Fundamentals of Information Transmission

Partially selected from J. G. Proakis and M Salehi, "Digital Communications", 5th ed. New York: McGraw-Hill, 2007

Chenggao HAN

Contents

Chapter 1

Introduction

In this book, we present the basic principles that underlie the analysis and design of digital communication systems. The subject of digital communications involves the transmission of information in digital form from a source that generates the information to one or more destinations. Of particular importance in the analysis and design of communication systems are the characteristics of the physical channels through which the information is transmitted. The characteristics of the channel generally affect the design of the basic building blocks of the communication system. Below, we describe the elements of a communication system and their functions.

1.1 Elements of a Digital Communication System

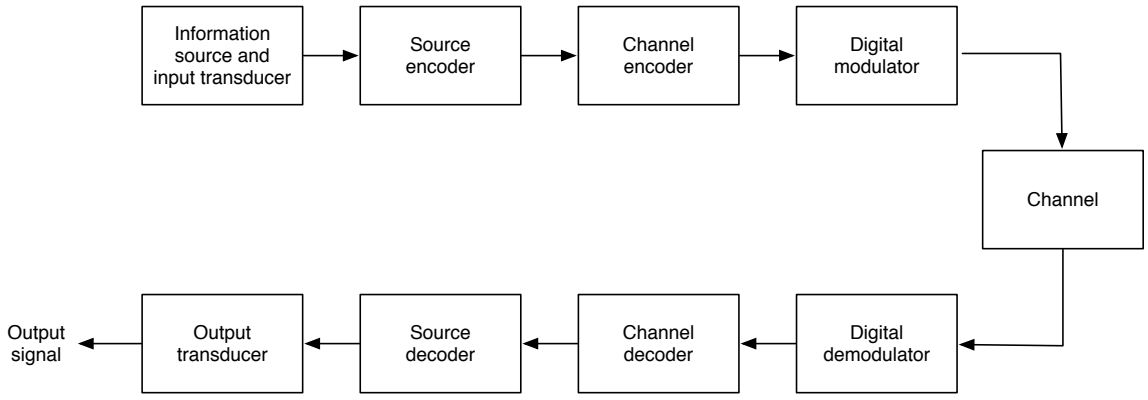

Figure 1-1: Basic elements of a digital communication system.

Figure 1-9 illustrates the functional diagram and the basic elements of a digital communication system.

- The source output may be either an analog signal, such an audio or video signal, or a digital signal, such as the output of a computer, that is discrete in time and has a finite number of output characters. In a digital communication system, the messages produced by the source are converted into a sequence of binary digits. Ideally, we should like to represent the source output (message) by as few binary digits as possible. In other words we seek

an efficient of the source that results in little or no redundancy. The process of efficiently converting the output of either an analog or digital source into a sequence of binary digits is called *source encoding* or *data compression*.

- The sequence of binary digits from the source encoder, which we call the *information sequence*, is passed to the channel encoder. The purpose of the channel encoder is to introduce, in a controlled manner, some redundancy in the binary information sequence that can be used at the receiver to overcome the effects of noise and interference encountered in the transmission of the signal through the channel. Thus the added redundancy serves to increase the reliability of the received data and improves the fidelity of the received signal. In effect, redundancy in the information sequence aids the receiver in decoding the desired information sequence. For example, a (trivial) form of encoding or the binary information sequence is simply to repeat each binary digit m times, where m is some positive integer. More sophisticated (nontrivial) encoding involves taking k information bits at time and mapping each k-bit sequence into a unique n-bit sequence, called a *code word*. The amount of redundancy introduced by encoding the data in this manner is measured by the ratio n/k. The reciprocal of this ratio, namely k/n, is called the rate of the code or, simply, the *code rate*.

- The binary sequence at the output of the channel encoder is passed to the *digital modulator*, which serves as the interface to the communication channel. Since nearly all the communication channels encountered in practice are capable or transmitting electrical signals (waveforms), the primary purpose of the digital modulator is to map the binary information sequence into signal waveforms. To elaborate on this point, let us suppose that the coded information is to be transmitted one bit at a time at some uniform rate R bits per second (bits/s). The digital modulator may simply map the binary digit 0 into a waveform $s_0(t)$ and the binary digit 1 into a waveform $s_1(t)$. In this manner, each bit from the channel encoder is transmitted separately. We call this *binary modulation*. Alternatively, the modulator may transmit b coded information bits at a time by using $M = 2^b$ distinct waveforms $s_i(t), i = 0, 1, \cdots, .M-1$, one waveform for each of the 2^b possible b-bit sequences. We call this *M-ary modulation* ($M > 2$). Note that a new b-bit sequence enters the modulator every b/R seconds. Hence, when the channel bit rate R is fixed, the amount or time available to transmit one of the M waveforms corresponding to a b-bit sequence is b times the time period in a system that uses binary modulation.

- The *communication channel* is the physical medium that is used to send the signal from the transmitter to the receiver. In wireless the channel may be the atmosphere (free space). On the other hand, telephone channels usually employ a variety of physical media, including wire lines, optical fiber cables, and wireless (microwave radio). Whatever the physical medium used for transmission of the information, the essential feature is that the transmitted signal is corrupted in a random manner by a variety of possible mechanisms, such as additive thermal noise generated by electronic devices; man-made noise, *e.g.*, automobile ignition noise; and atmospheric noise, *e.g.*, electrical lightning discharges during thunderstorms.

- At the receiving end of a digital communication system, the *digital demodulator* processes the channel-corrupted transmitted waveform and reduces the waveforms to a sequence of numbers that represent estimates of the transmitted data symbols (binary or M-ary). This sequence of numbers is passed to the channel decoder, which attempts to reconstruct the

original information sequence from knowledge of the code used by the channel encoder and the redundancy contained in the received data.

A measure of how well the demodulator and decoder perform is the frequency with which errors occur in the decoded sequence. More precisely, the average probability of a bit-error at the output of the decoder is a measure of the performance of the demodulator-decoder combination. In general, the probability of error is a function of the code characteristics, the types of waveforms used to transmit the information over the channel, the transmitter power, the characteristics of the channel (*i.e.*, the amount of noise, the nature of the interference), and the method of demodulation and decoding. These items and their effect on performance will be discussed in detail in subsequent chapters.

- As a final step, when an analog output is desired, the source decoder accepts the output sequence from the channel decoder and, from knowledge of the source encoding method used, attempts to reconstruct the original signal from the source. Because of channel decoding errors and possible distortion introduced by the source encoder, and perhaps, the source decoder, the signal at the output of the source decoder is an approximation to the original source output. The difference or some function of the difference between the original signal and the reconstructed signal is a measure of the distortion introduced by the digital communication system.

1.2 Communication Channels and Their Characteristics

As indicated in the preceding discussion, the communication channel provides the connection between the transmitter and the receiver. The physical channel may be a pair of wires that carry the electrical signal, or an optical fiber that carries the information on a modulated light beam, or an underwater ocean channel in which the information is transmitted acoustically, or free space over which the information-bearing signal is radiated by use of an antenna. Other media that can be can be characterized as communication channels are data storage media, such as magnetic tape, magnetic disks, and optical disks.

