the outputs occur with equal probability, an intuitively satisfying result.

The entropy function for a discrete source generalizes readily to the case of a continuous information source, for example, a source whose output is a voltage that can have any value over a continuous range. Let us say that we have such a source and that a sample of its output has a probability density function $p(x)$. The entropy of this source is given by

$$H = -\int_{-\infty}^{\infty} p(x)\log[\,p(x)]\,dx$$

where the summation in Eq. (1.1) has simply been converted to an integral.

Since we want the entropy function of a continuous source to accurately describe a physical reality, such as the information content of a signal voltage, we need to place reasonable constraints on $p(x)$. It is interesting to consider several such constraints and ask what $p(x)$ maximizes the entropy. For instance, it can be shown easily that if the only constraint is to confine the output to some finite interval, the entropy is maximized by letting $p(x)$ be uniform over the given interval. This might reasonably have been guessed by generalization from the case of a discrete M-ary symbol source. Another important case is one that we consider in the following example.

EXAMPLE: A CONTINUOUS INFORMATION SOURCE WITH FIXED VARIANCE. Consider a continuous information source whose output x has the probability density function $p(x)$. We wish to find the function $p(x)$ that maximizes the entropy function given only the constraint that the variance of $p(x)$ is fixed at a given value σ^2. This is equivalent to fixing the average power of a signal voltage x. We want to maximize the integral

$$H = -\int_{-\infty}^{\infty} p(x)\log[\,p(x)]\,dx$$

subject to the constraints

$$\int_{-\infty}^{\infty} p(x)\,dx = 1 \tag{1.2}$$

and

$$\int_{-\infty}^{\infty} x^2 p(x)\,dx = \sigma^2 \tag{1.3}$$

This is done by straightforward application of the method of Lagrange

multipliers [66]. We form the integral

$$\int_{-\infty}^{\infty} p(x)\{-\log[p(x)] + L + Mx^2\}\, dx \tag{1.4}$$

where L and M are undetermined multipliers of the integrals in Eqs. (1.2) and (1.3), which define the constraints. We want to maximize the integral in Eq. (1.4) by selecting the appropriate function $p(x)$. To do this we differentiate with respect to $p(x)$ and set the result equal to zero, producing the condition

$$-1 - \log[p(x)] + L + Mx^2 = 0$$

which in turn yields

$$p(x) = e^{L-1}e^{Mx^2} \tag{1.5}$$

The multipliers L and M are determined by substituting Eq. (1.5) into Eqs. (1.2) and (1.3), and we find

$$p(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-x^2/2\sigma^2}$$

which is seen to be the zero-mean Gaussian density function. Therefore, subject to a power constraint, the continuous information source having the greatest entropy is the Gaussian source. Its entropy is given by

$$\begin{aligned} H &= \log(\sqrt{2\pi e}\,\sigma) \\ &= \tfrac{1}{2}\log(2\pi e\sigma^2) \end{aligned} \tag{1.6}$$

which has units of bits per sample, and where we have used the property $\log_2 a = (\ln a)(\log_2 e)$.

We shall return to the result in Eq. (1.6) after further developing the concept of entropy as applied to a communication channel.

1.4.2. Transfer of Information Through a Channel

We have said that it is useful to view the entropy H of an information source as a measure of the prior uncertainty about the information produced by the source. Therefore, from the point of view of the intended recipient of the information, H represents the state of uncertainty before receiving information from the source. For the case of error-free discrete transmission, the user receives uncorrupted source symbols, and the uncertainty about the source information vanishes completely as symbols are received. In this ideal

case, the channel transfers information from the source to the user at an average rate of H bits per symbol, exactly the amount of uncertainty prior to transmission. We can quantify this notion by observing that after receipt of each symbol, the prior probabilities of the symbols, $p_1, p_2, \ldots, p_M$, are replaced with a distribution having probability 1 for the symbol received and 0 for all other symbols. If we calculate the entropy after reception we have 0, since $\log(1) = 0$.

In all realistic cases, of course, transmission through the channel is not error-free, and thus after reception the user is left with some residual uncertainty concerning the exact identity of the transmitted information. Minimizing this residual uncertainty while making efficient use of signal energy is the essence of the communication system design problem. Given the use of the entropy function to measure a priori uncertainty of the source information, it is logical to use a corresponding function to measure the a posteriori uncertainty of the information after reception on a noisy channel. To express this more formally, we consider a discrete memoryless channel characterized by a set of inputs $x_1, x_2, \ldots, x_M$, having a priori probabilities $\{P(x_i)\}$, a set of outputs $y_1, y_2, \ldots, y_Q$, and a set of transition probabilities $\{P(y_j|x_i)\}$ specifying the probability of receiving a symbol y_j given that a symbol x_i was transmitted. This channel model is illustrated in Fig. 1.4. Using the elementary rules of probability, we can write the *a posteriori* probability of x_i given the receipt of y_j as

$$P(x_i|y_j) = \frac{P(x_i, y_j)}{P(y_j)}$$

$$= \frac{P(y_j|x_i)P(x_i)}{P(y_j)}$$

Recalling that the a priori uncertainty about x_i is measured by $-\log P(x_i)$, we can define the a posteriori uncertainty about x_i given the receipt of y_j as

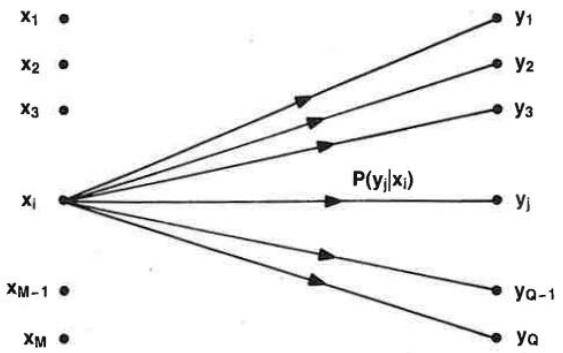

FIGURE 1.4. Model of an M-input, Q-output discrete memoryless channel.

$-\log P(x_i|y_j)$. The difference between these two values is defined as the *mutual information* associated with the transmission of x_i and reception of y_j and is given by

$$I(x_i; y_j) = \log \frac{P(y_j|x_i)}{P(y_j)}$$

We now calculate the overall rate of transfer of information through the channel by simply averaging $I(x_i; y_j)$ over all possible values of x_i and y_j, giving us the *average mutual information*, defined by

$$I(X; Y) = \sum_{i=1}^{M} \sum_{j=1}^{Q} P(x_i, y_j) \log \frac{P(y_j|x_i)}{P(y_j)} \tag{1.7}$$

which has units of bits per symbol when the logarithm base is 2. Since the output symbol probability $P(y_j)$ is given by

$$P(y_j) = \sum_{i=1}^{M} P(x_i) P(y_j|x_i)$$

it is seen that $I(X; Y)$ follows directly from specification of the probabilities of occurrence $P(x_i)$ of source symbols and the set of transition probabilities $P(y_j|x_i)$. The average mutual information is the average rate of transfer of information through the channel, given the distribution of source symbols and the channel transition probabilities. If we are given the transition probabilities defining the channel, the information transfer rate through the channel will depend upon the probabilities of occurrence of the input symbols. The *capacity* of the channel is defined as the maximum value of $I(X; Y)$ with respect to all input distributions, that is,

$$C = \max_{P(x_i)} I(X; Y) \tag{1.8}$$

given in bits per transmitted channel symbol. (Channel capacity can be expressed with various units of measurement, and the separate forms are useful in different contexts. We shall be careful to specify the different forms in our discussions.)

Thus we see that given any channel that is defined by the input-to-output transition probabilities, one can determine the maximum achievable information transfer rate through the channel by performing the indicated maximization over all possible input distributions. Let us consider a simple example.

EXAMPLE: CAPACITY OF THE BINARY SYMMETRIC CHANNEL. The binary symmetric channel was discussed in Section 1.3 and is described by the

transition diagram in Fig. 1.3*a*. It is readily seen that the channel is completely characterized by the crossover probability p. If we denote the input symbols as $x = 0, 1$ and the corresponding output symbols as $y = 0, 1$, the transition probabilities $P(y_j|x_i)$ are simply

$$P(0|0) = P(1|1) = 1 - p$$

$$P(1|0) = P(0|1) = p$$

The information transfer rate of this simple channel is maximized when the input symbols are equally likely, that is $P(0) = P(1) = 0.5$, which results in the following expression for capacity:

$$C = 1 + p \log p + (1 - p)\log(1 - p)$$

A plot of this expression is shown in Fig. 1.5. Note that for $p = 0$, corresponding to error-free transmission, C equals one bit per transmitted symbol, which is the entropy of the source with $P(0) = P(1) = 0.5$. For $p = 0.5$, either value of y is received with the same probability regardless of which value of x is transmitted. Thus $C = 0$, and the channel fails to transmit any information about the source symbols to the user. Note also that the capacity curve for the BSC is symmetric about $p = 0.5$. The portion of the curve for $p > 0.5$ has the same meaning as the region $p < 0.5$ if we interchange the assignments of the values 0 and 1 for the output symbols. Therefore, in using the binary symmetric channel model we need only consider values of p in the range $0 \leq p \leq 0.5$.