One common problem in signal transmission through any channel is additive noise. In general, additive noise is generated internally by components such as resistors and solid-state devices used to implement the communication system. This is sometimes called *thermal noise*. Other sources of noise and interference may arise externally to the system, such as interference from other users of the channel. When such noise and interference occupy the same frequency band as the desired signal, their effect can be minimized by the proper design of the transmitted signal and its demodulator at the receiver. Other types of signal degradations that may be encountered in transmission over the channel are signal attenuation, amplitude and phase distortion, and multipath distortion.

The effects of noise may be minimized by increasing the power in the transmitted signal. However, equipment and other practical constraint limit the power level in the transmitted signal. Another basic limitation is the available channel bandwidth. A bandwidth constraint is usually due to the physical limitations of the medium and the electronic components used to implement the transmitter and the receiver. These two limitations constrain the amount of data that can be transmitted reliably over any communication channel as we shall observe in later chapters. Below, we describe some of the important characteristics or several communication channels.

Wireline Channels

The telephone network makes extensive use of wire lines for voice signal transmission, as well as data and video transmission. Twisted-pair wire lines and coaxial cable are basically guided electromagnetic channels that provide relatively modest bandwidths. Telephone wire generally used to connect a customer to a central office has a bandwidth or several hundred kilohertz (kHz). On the other hand. coaxial cable has a usable bandwidth or several megahertz (MHz). Figure 1-2 illustrates the frequency range of guided electromagnetic channels, which include waveguides and optical fibers.

Signals transmitted through such channels are distorted in both amplitude and phase and further corrupted by additive noise. Twisted-pair wireline channels are also prone to crosstalk interference from physically adjacent channels. Because wireline channels carry a large percentage of our daily communications around the country and the world, much research has been performed on the characterization of their transmission properties and on methods for mitigating the amplitude and phase distortion encountered in signal transmission. In Chapter 4, we describe methods for designing optimum transmitted signals and their demodulation; in Chapter **??**, we consider the design of channel equalizers that compensate for amplitude and phase distortion on these channels.

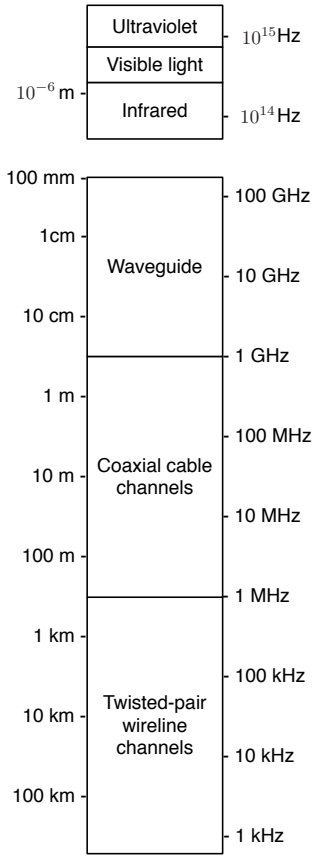

Figure 1-2: Frequency range for guided wire channel.

Fiber-Optic Channels

Optical fibers offer the communication system designer a channel bandwidth that is several orders of magnitude larger than coaxial cable channels. During the past two decades, optical fiber cables have been developed that have a relatively low signal attenuation, and highly reliable photonic devices have been developed for signal generation and signal detection. These technological advances have resulted in a rapid deployment of optical fiber channels, both in domestic telecommunication systems as well as for transcontinental communication. With the large bandwidth available on fiber-optic channels, it is possible for telephone companies to offer subscribers a wide array of telecommunication services, including voice, data, facsimile, and video.

The transmitter or modulator in a fiber-optic communication system is a light source, either a light-emitting diode (LED) or a laser. Information is transmitted by varying (modulating) the intensity of the light source with the message signal. The light propagates through the fiber as a light wave and is amplified periodically (in the case of digital transmission, it is detected and regenerated by repeaters) along the transmission path to compensate for signal attenuation. At the receiver, the light intensity is detected by a photodiode, whose output is an electrical signal that varies in direct proportion to the power of the light impinging on the photodiode. Sources of noise in fiber-optic channels are photodiodes and electronic amplifiers.

Wireless Electromagnetic Channels

In wireless communication systems, electromagnetic energy is coupled to the propagation medium by an antenna which serves as the radiator. The physical size and the configuration of the antenna depend primarily on the frequency of operation. To obtain efficient radiation of electromagnetic energy, the antenna must be longer than 1/10 of the wavelength. Consequently, a radio station transmitting in the amplitude-modulated (AM) frequency band, say at $f_c = 1$ MHz [corresponding to a wavelength of $\lambda = c/f_c = 300$ meters (m)], requires an antenna of at least 30 m. Other important characteristics and attributes of antennas for wireless transmission are described in Chapter 3.

Figure 1-3 illustrates the various frequency bans of the electromagnetic spectrum. The mode

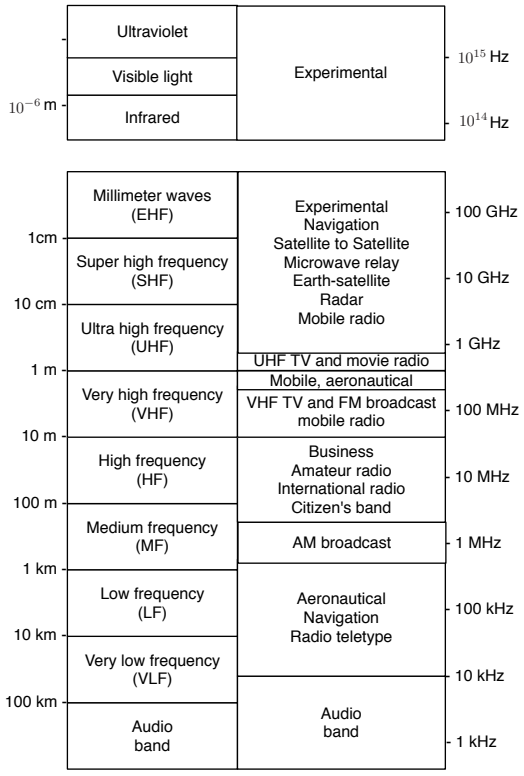

Figure 1-3: Frequency range for wireless electromagnetic channels. [Adapted from Carlson (1975). 2nd edition. ©McGraw-Hill Book Company Co. Reprinted with permission of the publisher.]

of propagation of electromagnetic waves in the atmosphere and in free space may be subdivided into three categories, namely, ground-wave propagation, sky-wave propagation, and line-of-sight (LOS) propagation. In the very low frequency (VLF) and audio frequency bands, where the wavelengths exceed 10 km, the earth and the ionosphere act as a waveguide for electromagnetic wave propagation. In these frequency ranges, communication signals practically propagate around the globe. For this reason, these frequency bands are primarily used to provide navigational aids from shore to ships around the world. The channel bandwidths available in these frequency bands are relatively small (usually 1-10 percent of the center frequency), and hence the information that is transmitted through these channels is of relatively slow speed and generally confined to digital transmission. A dominant type of noise at these frequencies is generated from thunderstorm

activity around the globe, especially in tropical regions. Interference results from the many users of these frequency bands.

Ground-wave propagation, as illustrated in Figure 1-4, is the dominant mode of propagation for frequencies in the medium frequency (MF) band (0.3-3 MHz). This is the frequency band used for AM broadcasting and maritime radio broadcasting. In AM broadcasting, the range with ground-wave propagation of even the more powerful radio stations is limited to about 150 km. Atmospheric noise, man-made noise, and thermal noise from electronic components at the receiver are dominant disturbances for signal transmission in the MF band.

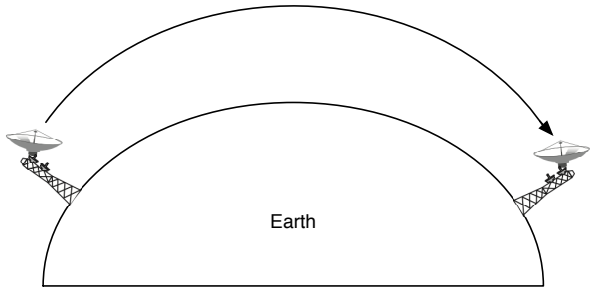

Figure 1-4: Illustration of ground-wave propagation.

Sky-wave propagation, as illustrated in Figure 1-5, results from transmitted signals being reflected (bent or refracted) from the ionosphere, which consists of several layers of charged particles ranging in altitude from 50 to 400 km above the surface of the earth. During the daytime hours, the heating of the lower atmosphere by the sun causes the formation of the lower layers at altitudes below 120 km. These lower layers, especially the D-layer, serve to absorb frequencies below 2 MHz, thus severely limiting sky-wave propagation of AM radio broadcast. However, during the nighttime hours, electron density in the lower layers of the ionosphere drops sharply and the frequency absorption that occurs during the daytime is significantly reduced. As a consequence, powerful AM radio broadcast stations can propagate over large distances via sky wave over the F-layer of the ionosphere, which ranges from 140 to 400 km above the surface of the earth.

A frequently occurring problem with electromagnetic wave propagation via sky wave in the high frequency (HF) range is *signal multipath*. Signal multipath occurs when the transmitted signal arrives at the receiver via multiple propagation paths at different delays. It generally results in intersymbol interference in a digital communication system. Moreover, the signal components arriving via different propagation path may add destructively, resulting in a phenomenon called *signal fading*, which most people have experienced when listening to a distant radio station at night when sky wave is the dominant propagation mode. Additive noise in the HF range is a combination of atmospheric noise and thermal noise.

Sky-wave ionospheric propagation ceases to exist at frequencies above approximately 30 MHz, which is the end of the HF band. However, it is possible to have ionospheric scatter propagation

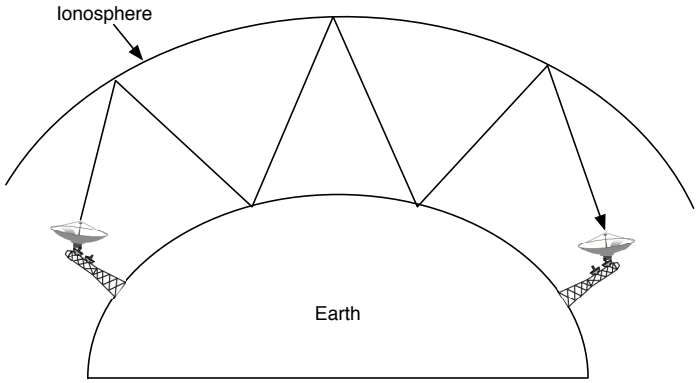

Figure 1-5: Illustration of sky-wave propagation.

at frequency in the range 30-60 MHz, resulting from signal scattering from the lower ionosphere. It is also possible to communicate over distances of several hundred miles by use or tropospheric scattering at frequencies in the range 40-300 MHz. Troposcatter results from signal scattering due to particles in the atmosphere at altitudes of 10 miles or less. Generally, ionospheric and tropospheric scatter involve large signal propagation losses and require a large amount of transmitter power and relatively large antennas.

Frequencies above 30 MHz propagate through the ionosphere with relatively little loss and make satellite and extraterrestrial communications possible. Hence, at frequencies in the very high frequency (VHF) band and higher, the dominant mode or electromagnetic propagation is LOS propagation. For terrestrial communication systems, this means that the transmitter and receiver antennas must be in direct LOS with relatively little or no obstruction. For this reason. television stations transmitting in the VHF and ultra high frequency (UHF) bands mount their antennas on high towers to achieve a broad coverage area.

In general, the coverage area for LOS propagation is limited by the curvature of the earth. If the transmitting antenna is mounted at a height h m above the surface of earth, distance to the radio horizon, assuming no physical obstructions such as mountains, is approximately $d = \sqrt{15h}$ km. For example, a television antenna mounted on a tower of 300 m in height provides a coverage of approximately 67 km. As another example, microwave radio relay systems used extensively for telephone and video transmission at frequencies above 1 gigahertz (GHz) have antennas mounted on tall towers or on the top of tall buildings.

The dominant noise limiting the performance of a communication system in VHF and UHF ranges is thermal noise generated in the receiver front end and cosmic noise picked up by the antenna. At in the super high frequency (SHF) band above 10 GHz, atmospheric conditions play a major role in signal propagation. For example, at 10 GHz, the attenuation ranges from about 0.003 decibel per kilometer (dB/km) in light rain to about 0.3 dB/km in heavy rain. At 100 GHz, the attenuation ranges from about 0.1dB/km in light rain to about 6dB/km in heavy rain.

Hence, in this frequency range, heavy rain introduces extremely high propagation losses that can result in service outages (total breakdown in the communication system).

At frequencies above the extremely high frequency (EHF) band, we have the infrared and visible light regions of the electromagnetic spectrum, which can be used to provide LOS optical communication in free space. To date, these frequency bands have been used in experimental communication systems, such as satellite-to-satellite links.

Underwater Acoustic Channels

Over the past few decades, ocean exploration activity has been steadily increasing. Coupled with this increase is the need to transmit data, collected by sensors placed under water, to the surface of the ocean. From there, it is possible to relay the data via a satellite to a data collection center.

Electromagnetic waves do not propagate over long distances under water except at extremely low frequencies. However, the transmission of signals at such low frequencies is prohibitively expensive because of the large and powerful transmitters required. The attenuation of electromagnetic waves in water can be expressed in terms of the *skin depth*, which is the distance a signal is attenuated by $1/e$. For seawater, the skin depth $\delta = 250/\sqrt{f}$, where f is expressed in Hz and δ is in m. For example, at 10 kHz, the skin depth is 2.5 m. In contrast, acoustic signals propagate over distances of tens and even hundreds of kilometers.

An underwater acoustic channel is characterized as a multipath channel due to signal reflections from the surface and the bottom of the sea. Because of wave motion, the signal multipath components undergo time-varying propagation delays that result in signal fading. In addition, there is frequency-dependent attenuation, which is approximately proportional to the square of the signal frequency. The sound velocity is nominally about 1500 m/s, but the actual value will vary either above or below the nominal value depending on the depth at which the signal propagates.

Ambient ocean acoustic noise is caused by shrimp, fish, and various mammals. Near harbors, there is also man-made acoustic noise in addition to the ambient noise. In spite of this hostile environment, it is possible to design and implement efficient and highly reliable underwater acoustic communication systems for transmitting digital signals over large distances.

Storage Channels

Information storage and retrieval systems constitute a very significant part of data-handling activities on a daily basis. Magnetic tape, including digital audiotape and videotape, magnetic disks used for storing large amounts of computer data, optical disks used for computer data storage, and compact disks are examples of data storage systems that can be characterized as communication channels. The process of storing data on a magnetic tape or a magnetic or optical disk is equivalent to transmitting a signal over a telephone or a radio channel. The readback process and the signal processing involved in storage systems to recover the stored information are equivalent to the functions performed by a receiver in a telephone or radio communication system to recover the transmitted information.