It is useful to write the expression for average information transfer given in Eq. (1.7) in either of the forms

$$I(X; Y) = H(X) - H(X|Y)$$

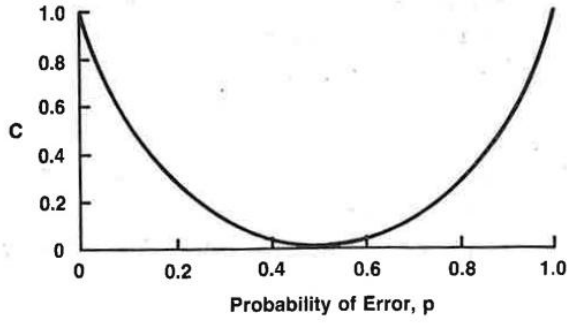

FIGURE 1.5. Capacity of the binary symmetric channel. The units of C are bits per transmitted symbol.

or

$$I(X;Y) = H(Y) - H(Y|X) \tag{1.9}$$

That is, $I(X;Y)$ can be expressed in terms of entropy functions calculated with unconditional or conditional probability distributions of input and output symbols. The second form, Eq. (1.9), is particularly instructive, with

$$H(Y) = -\sum_{j=1}^{Q} P(y_j)\log P(y_j) \tag{1.10}$$

and

$$H(Y|X) = -\sum_{i=1}^{M}\sum_{j=1}^{Q} P(x_i, y_j)\log P(y_j|x_i) \tag{1.11}$$

What is important about the use of Eqs. (1.9) to (1.11) for expressing average information transfer is that for certain types of channels the determination of channel capacity is relatively straightforward. In particular, there are channels described as *uniform from the input*, defined by the property that the sets of transition probabilities from the various channel inputs are permutations of the same set of numbers $p_1, p_2, \ldots, p_Q$. The importance of this property is that for any channel of this type the conditional entropy $H(Y|x_i)$ is given by

$$H(Y|x_i) = -\sum_{j=1}^{Q} P(y_j|x_i)\log P(y_j|x_i)$$

$$= -\sum_{j=1}^{Q} p_j \log p_j$$

which has the same value for any channel input x_i. Therefore, the average conditional entropy given by Eq. (1.11) is independent of the distribution of input symbols and may be written as

$$H(Y|X) = -\sum_{j=1}^{Q} p_j \log p_j$$

The practical interpretation of this model is that when a channel is uniform from the input, the transmission of any of the input symbols is disturbed to the same extent by channel noise. Thus, if Eq. (1.9) is used to derive the capacity, we see that only the first term, $H(Y)$, varies with the input symbol distribution,

and we may write the definition of capacity using Eqs. (1.8) and (1.9) as

$$C = \max_{P(x_i)} [H(Y)] - H(Y|X) \tag{1.12}$$

We can now see that for any channel of this type, the information transfer rate is maximized by choosing the input symbol distribution to maximize the entropy of the set of output symbols. From Fig. 1.3 it is readily seen that the BSC and BSEC models are uniform from the input. The capacity of the BSC has already been given. For the BSEC, the reader should verify that $H(Y)$ is maximized when the two input symbols are equally likely, and show that the capacity is

$$C = (1 - q)[1 - \log(1 - q)] + (1 - p - q)\log(1 - p - q) + p \log p$$

For discrete channels with larger numbers of input and output symbols, the uniform distribution of input symbols does not necessarily achieve capacity, even in cases where there is uniformity from the input. However, for a type of channel called a *doubly uniform channel*, capacity is always achieved with a uniform input distribution. A doubly uniform channel is not only uniform from the input but also *uniform from the output*. By the latter property we mean that for all the Q output symbols the sets of transition probabilities from the M inputs are permutations of the same set of M numbers. For any channel of this type it can be shown that capacity is given by

$$C = \log Q + \sum_{j=1}^{Q} p_j \log p_j$$

where the $\{p_j\}$ are the transition probabilities conditioned on any one of the M input symbols. It should be noted that the BSC is a doubly uniform channel, but the BSEC is uniform only from the input.

For channels with continuous rather than discrete input or output signals, the summations in Eqs. (1.10) and (1.11) are converted to appropriate integrals involving probability density functions, and equivalent arguments apply.

1.4.3. Capacity of the Continuous AWGN Channel

We now discuss a channel in which communication is accomplished by transmission of continuous waveforms rather than discrete symbols. In particular, consider a band-limited AWGN channel, and let the bandwidth be denoted by W, given in hertz (Hz). We would like to know the maximum information transfer rate for the channel, that is, the channel capacity, and also what transmission waveforms should be used to achieve capacity. The only constraint we place on the allowable signals is that they have some finite average power. We shall provide a brief outline of the derivation of capacity

for this channel, following a similar presentation given by Woodward [183]. More complete and rigorous derivations may be found in a number of texts on information theory, including references already cited in this chapter.

By well-known principles of sampling theory, the signal and noise waveforms received in the band-limited channel can be represented by independent samples taken at the *Nyquist rate* $2W$, where W is the channel bandwidth. Let us denote the transmitted waveform as $x(t)$, the additive white Gaussian noise as $n(t)$, and the noisy received waveform as $y(t) = x(t) + n(t)$. Also let us denote the average power of a received signal or noise sample as, respectively,

$$\bar{x}^2 = S \qquad \text{and} \qquad \bar{n}^2 = N$$

and let us suppose that these average power levels are fixed. Since we are considering continuous rather than discrete signals, we can modify the expressions developed in Section 1.4.2 by replacing the input and output symbol distributions $\{P(x_i)\}$ and $\{P(y_j)\}$ with probability density functions $p(x)$ and $p(y)$, respectively. We also replace the transition probability $P(y_j|x_i)$ with the conditional probability density function $p(y|x)$. Now, for an amplitude-continuous additive noise channel, it is clear that $p(y|x)$ depends on y and x only through their difference $y - x$, and for the case of Gaussian noise we have

$$p(y|x) = \frac{1}{\sqrt{2\pi N}} e^{-(y-x)^2/2N}$$

where N is the average power or variance of a noise sample. Noise additivity is a property of continuous channels that corresponds directly to the property of uniformity from the input discussed in the previous section for discrete channels. Therefore, we can derive channel capacity using Eq. (1.12), where the output entropy $H(Y)$ is calculated for sequences of samples of the received noisy signal $y(t)$, and the conditional entropy $H(Y|X)$ depends only on the statistical distribution of the noise samples, that is, sequences of samples of $n(t)$. We rewrite Eq. (1.12) as

$$C = \max_{p(x)} [H(y)] - H(n)$$

where y denotes a sequence or vector of received noisy signal samples and n denotes a sequence or vector of noise samples. Now we can compactly define the capacity of the continuous-waveform AWGN channel as the maximum entropy of the total received signal minus the entropy of the noise. The maximization is to be done by appropriate selection of the statistics of the transmitted waveforms $x(t)$.

First we consider $H(n)$. Recall from the discussion leading to Eq. (1.6) that the entropy of a sample drawn from a Gaussian distribution is $\frac{1}{2}\log(2\pi e\sigma^2)$, where σ^2 is the variance of the distribution. Thus, here the entropy of a sample of the Gaussian noise is $\frac{1}{2}\log(2\pi eN)$, where N is the variance, or average

power, of a noise sample. The entropy of a sequence of $2WT$ noise samples taken in an interval of time equal to T seconds is then

$$H(n) = WT \log(2\pi eN)$$

We now consider $H(y)$ and recall from the same discussion leading to Eq. (1.6) that if we constrain a continuous random variable to have a given fixed variance, the entropy of the variable is maximized by letting its probability density function be Gaussian. Therefore, since we have constrained the signal plus noise $y(t)$ to have average power $S + N$, it follows that the received signal entropy $H(y)$ will be maximized by letting $y(t)$ have the statistics of Gaussian noise, each sample having variance $S + N$. We can make this happen by transmitting signals that are Gaussian noise waveforms, and by reapplication of Eq. (1.6) we have for a sequence of $2WT$ samples,

$$\max_{p(x)} H(y) = WT \log[2\pi e(S + N)]$$

Therefore the maximum information transfer rate is

$$I_{\max} = WT \log\left(\frac{S + N}{N}\right)$$

which is measured in bits transmitted per T seconds. Equivalently, capacity per unit time is given by

$$C = W \log\left(\frac{S + N}{N}\right) = W \log\left(1 + \frac{S}{N}\right) \text{ bits/s} \tag{1.13}$$

This is Shannon's capacity formula for the band-limited continuous AWGN channel. We shall see shortly that the practical implication of this and another closely related result is that the most efficient use of the channel is achieved by setting up a suitable correspondence between long sequences of source information digits and long noiselike signaling waveforms $x(t)$ for transmission on the channel. The most remarkable of Shannon's results, which we shall not derive here, is the following:

If we take increasingly long sequences of source digits and map them into correspondingly long transmission waveforms, the error rate in the data delivered can be brought arbitrarily close to zero, as long as we do not attempt to transmit data at a rate higher than C. Therefore, at any nonzero level of channel signal-to-noise ratio S/N, there is some nonzero information transfer rate below which arbitrarily accurate communication can in principle be achieved.