Additive noise generated by the electronic components and interference from adjacent tracks is generally present in the readback signal of a storage system, just as is the case in a telephone or a radio communication system.

The amount of data that can be stored is generally limited by the size of the disk or tape and the density (number of bits stored per square inch) that can be achieved by the write/read electronic systems and heads. For example, a packing density of 10^9 bits per square inch has

been demonstrated in magnetic disk storage systems. The speed at which data can be written on a disk or tape and the speed at which it can be read back are also limited by the associated mechanical and electrical subsystems that constitute an information storage system.

Channel coding and modulation arc essential components of a well designed digital magnetic or optical storage system. In the readback process, the signal is demodulated and the added redundancy introduced by the channel encoder is used to correct errors in the readback signal.

1.3 Mathematical Models for Communication Channels

In the design of communication systems for transmitting information through physical channels, we find it convenient to construct mathematical models that reflect the most important characteristics of the transmission medium. Then, the mathematical model for the channel is used in the design of the channel encoder and modulator at the transmitter and the demodulator and channel decoder at the receiver. Below, we provide a brief description of the channel models that are frequently used to characterize many of the physical channels that we encounter in practice.

The Additive Noise Channel

The simplest mathematical model for a communication channel is the additive noise channel, illustrated in Figure 1-6. In this model. the transmitted signal $s(t)$ is corrupted an additive random noise process $n(t)$. Physically, the additive noise process may arise from electronic components and amplifiers at the receiver of the communication system or from interference encountered in transmission (as in the case of radio signal transmission).

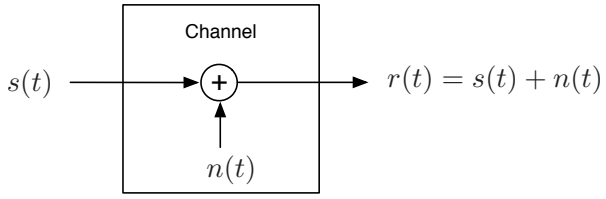

Figure 1-6: The additive noise channel.

If the noise is introduced primarily by electronic components and amplifiers at the receiver, it may be characterized as thermal noise. This type of noise is characterized statistically as a *Gaussian noise process*. Hence, the resulting mathematical model for the channel is usually called the *additive Gaussian noise channel*. Because this channel model applies to a broad class of physical communication channels and because of its mathematical tractability, this is the predominant channel model used in our communication system analysis and design. Channel attenuation is easily incorporated into the model. When the signal undergoes attenuation in transmission through the channel, the received signal is

$$r(t) = \alpha s(t) + n(t) \tag{1-1}$$

where α is the attenuation factor.

The Linear Filter Channel

In some physical channels, such as wireline telephone channels, filters are used to ensure that the transmitted signals do not exceed specified bandwidth limitations and thus do not interfere with one another. Such channels are generally characterized mathematically as linear filter channels with additive noise, as illustrated in Figure 1-7. Hence, if the channel input is the signal $s(t)$, the channel output is the signal

$$r(t) \quad = \quad s(t) * c(t) + n(t)$$

$$= \int_{-\infty}^{\infty} c(\tau)s(t-\tau)d\tau + n(t) \tag{1-2}$$

where $c(t)$ is the impulse response of the linear filter and $*$ denotes convolution.

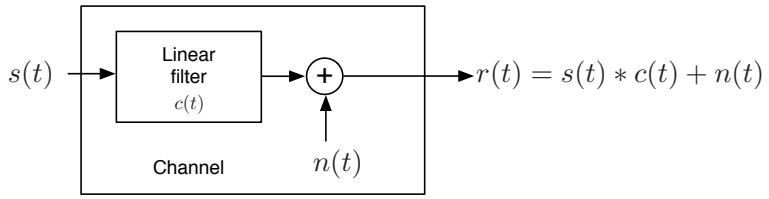

Figure 1-7: The linear filter channel with additive noise.

The Linear Time-Variant Filter Channel

Physical channels such as underwater acoustic channels and ionospheric radio channels that result in time-variant multipath propagation of the transmitted signal may be characterized mathematically as time-variant linear filters. Such linear filters are characterized by a time-variant channel impulse response $c(\tau;t)$, where $c(\tau;t)$ is the response of the channel at time t due to an impulse applied at time $t-\tau$. Thus, τ represents the "age" (elapsed-time) variable. The linear time-variant filter channel with additive noise is illustrated in Figure 1-8. For an input signal $s(t)$, the channel output signal is

$$\begin{aligned} r(t) &= s(t) * c(\tau;t) + n(t) \\ &= \int_{-\infty}^{\infty} c(\tau;t)s(t-\tau)d\tau + n(t) \end{aligned} \tag{1-3}$$

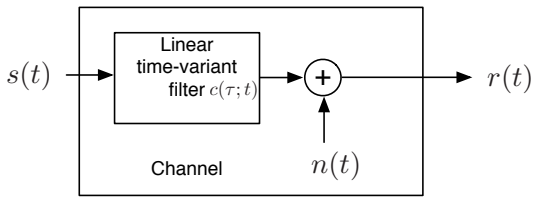

Figure 1-8: Linear time-variant filter channel with additive noise.

A good model for multipath signal propagation through physical channels, such as the ionosphere (at frequencies below 30MHz) and mobile cellular radio channels, is a special case of (1-3) in which the time-variant impulse response has the form

$$c(\tau;t) = \sum_{k=1}^{L} a_k(t)\delta(\tau - \tau_k) \tag{1-4}$$

where the $\{a_k(t)\}$ represents the possibly time-variant attenuation factors for the L multipath propagation paths and $\{\tau_k\}$ are the corresponding time delays. If (1-4) is substituted into (1-3),

the received signal has the form

$$r(t) = \sum_{k=1}^{L} a_k(t)s(t - \tau_k) + n(t) \tag{1-5}$$

Hence, the received signals of L multipath components, where the kth component is attenuated by $a_k(t)$ and delayed by τ_k.

The three mathematical models described above adequately characterize the great majority of the physical channels encountered in practice. These three channel models are used in this text for the analysis and design of communication systems.

1.4 A Historical Perspective in the Development of Digital Communications

It is remarkable that the earliest form of electrical communication, namely *telegraphy*, was a digital communication system. The electric telegraph was developed by Samuel Morse and was demonstrated in 1837. Morse devised the variable-length binary code in which letters of the English alphabet are represented by a sequence of dots and dashes (code words). In this code, more frequently occurring letters are represented by short code words, while letters occurring less frequently are represented by longer code words. Thus, the *Morse code* was the precursor of the variable-length source coding methods described in Chapter **??**.

Nearly 40 years later, in 1875, Emile Baudot devised a code for telegraphy in which every letter was encoded into fixed-length binary code words of length 5. In the *Baudot code*, binary code elements are of equal length and designated as mark and space.

Although Morse is responsible for the development of the first electrical digital communication system (telegraphy), the beginnings of what we now regard as modern digital communications stem from the work of Nyquist (1924), who investigated the problem of determining the maximum signaling rate that can be used over a telegraph channel of a given bandwidth without intersymbol interference. He formulated a model of a telegraph system in which a transmitted signal has the general form

$$s(t) = \sum_n a_n(t)g(t - nT) \tag{1-6}$$

where $g(t)$ represents a basic pulse shape and $\{a_n\}$ is the binary data sequence of $\{\pm 1\}$ transmitted at a rate of $1/T$ bits/s. Nyquist set out to determine the optimum pulse shape that was band-limited to W Hz and maximized the bit rate under the constraint that the pulse caused no intersymbol interference at the sampling time k/T, $k = 0, \pm 1, \pm 2, \cdots$. His studies led him to conclude that the maximum pulse rate is $2W$ pulses/s. This rate is now called the *Nyquist rate*. Moreover, this pulse rate can be achieved by using the pulses $g(t) = (\sin 2\pi Wt)/2\pi Wt$. This pulse shape allows recovery of the data without intersymbol interference at the sampling instants. Nyquist's result is equivalent to a version of the sampling theorem for band-limited signals, which was later stated precisely by Shannon (I948b). The sampling theorem states that a signal of bandwidth W can be reconstructed from samples taken at the Nyquist rate of $2W$ samples/s using the interpolation formula

$$s(t) = \sum_n s\left(\frac{n}{2W}\right) \frac{\sin\left[2\pi W(t - n/2W)\right]}{2\pi W(t - n/2W)} \tag{1-7}$$

In light of Nyquist's work, Hartley (1928) considered the issue of the amount of data that can be transmitted reliably over a band-limited channel when multiple amplitude levels are used. Because of the presence of noise and other interference, Hartley postulated that the receiver can reliably estimate the received signal amplitude to some accuracy, say A_δ. This investigation led Hartley to conclude that there is a maximum data rate that can be communicated reliably over a band-limited channel when the maximum signal amplitude is limited to $A_{\max}$ (fixed power constraint) and the amplitude resolution is A_δ.

Another significant advance in the development of communications was the work of Kolmogorov (1939) and Wiener (1942), who considered the problem of estimating a desired waveform $s(t)$ in the presence of additive noise $n(t)$, based on observation of the signal $r(t) = s(t) + n(t)$. This problem arises in signal demodulation. Kolmogorov and Wiener determined the linear filter

whose output is the best mean-square approximation to the desired signal $s(t)$. The resulting filter is called the *optimum linear (Kolmogorov-Wiener) filter.*

Hartley's and Nyquist's results on the maximum transmission rate of digital information were precursors to the work of Shannon (1948a,b), who established the mathematical foundations for information transmission and derived the fundamental limits for digital communication systems. In his pioneering work, Shannon formulated the basic problem of reliable transmission of information in statistical terms, using probabilistic models for information sources and communication channels. Based on such a statistical formulation, he adopted a logarithmic measure for the information content of a source. He also demonstrated that the effect of a transmitter power constraint, a bandwidth constraint, and additive noise can be associated with the channel and incorporated into a single parameter, called the *channel capacity.* For example, in the case of an additive white (spectrally flat) Gaussian noise interference, an ideal band-limited channel of bandwidth W has a capacity C given by

$$C = W \log_2 \left(1 + \frac{P}{WN_0}\right) \text{ bits/s} \tag{1-8}$$

where P is the average transmitted power and N_0 is the power spectral density of the additive noise. The significance of the channel capacity is as follows: If the information rate R from the source is less than $C(R < C)$, then it is theoretically possible to achieve reliable (error-free) transmission through the channel by appropriate coding. On the other hand, if $R > C$, reliable transmission is not possible regardless of the amount of signal processing performed the transmitter and receiver. Thus, Shannon established basic limits on communication of information and gave birth to a new field that is now called *information theory.*

Another important to the field of digital communication is the work of Kotelnikov (1947), who provided a coherent analysis of various digital communication systems based on a geometrical approach. Kotelnikov's approach was later expanded by Wozencraft and Jacobs (1965).

Following Shannon's publications came the classic work of Hamming (1950) on error-detecting and error-correcting codes to combat the detrimental effect of channel noise. Hamming's work stimulated many researchers in the years that followed, and a variety of new and powerful codes where discovered, many of which are used today in the implementation of modern communication systems.

The increase in demand for data transmission during the last four decades, coupled with the development of more sophisticated integrated circuit, has led to the development of very efficient and more reliable digital communication systems. In the course of these developments, Shannon's original results and the generalization of this results on maximum transmission limils over a channel and on bounds on the performance achieved have served as benchmarks for any given communication design. The theoretical limits derived by Shannon and other researchers that contributed to the development of information theory serve as an ultimate goal in the continuing efforts to design and develop more efficient digital communication systems.

There have been many new advances in the area of digital communications following the early work of Shannon, Kotelnikov, and Hamming. Some of the notable advances are the following:

- The development or new block codes by Muller (1954), Reed (1954). Reed and Solomon (1960). Bose and Ray-Chaudhuri (I960a,b), and Goppa (1970, 1971).
- The development of concatenated codes by Forney (1966a).
- The development of computationally efficient decoding of Bose-Chaudhuri- Hocquenghem (BCH) codes, *e.g.*, the Berlekamp-Massey algorithm (see Chien, 1964; Berlekamp, 1968).

- The development of convolutional codes and decoding algorithms by Wozencraft and Reiffen (1961), Fano (1963), Zigangirov (1966), Jelinek (1969), Forney (1970b, 1972, 1974), and Viterbi (1967, 1971).
- The development of trellis-coded modulation by Ungerboeck (1982), Forney *et al.* (1984), Wei (1987), and others.
- The development of efficient source encodings algorithms for data compression, such as those devised by Ziv and Lempel (1977, 1978), and Linde *et al.* (1980).
- The development of low-density parity check (LDPC) codes and the sum-product decoding algorithm by Gallager (1963) .
- The development or turbo codes and iterative decoding by Berrou *et al.* (1993).

1.5 Elements of A Modern Communication System

The subject of a communication system is the transmission of information from a source that generates the information to one or more desired destinations as accuratelly as is possible. Of particular importance in the analysis and design of communication systems are the characteristics of the physical channels through which the information is transmitted. The characteristics of the channel generally affect the design of the basic building blocks of the communication system. Nowadays, the mediums used for information transmission are the cables which can carry an electrical or optical signals and electromagnet since those allow the signal transmission with the velocity of light. In order to transmit the information accuratelly, a modern communication system is carefully designed and based on the communication theory and involves many digital processings on each bulding block. Below, we describe the elements of a modern communication system and their functions.

Figure 1-9 illustrates the functional diagram and the basic elements of a modern communication system, also called *digital communication system*.

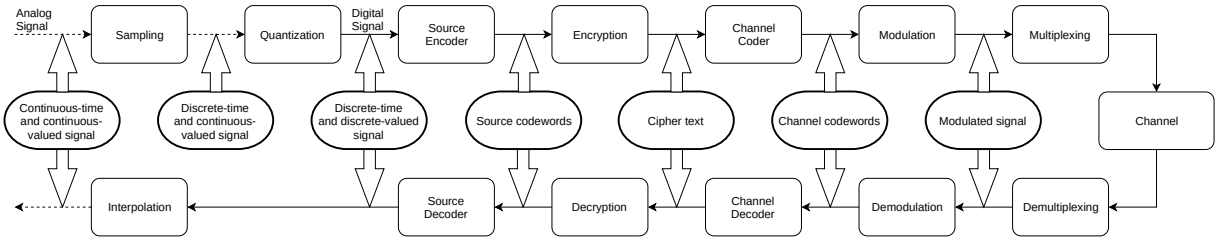

Figure 1-9: Basic elements of a modern communication system.