The essence of Shannon's result, which is called the *channel coding theorem*, is that noise in the channel does not inherently limit the accuracy with which communication can be achieved, but only the rate at which information can be

reliably transmitted [158,159]. We discuss this result in more detail in the following section.

1.4.4. Channel Coding Theorem

The channel coding theorem states that every channel has a channel capacity C, and that for any information transfer rate $R < C$ there exist codes of block length n and rate R having probability of incorrect decoding $P(E)$ bounded by

$$P(E) \leq 2^{-nE_b(R)} \tag{1.14}$$

where the exponent $E_b(R)$ is a positive function of R for $R < C$ and is determined solely by the characteristics of the channel. The implication of the bound in Eq. (1.14) is that for any information rate less than C, the error probability can be made arbitrarily small by increasing the code block length n while holding the code rate constant. A similar bound can be written for convolutional codes, where n is replaced by k, the code constraint length.

Shannon's derivation of the exponential error bound formula was based upon an analysis called the *random coding argument*. The bound is obtained by averaging the error probability over an ensemble of randomly selected codes. Since some codes in the ensemble must perform better than the average, the coding theorem guarantees the existence of codes capable of achieving the bound in Eq. (1.14).

Therefore, Shannon's work shows that it is not really necessary to transmit Gaussian noise waveforms in order to achieve capacity; rather, well-chosen codes can be used to produce the same result. For a coded communication system, sequences of information bits are mapped into long codewords by the error-control encoder and then into long digital waveforms by the modulator. The demodulator and decoder then utilize all the received signal energy during the transmission of a codeword in the decision-making process. Several practical issues need to be dealt with, however. First, the coding theorem provides no means for constructing effective codes. Second, requirements for very low error probabilities will compel the use of very long codes, and this in turn will lead to very complex decoding operations.

Because of the issues just outlined, much of the research in the coding field over the last three decades has dealt with two key problems: finding classes of codes that yield good performance over wide ranges of lengths, and designing decoding algorithms that realize the intrinsic code performance without prohibitive complexity. Chapters 2 through 11 will present some of the more important and useful results that have been obtained in addressing these questions.

However, no practical means of achieving channel capacity has yet been found. In Section 1.6 we discuss other implications of the channel capacity formula and outline a valuable concept closely related to channel capacity that

provides a calculable measure of modulation and coding performance achievable with practical implementations.

1.5. MODULATION PERFORMANCE ON THE AWGN CHANNEL

In this section we summarize formulas giving the performance that can be achieved in additive white Gaussian noise with several of the commonly used forms of digital modulation. Detailed derivations of these formulas can be found in a number of references, including Arthurs and Dym [4], Viterbi [172], and Proakis [139].

1.5.1. Phase-Shift Keying

Let us first consider binary signaling in the AWGN channel, so that we write the received signal as

$$r(t) = s_i(t) + n(t), \qquad 0 \le t \le T$$

where $s_i(t)$, with $i = 0$ or 1, is the transmitted waveform and $n(t)$ represents the white Gaussian noise waveform in the signaling interval T. For *phase-shift keying* (PSK), the two possible waveforms are chosen to be *antipodal*, so that

$$s_0(t) = -s_1(t), \qquad 0 \le t \le T$$

where $s_0(t)$ and $s_1(t)$ are sinusoids of the same frequency with fixed phases 180° apart. Binary PSK is sometimes called *biphase* modulation.

The transmission channel is assumed to have finite bandwidth, which places a practical upper limit on the pulse transmission rate. Let us say that PSK pulses are transmitted at the rate B pulses per second, which is equal to or less than the upper limit. If the average transmitted signal power is S, then the energy per PSK pulse is $E_b = S/B = ST$. Note that E_b was defined earlier to be the average energy transmitted per bit of source information. For uncoded binary transmission the signal energy per pulse equals the energy per source bit, and so we shall use the symbol E_b in the error-probability formulas that follow.

Optimum detection of binary PSK signals is done with a *matched filter* followed by a sampler, the filter being matched to either $s_0(t)$ or $s_1(t)$. Assuming perfect phase coherence, analysis of matched filter detection for PSK shows the probability of error p to be

$$p = \frac{1}{\sqrt{2\pi}} \int_{\sqrt{2E_b/N_0}}^{\infty} e^{-x^2/2}\, dx$$

which can be written more compactly as

$$p = \frac{1}{2}\operatorname{erfc}\left(\sqrt{A_b}\right) \tag{1.15}$$

where $\operatorname{erfc}(x)$ is the complementary error function, defined by

$$\operatorname{erfc}(x) = \frac{1}{\sqrt{\pi}} \int_x^\infty e^{-t^2}\, dt$$

and $A_b = E_b/N_0$ is the *signal-to-noise ratio* (SNR) per bit.

1.5.2. Differential PSK

Coherent PSK demodulation requires the use of a phase reference. This is usually done by extracting the carrier phase from the received signal with a *phase-locked loop*. In some systems, however, it is difficult to obtain a reliable phase reference due to characteristics of the transmission medium or the demodulation equipment itself. In such cases, PSK signaling can still be used if the signals to be transmitted are mapped into successive phase differences. The resulting modulated waveform is known as *differentially encoded PSK* or simply *differential PSK*, abbreviated as DPSK. In binary DPSK a 1 is transmitted by sending the pulse waveform which is 180° out of phase with the pulse sent in the previous interval, while a 0 is transmitted by sending a pulse in phase with the previous pulse. This is called *differential encoding* or *differential precoding*. *Differentially coherent demodulation* of DPSK signals is implemented by detection of the phase difference between received pulses in successive signaling intervals. For steady-signal reception of binary DPSK signals on an AWGN channel, the average probability of error is given by

$$p = \tfrac{1}{2}e^{-A_b} \tag{1.16}$$

where $A_b = E_b/N_0$ is the SNR per bit. The bit-error probability for DPSK was derived originally by Lawton [87].

Because of the partial overlap of signal and noise components involved in successive binary decisions, there is a tendency for errors in differentially coherent demodulation of DPSK to occur in clusters of two. This occurs because, especially under high-SNR, low-error-rate operation, an error is likely to be caused by a momentarily high noise level associated with a single pulse completely distorting its phase. Since any such single pulse is involved in two successive binary decisions, both may have a high likelihood of error. Such a clustering of errors may be of major significance when error-control coding is used, since the coding must then be designed to cope with clusters of errors rather than with statistically independent errors. Numerical results on error clustering behavior with DPSK are given in a paper by Salz and Saltzberg [152].

There is another form of DPSK that is sometimes used in systems where coherent demodulation can be done but where there is an unresolvable 180° ambiguity in the carrier phase. Here a DPSK signal is transmitted, and at the receiving end coherent PSK demodulation is performed on individual pulses, followed by *differential decoding* of the stream of demodulated bits. This scheme does not perform as well as strict-sense coherent PSK signaling and reception, since a single error leaving the PSK demodulator produces two errors after differential decoding. The probability of bit error for differentially encoded coherent PSK lies between that of strict-sense coherent PSK and that of DPSK; at high SNR, the probability of bit error is approximately twice that of coherent PSK. A detailed analysis of the performance of coherent demodulation of DPSK, including a table of computed bit-error probabilities, can be found in Lindsey and Simon [97].

1.5.3. Coherent Frequency Shift Keying

Frequency shift keying (FSK) is a special case of *orthogonal signaling*. A set of signals is said to be orthogonal over an interval T if any pair of different signals, say $s_i(t)$ and $s_j(t)$, have zero crosstalk, that is, if

$$\int_0^T s_i(t)s_j(t)\,dt = 0, \qquad i \neq j$$

With FSK a symbol is transmitted by sending one of a set of tones, where the tone frequencies are chosen so that any pair of tones at different frequencies are orthogonal over the signaling interval T. In binary FSK, two tones are used, and if the tones are both generated and demodulated with known phases it can be shown that orthogonality is obtained with any tone spacing equal to an integer multiple of $1/2T$. In practice, the tone spacing is usually chosen as $1/2T$ in the interest of minimizing bandwidth. Optimum demodulation is done with a pair of coherent matched filters. The bit-error probability for coherent FSK detection in AWGN is

$$p = \tfrac{1}{2}\text{erfc}\left(\sqrt{\frac{A_b}{2}}\right) \tag{1.17}$$

where $A_b = E_b/N_0$ is the SNR per bit.

If we compare this expression with Eq. (1.15) for coherent PSK, we see that equal bit-error probabilities are obtained with 3 dB greater SNR in the case of coherent FSK. In other words, binary coherent FSK has 3 dB poorer performance than binary coherent PSK.

If, in generating the coherent FSK signal, phase continuity is maintained from one pulse to the next, the resulting modulation is called *continuous-phase FSK* (CPFSK). In many applications CPFSK is an attractive form of modulation, because the phase continuity results in a signal spectrum that rolls off more rapidly than spectra for other forms of modulation. There is an important special case of CPFSK modulation and demodulation that achieves

performance identical to coherent PSK. The scheme is called *minimum-shift keying* (MSK). In MSK, the two tones have a frequency separation equal to $1/2T$, hence the term "minimum shift." Optimum performance for MSK is achieved by exploiting a special property of the MSK waveform, namely that it can be constructed as two binary phase-modulated pulse streams or "subchannels" in phase quadrature, each subchannel carrying sinusoidally shaped pulses of duration $2T$, with the two subchannels offset by T seconds. That is, MSK can be viewed as a particular way to implement antipodal signaling on two orthogonal subchannels. The optimum demodulator performs coherent matched-filter detection in each of the quadrature subchannels, producing bit decisions at times $0, 2T, 4T, \ldots$ in one subchannel and at times $T, 3T, 5T, \ldots$ in the other subchannel. It can be shown that the signal energy usable in each subchannel demodulation is E_b, so that the average probability of bit error is the same as for coherent PSK and thus is given by Eq. (1.15).