- Generally, the source to transmit is an analog signal, such an audio or video signal, which commonly represented by electrical voltages varing with time. Here, the term *analog signal* means that the signal takes values from a infinite set, usually from a subset of the set of real number $\mathbb{R}$, while the set a digital signal takes values has finite members. An original audio signal is analog since it is continuous in time and value, even their intervals are finite. The followings are some examples of sets for analog and digital signals.

  – analog: $\mathbb{R}$, [0 1], $\mathbb{Z}$
  – digital: $\{0, 1\}$, $\{\pi/2, \pi\}$

  In early communication systems, an analog signal is directly transmitted over a communication cable, such as voice telephone line. However, while the transmitted analog signal is distorted in the transmission in a random manner, we can not recover the original signal from the received signal. If the signal is transmitted in a digital form, on the other hand, we may eliminate impassible candidates to correct some errors occurred during transmission. In contrast, the main disadvantage of the digital communication system is complexity. To transmit the information efficiently and to correct the errors, we need to design the transmitting signal carefully and perform a complex processing for the received signal. As the hardware aspect, the progression of microelectronics allows the receiver performs a complex task in a short time. On the other hand, the progression of data compression, channel

coding, and signal processing technologies provided software soluiton to overcome the difficulties of the digital communication. By this reason, the most modern communication systems transmit information in a digital form.

- While we transmit a analog signal over a digital communication system, the first process is called *sampling*, the discretization of the signal in time domain. Let $x(t)$ be the analog signal for $t \in [0\ T]$. Then, the task at this step is to generate a signal serial x_i, $i = 0, 1, \cdots$ from $x(t)$. Assuming we generate the signal serial with evenly spaced time interval T_S as $x_i = x(iT_S)$, obviously, the shorter we set T_S, the more accurate x_i presents $x(t)$ and the larger number of signals need to transmit. The question naturally arises: Can we recover the origial signal $x(t)$ from x_i? If it is possible, what is the maximum value of T_S and how to interpolate the unsampled signal from x_i? The following answer is given by Nyquist and is known as *sampling theorem*.

  If a function $x(t)$ contains no frequencies higher than B hertz, it can be completely recovered from a series sampled by $T_S = 1/(2B)$ using sinc **function interpolation.**

- Let $x_{\min}$ and $x_{\max}$ be the minimum and maximum values of $x(t)$, $0 \le t \le T$. Then, we have $x_i \in [x_{\min}\ x_{\max}]$ and it still is an analog signal since the output falls into a set with infinite numbers. In order to transmit the sampled signal using a digital communication, the putput of the sampler is further inputed into *quantizer*. The quantizer discretizes the set $[x_{\min}\ x_{\max}]$ into a finite set $\mathcal{Y} = \{y_0, y_1, \cdots, y_{J-1}\}$ and outputs the closest number from $\mathcal{Y}$ for each x_i. Unlike the previous sampling, we can not recover the sampled signal from the output of the quantizer. The difference between input and recovered signal is called the *quantization noise*. A counterpart to suppress the quantization noise is the increment of the quantizer memory. If a linear quantizer, in which $[x_{\min}\ x_{\max}]$ is divided evenly, is utilized, the signal-to-noise ratio improves 6dB per memory bit. To recover the transmitted $y_j \in \mathcal{Y}$, the set $\mathcal{Y}$ is shared with the receiver. For a qantized signal y_j, insteed of the value tself, the transmitter transmits the index j to the receiver since y_j may be an irrational number which can not be transmitted over digital communication system.

- The input of a digital communication system is a digital signal that is discrete in time and has a finite number of output characters, such as serial of integers, called *messages*. The messages produced by the source are converted into a sequence of binary digits. Ideally, we should like to represent the source output (message) by as few binary digits as possible. In other words we seek an efficient of the source that results in little or no redundancy. The process of efficiently converting the output of either an analog or digital source into a sequence of binary digits is called *source encoding* or *data compression*. While the redundancies are completely removed from the messages, the output is a binary information sequence whose whole elements have the *independent idendical distribution* (i.i.d.) with probability 1/2.

- To prevent some undesired users intercept the transmitted messages, in a secure communication system, the information sequence is further encrypted using a key. The encryption is classified into the secret and public key cryptosystems. In the secret key cryptosystem, the key generated at the transmitter has to delivered to the desired users in advance through a secure channel. The public key cryptosystem is proposed to resolve the challenge that how to obtain a secure channel for the secret key deliver. In the public key cryptosystem, the receiver generates the private and public keys and disclose the public key to the public.

Transmitter encrypts the messages using the disired user's public key and sends the resultant ciphertext over a unsecure channel. The private key is used for decoding ciphertext and designed to be unguessable from the public key.

- We assume the ciphertext is passed to the channel encoder in a binary sequence form. The purpose of the channel encoder is to introduce, in a controlled manner, some redundancy in the ciphertext that can be used at the receiver to overcome the effects of noise, interference, and fading encountered in the transmission of the signal through the channel. Thus the added redundancy serves to increase the reliability of the received data and improves the fidelity of the received signal. In effect, redundancy in the information sequence aids the receiver in decoding the desired information sequence. For example, a (trivial) form of encoding or the binary information sequence is simply to repeat each binary digit several times. More sophisticated (nontrivial) encoding involves taking K information bits at time and mapping each K-bit sequence into a unique N-bit sequence, called a *code word*. The amount of redundancy introduced by encoding the data in this manner is measured by the ratio N/K. The reciprocal of this ratio, namely K/N, is called the rate of the code or, simply, the *code rate*. The channel decoder attempts to reconstruct the original binary sequence from knowledge of the redundancy of code used by the channel encoder based on the corrupted data inputed from demodulator.

- The binary sequence at the output of the channel encoder is passed to the *digital modulator*, which serves as the interface to the communication channel. Since nearly all the communication channels encountered in practice are capable or transmitting electrical signals (waveforms), the primary purpose of the digital modulator is to map the binary information sequence into signal waveforms. To elaborate on this point, let us suppose that the coded information is to be transmitted one bit at a time at some uniform rate R bits per second (bits/s). The digital modulator may simply map the binary digit 0 into a waveform $s_0(t)$ and the binary digit 1 into a waveform $s_1(t)$. In this manner, each bit from the channel encoder is transmitted separately. We call this *binary modulation*. Alternatively, the modulator may transmit Q coded information bits at a time by using $M = 2^Q$ distinct waveforms $s_i(t), i = 0, 1, \cdots, .M - 1$, one waveform for each of the 2^Q possible Q-bit sequences. We call this *M-ary modulation* ($M > 2$). Note that a new Q-bit sequence enters the modulator every Q/R seconds. Hence, when the channel bit rate R is fixed, the amount or time available to transmit one of the M waveforms corresponding to a Q-bit sequence is Q times the time period in a system that uses binary modulation. The *digital demodulator* processes the channel-corrupted transmitted waveform and reduces the waveforms to a sequence of numbers that represent estimates of the transmitted data symbols (binary or M-ary).

- While a public channel is used for the information transmission, the modulated signal is multiplexed with other user's signals. Although the details of multiplexing depend on the type of public channel, *e.g.*wireless, telephone, and internet *etc.*, time, frequencies, and spreading codes are common resources what used to distinguish users. If a resource is assigned to a user, the leaked signal from other resources at receiver is called *interference*. How to suppress or cancel interference is the main task of the (de)multiplexing processes.