As with PSK, the data bits may be differentially precoded prior to quadrature-channel MSK modulation, with the objective of again providing resistance to channel phase inversions. This produces CPFSK, the frequencies of the transmitted pulses having a one-to-one correspondence to the source data bits. (This is not the case if differential precoding is omitted.) This form of MSK modulation has also been called *fast frequency-shift keying* (FFSK). If this signal is demodulated with a coherent matched filter correlating over one T-second pulse at a time, the probability of bit error is that already given for coherent FSK in Eq. (1.17). However, the signal can be optimally demodulated using the coherent quadrature channel method outlined earlier, but followed now by differential decoding. With this modulation and demodulation scheme, the bit-error probability is the same as that observed with differentially encoded, coherently detected PSK, discussed at the end of Section 1.5.2.

MSK has found application in a number of systems for which efficient spectral utilization and low crosstalk are important requirements. The interested reader may refer to a tutorial paper on MSK by Pasupathy and other references cited there [126].

1.5.4. Noncoherent Binary FSK

In most applications, binary FSK signals are demodulated noncoherently, that is, with a detector that operates without knowledge of the received carrier phase. Then it is necessary that the tones be spaced by an integer multiple of $1/T$ Hz, which ensures orthogonality even if the phases are arbitrary. Demodulation is usually done with two pairs of quadrature filters, one pair matched to each frequency, followed by a comparison of envelopes or squared envelopes of the outputs of the two filter pairs. For steady-signal reception in AWGN, the bit-error probability is

$$p = \tfrac{1}{2}e^{-A_b/2}$$

where $A_b = E_b/N_0$ is the SNR per bit.

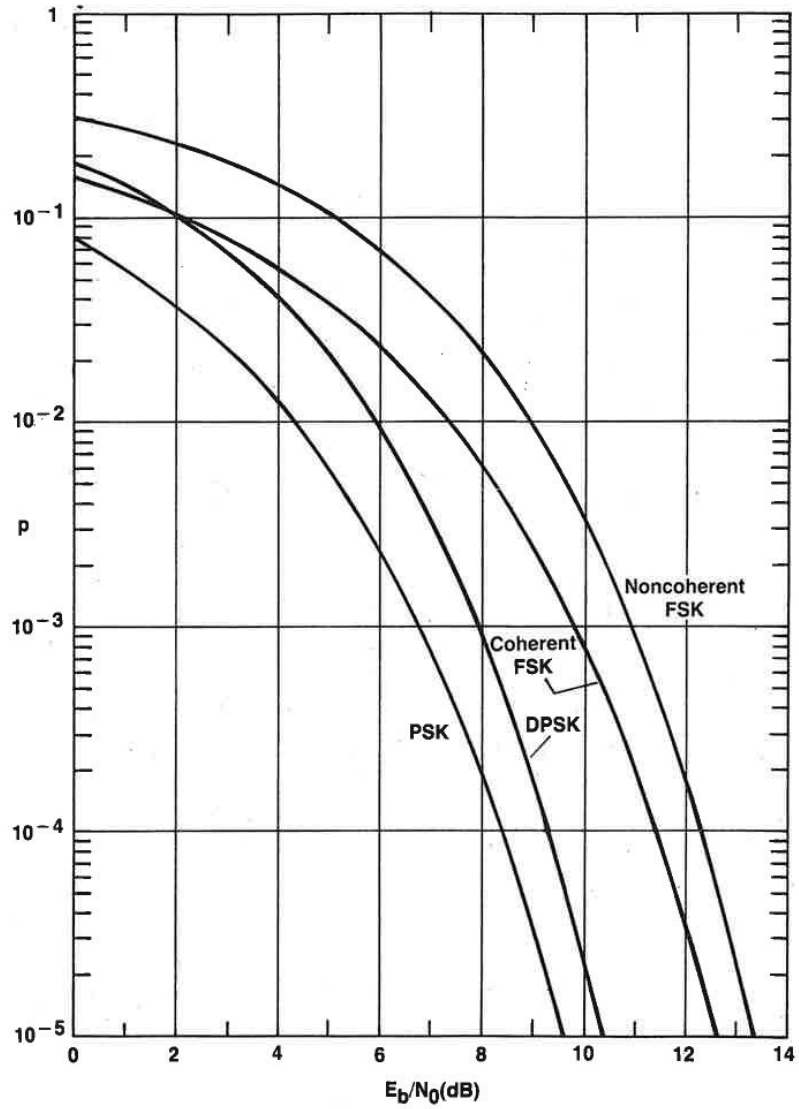

FIGURE 1.6. Probability of bit error for several binary modulation methods.

Comparing the above expression with Eq. (1.16), one sees that the performance of noncoherent FSK is 3 dB poorer than that of DPSK.

Figure 1.6 shows the probability of bit error versus SNR per bit for coherent PSK, DPSK, coherent FSK, and noncoherent FSK.

1.5.5. *M*-ary Signaling on the Gaussian Noise Channel

The basic types of signaling waveforms outlined above, PSK and FSK, are readily generalized from binary to *M*-ary forms by using *M* signal phases or frequencies, respectively. For *M*-ary PSK transmission, the set of transmitted

waveforms consists of sinusoidal carriers with relative phases $0°, 360°/M$, $2 \cdot 360°/M, \ldots, (M-1) \cdot 360°/M$. In describing the performance of M-ary modulation techniques, we need to distinguish between the *signal energy per channel symbol*, which we denote by E_s, and the energy per bit E_b, already defined. In most applications, M is a power of 2, say $M = 2^m$, so that each M-ary symbol represents m channel bits and therefore $E_s = mE_b$. Performance is most readily derived in terms of the M-ary *symbol-error probability*. For four-phase PSK with coherent detection and AWGN, the 4-ary symbol error probability can be shown to be

$$P_4 = \operatorname{erfc}\left(\sqrt{A_b}\right)\left[1 - \tfrac{1}{4}\operatorname{erfc}\left(\sqrt{A_b}\right)\right] \tag{1.18}$$

where $A_b = E_b/N_0$ is the SNR per bit. Note that since $M = 4$, $A_b = 0.5A_s$, where $A_s = E_s/N_0$. For values of M other than 2 or 4, the error probability formulas cannot be presented in such compact forms as Eqs. (1.15) and (1.18), but rather the symbol-error probability in each case must be obtained by numerical integration. However, it can be shown that for $A_b \gg 1$ the M-ary symbol-error probability is given approximately by

$$P_M \approx \operatorname{erfc}\left(\sqrt{mA_b}\sin\frac{\pi}{M}\right) \tag{1.19}$$

where $m = \log_2 M$.

The relationship between M-ary symbol-error probability and bit-error probability in the corresponding m-bit groups depends upon the assignment of bit groups to symbols, which is up to the designer. The preferred mapping in most cases is *Gray coding*, in which the bit groups assigned to adjacent phases differ by only one bit. For example, a Gray code for 4-ary PSK would assign the bit groups 00, 01, 11, and 10 to successive phases. With the use of a Gray code, and given a practical level of SNR, symbol errors will be predominantly transitions to adjacent phases, and therefore the bit-error probability corresponding to a given level of symbol-error probability is reasonably well approximated by

$$p \approx \frac{1}{m}P_M$$

Curves of P_M versus SNR per bit are shown in Fig. 1.7 for several values of M.

For M-ary DPSK, the error-probability formulas are rather tedious to derive, and even for the relatively simple case of $M = 4$, the error probability must be expressed in terms of higher transcendental functions. We do not wish to take the space to present these formulas here, but instead refer the interested reader to reference [139] or [4].