- The *communication channel* is the physical medium that is used to send the signal from the transmitter to the receiver. In wireless the channel may be the atmosphere (free space). On the other hand, telephone channels usually employ a variety of physical media, including wire lines, optical fiber cables, and wireless (microwave radio). Whatever the physical

medium used for transmission of the information, the essential feature is that the transmitted signal is corrupted in a random manner by a variety of possible mechanisms, such as additive thermal noise generated by electronic devices; man-made noise, *e.g.*, automobile ignition noise; and atmospheric noise, *e.g.*, electrical lightning discharges during thunderstorms.

The performance of a digital communication is usually evaluated for the cipher text under assumption of i.i.d. binary sequence input. A measure of how well the demultiplexing, demodulator, and decoder perform is the frequency with which errors occur in the decoded sequence. More precisely, the average probability of a bit-error at the output of the decoder is a measure of the performance of the demultiplexing-demodulator-decoder combination. In general, the probability of error is a function of the code characteristics, the types of waveforms used to transmit the information over the channel, the utilized multiplexing method, the transmitter power, the characteristics of the channel (*i.e.*, the amount of noise, the nature of the interference), and the method of demultiplexing, demodulation, and decoding.

Chapter 2

Matrices and Random Processes

2.1 Matrices

A matrix is a rectangular array of real or complex numbers called the *elements of the matrix*. An $n \times m$ matrix has n rows and m columns. If $m = n$, the matrix is called a *square matrix*. An vector may be viewed as an $n \times 1$ matrix. An $n \times m$ matrix may be viewed as having m n-dimensional vectors as its rows or m n-dimensional vectors as its columns.

The complex *conjugate* and the *transpose* of a matrix $\boldsymbol{A}$ are denoted as $\boldsymbol{A}^*$ and $\boldsymbol{A}^T$, respectively. The *conjugate transpose* of a matrix with complex elements is denoted as $\boldsymbol{A}^H$; that is, $\boldsymbol{A}^H = [\boldsymbol{A}^*]^T = \left[\boldsymbol{A}^T\right]^*$.

A square matrix $\boldsymbol{A}$ is said to be *symmetric* if $\boldsymbol{A}^T = \boldsymbol{A}$. A square matrix $\boldsymbol{A}$ with complex elements is said to be *Hermitian* if $\boldsymbol{A}^H = \boldsymbol{A}$. If $\boldsymbol{A}$ is a square matrix, then $\boldsymbol{A}^{-1}$ designates the *inverse* of $\boldsymbol{A}$ (if one exists), having the property that

$$\boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{I}_n \tag{2-1}$$

where $\boldsymbol{I}_n$ is the $n \times n$ *identity* matrix, *i.e.*, a square matrix whose diagonal elements are unity and off-diagonal elements are zero. If $\boldsymbol{A}$ has no inverse, it is said to be *singular*.

The *trace* of a square matrix $\boldsymbol{A}$ is denoted as $\mathrm{tr}(\boldsymbol{A})$ and is defined as the sum of the diagonal elements, *i.e.*,

$$\mathrm{tr}(\boldsymbol{A}) = \sum_{i=1}^{n} a_{ii} \tag{2-2}$$

The *rank* of an $n \times m$ matrix $\boldsymbol{A}$ is the maximum number of linearly independent columns or rows in the matrix (it makes no difference whether we take rows or columns). A matrix is said to be of *full rank* if its rank is equal to the number of rows or columns, whichever is smaller.

The following are some additional matrix properties (lowercase letters denote vectors):

$$\begin{array}{rclcrcl} (\boldsymbol{A}\boldsymbol{v})^T &=& \boldsymbol{v}^T\boldsymbol{A}^T & & (\boldsymbol{A}\boldsymbol{B})^{-1} &=& \boldsymbol{B}^{-1}\boldsymbol{A}^{-1} \\ (\boldsymbol{A}\boldsymbol{B})^T &=& \boldsymbol{B}^T\boldsymbol{A}^T & & (\boldsymbol{A}^T)^{-1} &=& \left(\boldsymbol{A}^{-1}\right)^T \end{array} \tag{2-3}$$

Linear dependence and independence

A set of vectors $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_K\}$ in a vector space is said to be *linearly dependent* if there exist coefficients $\{a_1, \cdots, a_K\}$, not all 0, in the underlying scalar field $\mathbb{F}$ such that

$$a_1\boldsymbol{x}_1 + a_2\boldsymbol{x}_2 + \cdots + a_K\boldsymbol{x}_K = 0$$

Equivalently, one of the $\boldsymbol{x}_i$ terms is a *linear combination*, with coefficients from $\mathbb{F}$, of the others. For example

$$\left\{[1,2,3]^T, [1,0,-1]^T, [2,2,2]^T\right\}$$

is a linearly dependent set in $\mathbb{R}^3$. A subset of $\mathcal{V}$ that is not linearly dependent over $\mathbb{F}$ is said to be *linearly independent*. For example,

$$\left\{[1,2,3]^T, [1,0,-1]^T\right\}$$

is a linearly independent set in $\mathbb{R}^3$. It is important to note that both concepts intrinsically pertain to *sets* of vectors. Any subset of a linearly independent set is linearly independent; $\{\mathbf{0}\}$ is a linearly dependent set; and hence any set which includes the $\mathbf{0}$ vector is linearly dependent. It can happen that a set of vectors is linearly dependent, while any proper subset of it is linearly independent.

2.2 Eigenvalues and Eigenvectors of a Matrix

Constrained extrema and eigenvalues

Nonzero vectors $\boldsymbol{v}$ such that $\boldsymbol{A}\boldsymbol{v}$ is a multiple of $\boldsymbol{v}$ play a major role in analyzing the structure of a general matrix or linear transformation, but such vectors arise in the more elementary context of maximizing (or minimizing) a real symmetric quadratic form subject to a geometric constraint:

$$\text{Maximize } \boldsymbol{v}^T\boldsymbol{A}\boldsymbol{v},\ s.t.\ \boldsymbol{v} \in \mathbb{R}^n, \boldsymbol{v}^T\boldsymbol{v} = 1$$

in which $\boldsymbol{A}^T = \boldsymbol{A}$ is given. A conventional approach to such a constrained optimization problem is to introduce the Lagrangian

$$L = \boldsymbol{v}^T\boldsymbol{A}\boldsymbol{v} - \lambda\boldsymbol{v}^T\boldsymbol{v}$$

Necessary conditions for an extremum then are

$$\Delta L = 2(\boldsymbol{A}\boldsymbol{v} - \lambda\boldsymbol{v}) = 0$$

Thus, if a vector $\boldsymbol{v} \in \mathbb{R}^n$ with $\boldsymbol{v}^T\boldsymbol{v} = 1$ is to be an extremum of $\boldsymbol{v}^T\boldsymbol{A}\boldsymbol{v}$, it must necessarily satisfy the equation $\boldsymbol{A}\boldsymbol{v} = \lambda\boldsymbol{v}$, and hence $\boldsymbol{A}\boldsymbol{v}$ is a multiple of $\boldsymbol{v}$.

The set of all $\lambda \in \mathbb{C}$ that are eigenvalues of a size $n \times n$ matrix $\boldsymbol{A}$ is called the *spectrum* of $\boldsymbol{A}$ and is denoted by $\sigma(\boldsymbol{A})$.