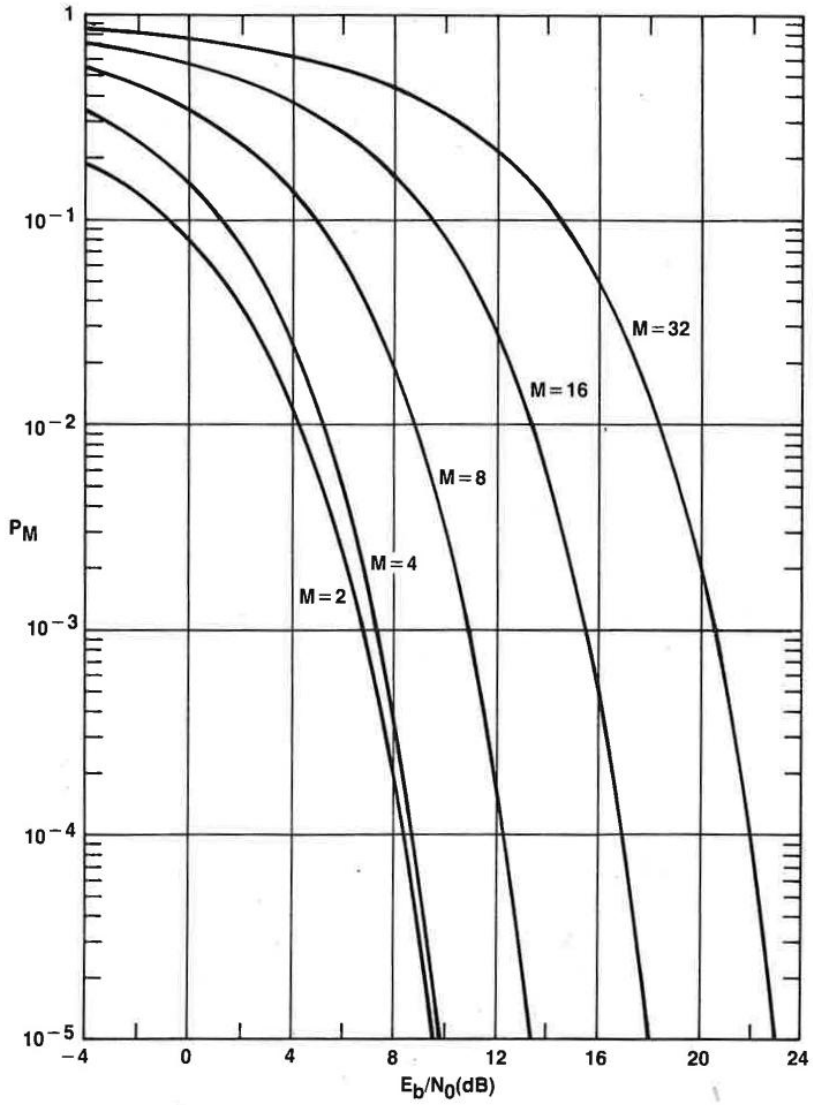

FIGURE 1.7. Probability of symbol error for M-ary PSK modulation.

For coherent detection of M-ary FSK in AWGN, the probability of M-ary symbol error is given by

$$P_M = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \left\{ 1 - \left[1 - \frac{1}{2} \operatorname{erfc}\left(\frac{y}{\sqrt{2}} \right) \right]^{M-1} \right\} e^{-(y-\sqrt{2mA_b})^2/2} \, dy$$

where $A_b = E_b/N_0$ is the SNR per bit, $M = 2^m$, and y is the variable of integration. Curves of P_M versus E_b/N_0 are shown in Fig. 1.8 for several values of M.

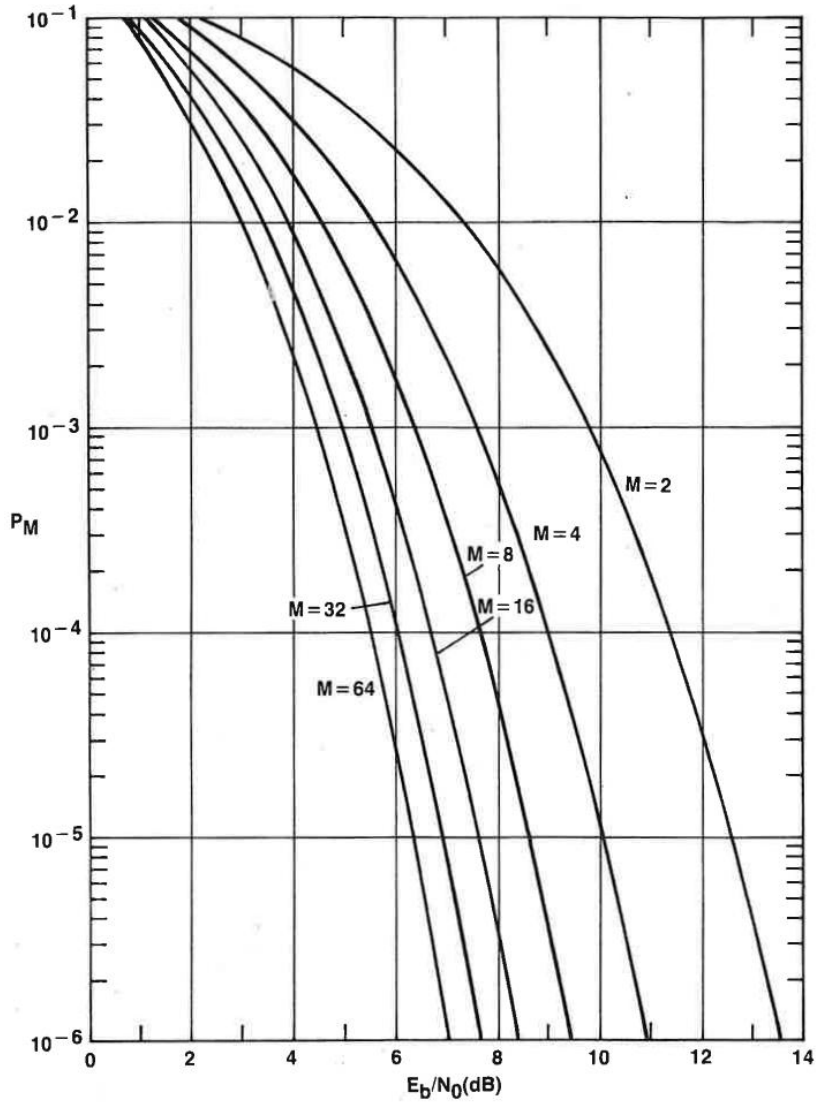

FIGURE 1.8. Probability of symbol error for coherent detection of M-ary FSK signals.

As in the case of binary FSK, M-ary FSK signals may be generated with phase continuity from pulse to pulse. The resulting modulation is called *M-ary CPFSK*. As a consequence of the phase continuity being maintained, the M-ary CPFSK signal has memory that in general can extend over a number of pulses. While it is possible to ignore this inherent memory in the detection process, a demodulator that makes use of this memory performs better than one that does not. See, for example, Schonhoff [156].

The error probability P_M for M-ary symbols can be converted to a bit-error probability p for the equivalent m-bit groups by assuming that when an M-ary symbol is in error, each of the $2^m - 1$ incorrect m-bit patterns is equally likely.

It can be shown easily that this leads to the relationship

$$p = \frac{2^{m-1}}{2^m - 1} P_M$$

For noncoherent detection of M-ary FSK in AWGN, the probability of symbol error is given by

$$P_M = \sum_{i=1}^{M-1} (-1)^{i+1} \binom{M-1}{i} \frac{1}{i+1} e^{-imA_b/(i+1)}$$

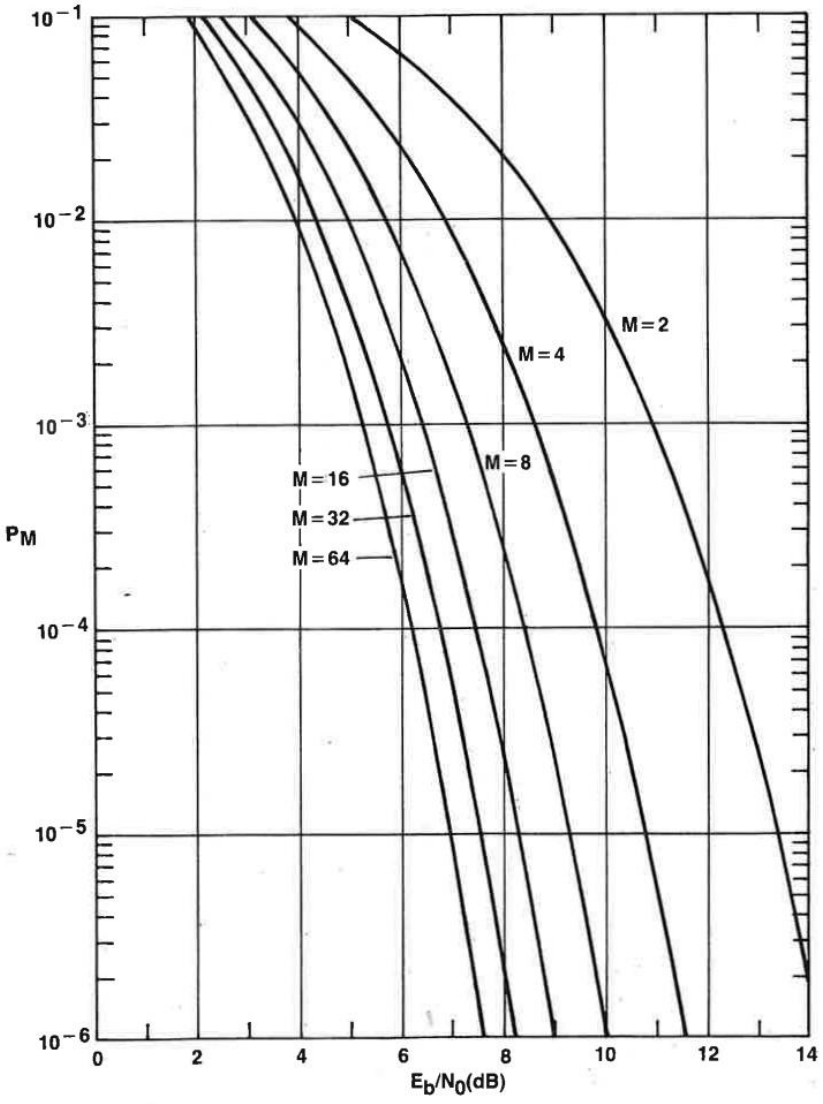

FIGURE 1.9. Probability of symbol error for noncoherent detection of M-ary FSK signals.

Symbol-error probability P_M is plotted in Fig. 1.9 as a function of SNR per bit for several values of M.

Tables of computed error probabilities for all the forms of binary and M-ary modulation described here can be found in Lindsey and Simon [97]. Tables of error probabilities for coherent demodulation of M-ary orthogonal signals can also be found in Viterbi [171].

1.5.6. Comparison of Binary Modulation Techniques

In the chapters that follow, the performance of many errror-control techniques will be presented. Particularly for the block-coded systems, it will be convenient to give results as a function of p, the bit-error rate in the channel, or P_{CE}, the symbol-error rate. These results can be related to channel SNR by inspecting the applicable modulation performance curve. A word of caution is in order, though, since the error-control results assume an independent-error channel. This assumption is valid for all the channel models considered here except binary and nonbinary DPSK and CPFSK. Some degradation can be expected due to the correlation described previously, but this degradation can be eliminated with interleaving (see Chapter 11).

There is an important observation to be made in comparing Figs. 1.7 and 1.8, which give error probabilities for M-ary PSK and M-ary coherent FSK, respectively. Note that for M-ary PSK, the error-rate curves shift to the right (increasing SNR per bit) with increasing M, while for M-ary coherent FSK the curves shift to the left (decreasing SNR per bit) with increasing M. Therefore we see that by providing an expansion of bandwidth, as we must do to increase the library of FSK tones, we can achieve improvements in communication efficiency as measured by the SNR per bit, E_b/N_0. That is, for a desired quality of service, measured here as the probability of a symbol error, the E_b/N_0 required is reduced. In fact, it can be shown that for coherent M-ary FSK signaling, as $M \to \infty$, arbitrarily small error rates can be provided for any $E_b/N_0 \geq \ln 2$, which is equal to -1.6 dB.

Unfortunately, the use of extremely large FSK tone libraries cannot be considered a practical design approach. To see this we consider a parameter called the *bandwidth expansion factor*, B_e, which is defined for any digital modulation scheme as

$$B_e = \frac{W}{R}$$

where W is the overall bandwidth of the set of modulation waveforms measured in hertz and R is the information rate of the modulation in bits per second. For M-ary FSK with tones spaced at $1/2T$ Hz, the overall required bandwidth is approximately $M/2T$, where T is the FSK pulse duration, and R is m/T, where $M = 2^m$. Therefore we see that for M-ary FSK signaling, the

bandwidth expansion factor is

$$B_e = \frac{W}{R} = \frac{M}{2\log_2 M}$$

Thus as $M \to \infty$, the bandwidth expansion required also goes to infinity. In addition, coherent reception of a very large number of orthogonal signals would be required, leading to a very complex design. Nonetheless, the asymptotic MFSK result demonstrates that utilizing additional bandwidth can provide highly efficient and reliable communication. However, by the use of simpler modulation schemes and error-control coding, one can construct waveforms that perform as well as orthogonal waveforms while requiring smaller bandwidth expansions. In other words, coded waveforms can be designed that are more efficient than orthogonal waveforms in their use of bandwidth. The coded waveforms are nonorthogonal, in general.

1.6. COMBINED MODULATION AND CODING FOR EFFICIENT SIGNAL DESIGN

It is becoming common in the field of digital communication to refer to the set of modulation, coding, and decoding techniques used in a system design by the overall name *signal design*. The use of this terminology reflects a growing awareness that the design of an efficient digital communication system is best approached by the selection of modulation and coding techniques jointly as part of an integral design. In this section we present this viewpoint by discussing certain implications of Shannon's capacity formula as well as a closely related capacity-like performance measure called R_0("R zero").

1.6.1. Implications of the Capacity Formula

The underlying principles of efficient signal design can best be seen by examining the capacity formula for the AWGN channel, which we rewrite from Eq. (1.13) in the form

$$\frac{C}{W} = \log\left(1 + \frac{S}{N_0 C}\frac{C}{W}\right) \text{bits/Hz}$$

where N_0 is the one-sided power spectral density of the noise.

This form describes channel capacity in terms of two convenient normalized parameters, S/N_0C and C/W. For transmission at capacity, the first parameter becomes

$$\frac{S}{N_0 C} = \frac{(E_b)_{\min}}{N_0}$$

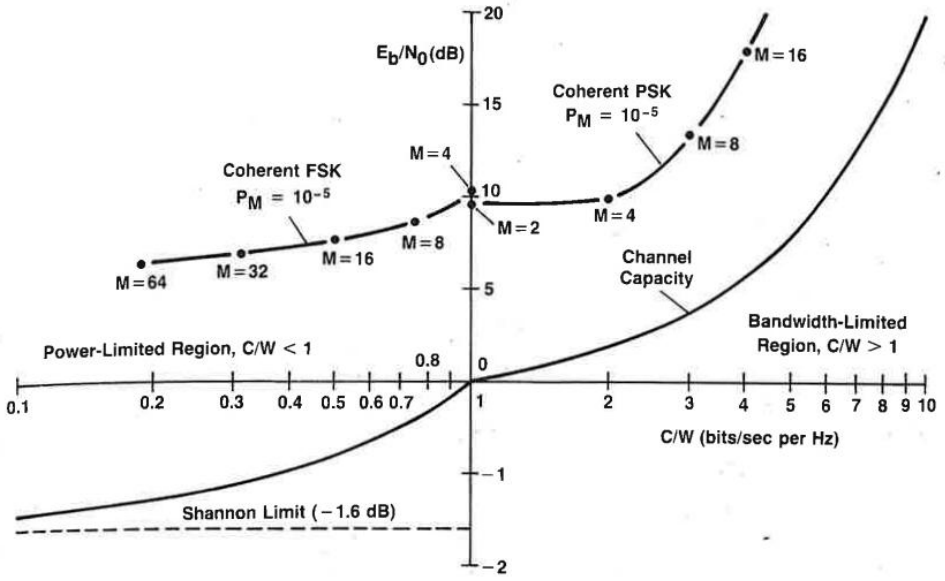

FIGURE 1.10. Channel capacity and a comparison of several modulation methods at symbol-error probability equal to 10^{-5}. Note that the vertical scale is expanded below the origin.

where $(E_b)_{\min}$ is the minimum energy per transmitted source information bit required for reliable communication. The second parameter C/W simply normalizes channel capacity with respect to an arbitrary bandwidth and is seen to be the reciprocal of the bandwidth expansion for operation at capacity.

Figure 1.10 shows the relationship between E_b/N_0 and C/W as given by the capacity formula. Note that the lower portion of the vertical scale is expanded for convenience in drawing the figure. It is conventional to call the region of $C/W < 1$ the *power-limited region* of operation and the region of $C/W > 1$ the *bandwidth-limited region*. As can be seen from the figure, the greatest energy efficiency is achieved when the bandwidth can be made large by comparison with the information rate. In the limiting case, very large bandwidths, C/W approaching zero, E_b/N_0 approaches ln 2 or -1.6 dB, which is called the *Shannon limit*. Thus, by invoking the coding theorem, we can say that for any SNR per information bit E_b/N_0 equal to or greater than -1.6 dB, the probability of error in delivered information can theoretically be made arbitrarily close to zero by use of a suitably chosen error-control code.

It is seen that the Shannon limit, -1.6 dB, is identical to the minimum level of E_b/N_0 achievable with coherent *M-ary FSK* as $M \to \infty$ (see Section 1.5.6). The use of error-control coding, however, has promise of less complex transmission and reception equipment, since long codewords are easier to generate and decode than long orthogonal M-ary waveforms. With M-ary FSK, energy efficiency is realized entirely with the modem and the complexity of demodulation grows very rapidly with M. Furthermore, M-ary FSK waveforms are relatively extravagent in their use of bandwidth when compared

with well-chosen codes as we noted in Section 1.5.6. A key point here is that the bandwidth consumed by MFSK waveforms provides orthogonality for every pair of waveforms in the set, but the achievement of energy-efficient communication does not require orthogonality. Design points for several MFSK modulations at symbol error probability 10^{-5} are shown in Fig. 1.10. Note that for practical values of M, $M \leq 32$, significant gains in communication efficiency can be obtained by increasing M.

The assumptions of severely limited power and practically unlimited bandwidth are valid in some applications, for example in the design of data links for space probes and some satellite systems. In these cases the relationship shown in Fig. 1.10 suggests the use of highly redundant coding in the power-limited region, where E_b/N_0 is to be minimized. For operation in this region, highly efficient communication systems have been designed using binary PSK modulation and low-rate block and convolutional codes decoded by powerful decoding algorithms, which will be described in later chapters. While the Shannon limit cannot be achieved in practice, there are practical schemes that can approach a computation-limited information transfer rate, termed R_0, which is exactly 3 dB above the Shannon limit for the AWGN channel. This point is addressed further in Section 1.6.2, and the coding schemes that we refer to are described in Chapters 10 and 11.

Many important communication channels can be characterized as bandwidth-limited — for example, wireline telephone circuits and radio channels in crowded regions of the radio spectrum. It is instructive, therefore, to examine the capacity formula in the bandwidth-limited region, which in turn implies high SNR if capacity is to be achieved.

It will be useful here to use another normalized form of the capacity formula:

$$\frac{C}{2W} = \frac{1}{2}\log\left(1 + \frac{S}{N_0 W}\right) \text{bits/symbol}$$

which expresses capacity in bits per transmission symbol, assuming signaling at the Nyquist rate of $2W$ symbols per second.

At high levels of SNR we see that this formula gives

$$C \cong \frac{1}{2}\log\left(\frac{S}{N_0 W}\right) \text{ bits/symbol}$$

The need for operation at high SNR in limited bandwidth suggests the use of nonbinary modulation, and so we consider the asymptotic performance of M-ary coherent PSK modulation in Gaussian noise as an example. An approximation to the probability of symbol error, P_M, obtained from Eq.

(1.19), is

$$P_M \cong \text{erfc}\left(\sqrt{\frac{E_s}{N_0}}\sin\frac{\pi}{M}\right)$$

where E_s is the energy per symbol, that is, per M-ary PSK pulse, and erfc(x) is the complementary error function. If $E_s/N_0 \gg 1$, this expression gives a very accurate approximation to P_M for arbitrary M. For large M and E_s/N_0, we have, using the small-angle formula,

$$P_M \cong \text{erfc}\left(\sqrt{\frac{E_s}{N_0}}\frac{\pi}{M}\right)$$

Therefore, as M becomes large, the SNR per pulse required to maintain a given probability of symbol error must increase as the square of the number of modulation states; that is, we have, for large E_s/N_0 and large M,

$$\frac{E_s}{N_0} \propto M^2$$

Since the information conveyed per transmission symbol is $\log_2 M$ bits, the information rate needed to maintain constant P_M is

$$R = \log_2 M \propto \log\left(\frac{E_s}{N_0}\right)\text{bits/symbol}$$

which in turn implies

$$R \propto \log\left(\frac{S}{N_0 W}\right)\text{bits/symbol}$$

Therefore the information rate that can be achieved for high-order PSK exhibits asymptotically the same function of SNR as channel capacity.

Design points for several uncoded modulation methods operating at symbol error probability 10^{-5} are shown in Fig. 1.10. Similar asymptotic behavior can also be shown for amplitude-modulated (AM) signals and combined AM/PSK signals.

It is seen, therefore, that for operation in the high-SNR, bandwidth-limited region, channel capacity can be approached by simply using large multiphase or multiamplitude alphabets without a need for coding. This underlines the point that as greater levels of signal power can be provided, system operating points move from the strictly power-limited region, where coding and bandwidth expansion must be employed to achieve reliable transmission at the highest possible rate with the least signal energy, toward the bandwidth-limited

region, where coding offers little or no performance gain with respect to uncoded high-order signaling alphabets. For operating points in the latter region, emphasis is placed on the design of bandwidth-efficient signals and the effects of bandwidth limitations on system performance.

The above discussion serves to outline some of the fundamental principles of modulation and coding design for digital communication channels. In the following section we describe a useful analytical methodology for arriving at specific modulation and coding designs that optimize the efficiency of a digital communication system.

1.6.2. The R_0 Criterion for Modulation and Coding Design

In section 1.6.1 we examined the capacity formula for the AWGN channel with a view toward understanding the conditions under which coding promises to be useful in the design of an energy-efficient communication system. However, we have already pointed out that there are practical limitations in trying to achieve capacity in an actual design and that in fact it has not been done. This would seem to leave us at an impasse, with informative theoretical bounds on information rate and error probability but no guidance for the means to approach the theoretical limits to some reasonable degree and with practical designs. As it happens, there is a line of analysis that provides exactly the design guidance that we would like to have. This analysis revolves around a parameter called the *cutoff rate* of the channel, denoted as R_0 ("R zero"). The cutoff rate R_0 is a capacity-like quantity defined for any discrete memoryless channel, whose value is always less than the channel capacity C. R_0 gives the practical limit on the rate at which information can be reliably transmitted through the channel.

It has been shown [173] that for a system using a convolutional code of constraint length k, the post-decoding error probability is bounded as

$$P_e < C_R 2^{-kR_0}, \qquad R \le R_0 \tag{1.20}$$

where R is the code rate and C_R is a small constant usually determined experimentally. The corresponding bound for block coding is

$$P_e < C_R 2^{-nR_0}, \qquad R \le R_0 \tag{1.21}$$

where n is the code block length. Together with these results it has been shown that for rates $R < R_0$ codes with long constraint length k or block length n can be decoded without suffering an unbounded growth in the number of decoding computations. Thus R_0 at once provides both an error-bound exponent and a practical limit on information transfer rate. Comparison of Eqs. (1.20) and (1.21) with Eq. (1.14) shows that R_0 is an exponential-bound parameter of nearly the form of $E_b(R)$. In fact, R_0 was first derived as a lower bound on $E_b(R)$ (see Gallager [49] and Wozencraft and Kennedy [186]). In the

development and analysis of powerful decoding algorithms for long-constraint-length convolutional codes, R_0 has been given a more descriptive name, the *computational cutoff rate*, denoted by R_{COMP} (see Chapters 10 and 11).

As we said earlier, R_0 is a quantity similar to capacity that can be evaluated for any given channel. Mathematically, the cutoff rate for an M-input, Q-output discrete memoryless channel is given by

$$R_0 = -\log_2 \left\{ \min_{P(x_i)} \sum_{j=0}^{Q-1} \left[\sum_{i=0}^{M-1} \sqrt{P(y_j|x_i)}\, P(x_i) \right]^2 \right\} \tag{1.22}$$

where the $\{P(y_j|x_i)\}$ denote the channel transition probabilities, and the minimization is taken over all possible probability distributions $P(x_i)$ on the channel input symbols. The cutoff rate has the same units as capacity — that is, in the form given above, source information bits transferred per symbol—and it is obtained by a process of optimization as in the derivation of capacity. However, unlike capacity, R_0 is optimized not only with respect to the distribution of symbols entering the channel, but also with respect to the boundaries between decision regions at the output of the demodulator. That is, for a given number Q of demodulator outputs (quantization intervals), the decision boundaries are chosen so as to maximize the value of R_0.

Thus we see from Eqs. (1.20) and (1.21) that maximizing R_0 serves two purposes. The practical limit on the rate of information transfer is maximized, and the bound on the probability of error is minimized. The complexity of the modulation and demodulation strategy is established as part of the optimizations of input symbol set and output decision regions.

Massey and other researchers have proposed a unified theory of modulation and coding design based upon R_0 as the fundamental channel parameter and have defined an approach to the design of optimum modulation systems to be used in conjunction with efficient decoders [106]. This approach has come to be known as designing according to the "R_0 criterion." The point of this approach is that a modulation system should be designed to achieve the highest possible value of R_0 rather than the lowest value of post-demodulation error probability.

It is instructive to calculate R_0 for a relatively simple case that is of considerable practical importance, that of binary antipodal signaling on the AWGN channel. Let us say that the two channel inputs are $x_0 = -a$ and $x_1 = a$ and that they occur with equal probability $P(x_0) = P(x_1) = \frac{1}{2}$. If we assume that the channel adds a sample of Gaussian noise with variance σ^2 to each transmitted digit and that the channel outputs are unquantized, then Eq. (1.22) is rewritten as

$$R_0 = -\log \int_{-\infty}^{\infty} \left(\frac{1}{2}\sqrt{p(y|x_0)} + \frac{1}{2}\sqrt{p(y|x_1)} \right)^2 dy. \tag{1.23}$$

where the transition probabilities are given by the Gaussian conditional probability density functions

$$p(y|x_0) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-(y+a)^2/2\sigma^2} \tag{1.24}$$

and

$$p(y|x_1) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-(y-a)^2/2\sigma^2} \tag{1.25}$$

Because of the symmetry in this simple binary case, the minimization over $P(x)$ is eliminated. Also, the demodulation decision boundary to maximize R_0 is obvious: $y = 0$. The integration in Eq. (1.23) is straightforward, yielding

$$R_0 = -\log\frac{1}{2}\left(1 + e^{-a^2/2\sigma^2}\right)$$

$$= 1 - \log\left(1 + e^{-a^2/2\sigma^2}\right)$$

If we now relate this channel model to the case of binary PSK signaling in AWGN, with unquantized coherent matched-filter detection, the quantity $a^2/2\sigma^2$ corresponds to E_s/N_0, where E_s is the signal energy per PSK pulse and N_0 is the one-sided noise spectral density. We can then write the cutoff rate for this channel as

$$R_0 = 1 - \log_2(1 + e^{-E_s/N_0}) \tag{1.26}$$

which has units of bits per channel symbol. Note that code rate is measured in the same units.

We would now like to determine the form of R_0 in the power-limited region (Fig. 1.10), where, with unlimited bandwidth, coding can be used to reduce the required signal energy to the lowest possible level. To do this we assume the use of a code with rate equal to R_0 and rewrite E_s/N_0 as

$$\frac{E_s}{N_0} = R_0 \frac{E_b}{N_0}$$

where E_b is the signal energy per source information bit. Now, letting the rate R_0 become very small, we rewrite Eq. (1.26) as

$$R_0 = 1 - \log_2(1 + e^{-R_0 E_b/N_0}),$$

$$\cong 1 - \log_2(2 - R_0 E_b/N_0)$$

$$\cong \frac{R_0}{2\ln 2}\left(\frac{E_b}{N_0}\right) \tag{1.27}$$

Solving Eq. (1.27) for E_b/N_0, we find that the minimum value of E_b/N_0 to

achieve the cutoff rate R_0 is

$$\frac{E_b}{N_0} = 2\ln 2 \qquad \text{or} \qquad \frac{E_b}{N_0} = 1.4 \text{ dB}$$

Referring to the discussion in Section 1.6.1, we see that this is exactly 3 dB above the Shannon limit ($\ln 2$ or -1.6 dB), the minimum value of E_b/N_0 needed to achieve channel capacity on the AWGN channel.

The reader may question this comparison, since R_0 was derived here for a binary-input AWGN channel, while capacity C (Eq. 1.13) was derived for a channel whose inputs were unconstrained except for a limit on the average signal power. If we were to compute C for the binary-input AWGN channel, we would convert Eq. (1.7) to an integral form and then use Eq. (1.8) to give

$$C = \frac{1}{2}\sum_{i=0}^{1}\int_{-\infty}^{\infty} p(y|x_i)\log\frac{p(y|x_i)}{p(y)}dy \tag{1.28}$$

with $p(y|x_0)$ and $p(y|x_1)$ given by Eqs. (1.24) and (1.25) and

$$p(y) = \frac{1}{2}p(y|x_0) + \frac{1}{2}p(y|x_1)$$

In this calculation, C has units of bits per channel symbol. For this case it is not possible to carry out the integral in Eq. (1.28) in closed form. However, it can be shown that for code rates R approaching zero, channel capacity can be accurately approximated by

$$C \cong \frac{R}{\ln 2}\left(\frac{E_b}{N_0}\right)$$

Again we set $C = R$ and find that the minimum SNR per information bit needed to achieve capacity is

$$\frac{E_b}{N_0} = \ln 2 \qquad \text{or} \qquad \frac{E_b}{N_0} = -1.6 \text{ dB}$$

This is the Shannon limit, the minimum value of E_b/N_0 found for the continuous AWGN channel. Thus in the power-limited region of operation (arbitrarily large bandwidths and code rates approaching zero), capacity is in principle achievable with binary modulation and coding.

To develop fully the concept of the R_0 criterion and discuss the extensive implications is beyond the scope of this book. However, it has served our purpose to state that by determining R_0 for any given channel, we establish an upper bound on the information transfer rate that can be achieved with practical coding implementations and also obtain an exponential bound on

post-decoding error probability. In Chapters 10 and 11 we shall see examples of error-control coding schemes that operate reliably at information transfer rates remarkably close to R_0. For schemes that operate at rates above R_0 but less than capacity, the computational complexity increases drastically. We shall also see that the particular figure of communication system performance used is not important; that is, low post-decoding error rates or high probability of correct decoding can both be provided for rates $R \leq R_0$.

1.7. SUMMARY AND CONCLUSIONS

In this introductory chapter we have outlined the problem of designing an efficient digital communication system and have reviewed the principal results of information theory in order to provide a framework for the remainder of the book. At this point it is useful to summarize the major points that have been made and to place them in their proper perspective for the sequel.

We can describe the communication system design problem succinctly as follows. We first want to remove all redundancy from the source information, so that the amount of data to be transmitted is minimized, and we also want to communicate this information reliably with the smallest possible expenditure of signal energy. The two key parameters here are the information rate R_s of the source, which is the minimum number of bits per second needed to represent the output of the source, and the channel capacity C, the maximum rate at which information can be transmitted through the channel and received reliably. The channel coding theorem provides us with the important result that if the source rate R_s does not exceed the channel capacity C, it is possible to deliver the source information with arbitrarily low probability of error. We do not concern ourselves with details of the source-coding function, that is, the reduction of a source output to a stream of bits occurring at the rate R_s. This is a large subject unto itself and is treated extensively by other authors. Rather, we assume that the information source produces a sequence of information bits completely free of redundancy and concentrate on the problem of communicating this information as reliably and efficiently as possible.

The central idea of efficient channel coding is to transform long sequences of source data bits into even longer coded channel sequences or signaling waveforms. That is, we must put well-structured redundancy back into the source data for transmission. The amount of redundancy required depends on the quality of the channel (the channel SNR or BER) and the desired level of reliability in the delivered information. At the receive side we use the known structure of the possible transmitted signals to detect and decode the output of the channel and to deliver a representation of the source data.

Information theory provides the fundamental limits on the reliability and efficiency of a digital communication system. For the important case of the AWGN channel, communication efficiency is measured by the SNR per source bit, E_b/N_0. For operation in the power-limited region, where we assume that

large bandwidth expansion can be used, the Shannon limit tells us that for all $E_b/N_0 > -1.6$ dB, arbitrarily reliable communication can theoretically be provided. Furthermore, very reliable communication at low SNRs can best be achieved by using an error-control code to construct the long channel sequences prior to modulation. On the other hand, for operation in the bandwidth-limited region, highly efficient operation can be obtained without resorting to coding, that is, by the use of complex modems that implement high-order modulation alphabets.

Thus we shall be concerned primarily with operation of a digital communication system in the power-limited region, and much of this book is devoted to the solution of two basic problems. The first is finding the best ways to transform the source information into the redundant channel sequences (the code design problem) and the second is finding ways to invert that transformation that are not unduly complex (the decoding problem). Thus, we shall be concerned primarily with finding good code designs and efficient decoding algorithms.

While channel capacity establishes the theoretical limit on the performance that can be achieved with a digital communication system, under conditions of severe signal power limitations operation at or near channel capacity may require an unacceptably complex system design. Channel capacity provides little in the way of practical design guidance. We have asserted that the SNR corresponding to the computational cutoff rate of a channel, R_0, gives us a limit that can be approached with a system design that is not unduly complex. For the case of antipodal signaling with coherent reception in AWGN, this corresponds to an SNR that is 3 dB above the Shannon limit, namely $E_b/N_0 = 1.4$ dB. In fact, in Chapters 10 and 11 we shall see examples of powerful code designs that come close to achieving this limit.

Although we have stressed the view that an improvement in communication efficiency is directly related to a reduction in the required level of signal power, there are other important practical interpretations. We have assumed that the time required for transmission of a message is held fixed, and consequently for coded operation, bandwidth expansion is needed. However, in many applications, both the available signal power and bandwidth are limited. In such situations, if highly reliable and efficient operation is to be provided, a lengthening of message transmission time is inescapable for both coded and uncoded systems. When transmission time is an adjustable parameter, comparison of alternative designs on the basis of communication efficiency is equally valid. Improvements in communication efficiency can in this case be related directly to a reduction in the required level of signal power, a reduction in transmission time, or a simultaneous reduction in both. In addition, for some applications, a reduction in required power levels to achieve the desired performance can be used to increase the effective range of the communication system.

For most applications encountered, operation near channel capacity or computational cutoff is not needed, and considerably less severe requirements

apply. We shall see in this book that error-control coding can also be used effectively at higher SNRs. There are, in fact, a number of coding techniques that provide sufficient improvements in communication efficiency to justify the added implementation complexity associated with coding. To apply error-control coding in a cost-effective design for any application, it is clearly necessary to determine both the complexity of the coding technique considered and the performance improvements that accrue. Therefore, the remainder of this book is concerned with the details of code design, encoding and decoding techniques, and the evaluation of performance of error-control-coded systems.

1.8. NOTES

The mathematical formulation of error-correcting codes was founded by R. W. Hamming, whose early work was cited in Shannon's 1948 paper [157] on the mathematical theory of communication. Apparently delayed because of patent considerations at Bell Telephone Laboratories [10], Hamming's own paper appeared in 1950 [60]. The papers of Hamming and Shannon represent, respectively, an essentially combinatorial discipline termed coding theory and an essentially statistical discipline known as information theory or *Shannon theory*.

An important area of information theory, and in particular an important aspect of the source coding problem, is that of *rate-distortion theory*. This theory addresses the question of what accuracy must be sacrificed in the delivery of information from a source to a user when the capacity of the intervening communication channel is less than the minimum information rate needed to completely reconstruct the output of the source. Much research has been devoted to this interesting subject, but a thorough discussion is beyond the scope of this book. The interested reader will find detailed treatments of the subject in books by Berger [6], Gallager [49], and Viterbi and Omura [175].

Many of the key papers in the development of information theory have been collected into an IEEE Reprint Series volume edited by Slepian [163]. Additional bibliographies and surveys of published research are also cited there. An excellent survey of the development of coding theory is provided in a companion volume edited by Berlekamp [10]. Other surveys of the field are also cited there. A book by MacWilliams and Sloane [109] provides an almost exhaustive treatment of the theory of block codes with many references to the literature up to 1981. A brief introduction to the theory of error-correcting block codes is the subject of a recent book by Pless [134]. Convolutional codes and sequential decoding are treated in detail in a text on information theory and coding by Gallager [49] as well as an earlier text by Wozencraft and Jacobs [185].