Let $\boldsymbol{A}$ be an $n \times n$ square matrix. A nonzero vector $\boldsymbol{v}$ is called an *eigenvector* of $\boldsymbol{A}$ and λ is the associated *eigenvalue* if

$$\boldsymbol{A}\boldsymbol{v} = \lambda\boldsymbol{v} \tag{2-4}$$

If $\boldsymbol{A}$ is a Hermitian $n \times n$ matrix, then there exist n mutually orthogonal eigenvectors $\boldsymbol{v}_i, i = 1, 2, \cdots, n$. Usually, we normalize each eigenvector to unit length, so that

$$\boldsymbol{v}_i^H \boldsymbol{v}_j = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \tag{2-5}$$

In such a case, the eigenvectors are orthonormal.

We define an $n \times n$ matrix $\boldsymbol{Q}$ whose ith column is the eigenvector $\boldsymbol{v}_i$. Then

$$\boldsymbol{Q}^H \boldsymbol{Q} = \boldsymbol{Q}\boldsymbol{Q}^H = \boldsymbol{I}_n \tag{2-6}$$

Furthermore, $\boldsymbol{A}$ may be represented (decomposed) as

$$\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^H \tag{2-7}$$

where $\boldsymbol{\Lambda}$ is an $n \times n$ diagonal matrix with elements equal to the eigenvalues of $\boldsymbol{A}$. This decomposition is called a *spectral decomposition* of a Hermitian matrix.

If $\boldsymbol{u}$ is an $n \times 1$ nonzero vector for which $\boldsymbol{A}\boldsymbol{u} = \boldsymbol{0}$, then $\boldsymbol{u}$ is called a *null vector* of $\boldsymbol{A}$. When $\boldsymbol{A}$ is Hermitian and $\boldsymbol{A}\boldsymbol{u} = \boldsymbol{0}$ for some vector $\boldsymbol{u}$, then $\boldsymbol{A}$ is singular. A singular Hermitian matrix has at least one zero eigenvalue.

Now, consider the scalar quadratic form $\boldsymbol{u}^H \boldsymbol{A}\boldsymbol{u}$ associate with the Hermitian matrix $\boldsymbol{A}$. If $\boldsymbol{u}^H \boldsymbol{A}\boldsymbol{u} > 0$, the matrix $\boldsymbol{A}$ is said to be *positive definite*. In such a case all the eigenvalues of $\boldsymbol{A}$ are positive. On the other hand, if $\boldsymbol{u}^H \boldsymbol{A}\boldsymbol{u} \geq 0$, matrix $\boldsymbol{A}$ is said to be *positive semidefinite*. In such a case, all the eigenvalues of $\boldsymbol{A}$ are nonnegative.

The following properties involving the eigenvalues of an arbitrary $n \times n$ matrix $\boldsymbol{A} = (a_{ij})_n$ hold:

$$\sum_{i=1}^{n} \lambda_i = \sum_{i=1}^{n} a_{ii} = \operatorname{tr}(\boldsymbol{A}) \tag{2-8}$$

$$\prod_{i=1}^{n} \lambda_i = \det(\boldsymbol{A}) \tag{2-9}$$

$$\sum_{i=1}^{n} \lambda_i^k = \operatorname{tr}\left(\boldsymbol{A}^k\right) \tag{2-10}$$

$$\operatorname{tr}\left(\boldsymbol{A}^T \boldsymbol{A}\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}^2 \geq \sum_{i=1}^{n} \lambda_i^2, \ \boldsymbol{A} \text{ real} \tag{2-11}$$

If we define a polynomial of degree k for a square matrix $\boldsymbol{A}$ as

$$p(\boldsymbol{A}) = \alpha_k \boldsymbol{A}^k + \alpha_{k-1}\boldsymbol{A}^{k-1} + \cdots + \alpha_1 \boldsymbol{A} + \alpha_0 \boldsymbol{I}$$

we have the following theorem.

Theorem 1. If λ is an eigenvalue of $\boldsymbol{A}$, while $\boldsymbol{v}$ is an accociated eigenvector, then $p(\lambda)$ is an eigenvalue of the matrix $p(\boldsymbol{A})$ and $\boldsymbol{v}$ is an eigenvector of $p(\boldsymbol{A})$ accociated with $p(\lambda)$.

2.3 Singular-valued Decomposition

The singular-valued decomposition (SVD) is another orthogonal decomposition of a matrix. Let us assume that $\boldsymbol{A}$ is an $n \times m$ matrix of rank r. Then there exist an $n \times r$ matrix $\boldsymbol{U}$, an $m \times r$ matrix $\boldsymbol{V}$, and an $r \times r$ diagonal matrix $\boldsymbol{\Sigma}$ such that $\boldsymbol{U}^H\boldsymbol{U} = \boldsymbol{V}^H\boldsymbol{V} = \boldsymbol{I}_r$, and

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^H \tag{2-12}$$

where $\boldsymbol{\Sigma} = \text{diag}\,(\sigma_1, \sigma_2, \cdots, \sigma_r)$. The r diagonal elements of $\boldsymbol{\Sigma}$ are strictly positive and are called the *singular values* of matrix $\boldsymbol{A}$. For convenience, we assume that $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r$.

The SVD of matrix $\boldsymbol{A}$ may be expressed as

$$\boldsymbol{A} = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^H \tag{2-13}$$

where $\boldsymbol{u}_i$ are the column vectors of $\boldsymbol{U}$, which are called the *left singular vectors* of $\boldsymbol{A}$, and $\boldsymbol{v}_i$ are the column vectors of $\boldsymbol{V}$, which are called the *right singular vectors* of $\boldsymbol{A}$.

The singular values $\{\sigma_i\}$ are the nonnegative square roots of the eigenvalues of matrix $\boldsymbol{A}^H\boldsymbol{A}$. To demonstrate this, we postmultiply Equation 2-12 by $\boldsymbol{V}$. Thus, we obtain

$$\boldsymbol{A}\boldsymbol{V} = \boldsymbol{U}\boldsymbol{\Sigma} \tag{2-14}$$

or, equivalently,

$$\boldsymbol{A}\boldsymbol{v}_i = \sigma_i \boldsymbol{u}_i,\ i = 1, 2, \cdots, r \tag{2-15}$$

Similarly, we postmultiply $\boldsymbol{A}^H = \boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{U}^H$ by $\boldsymbol{U}$. Thus, we obtain

$$\boldsymbol{A}^H\boldsymbol{U} = \boldsymbol{V}\boldsymbol{\Sigma} \tag{2-16}$$

or equivalently,

$$\boldsymbol{A}^H\boldsymbol{u}_i = \sigma \boldsymbol{u}_i,\ i = 1, 2, \cdots, r \tag{2-17}$$

Then, by premultiplying both sides of Equation 2-15 with $\boldsymbol{A}^H$ and using Equation 2-17, we obtain

$$\boldsymbol{A}^H\boldsymbol{A}\boldsymbol{v}_i = \sigma_i^2 \boldsymbol{v}_i,\ i = 1, 2, \cdots, r \tag{2-18}$$

This demonstrates that the r nonzero eigenvalues of $\boldsymbol{A}^H\boldsymbol{A}$ are the squares of the singular values of $\boldsymbol{A}$, and the corresponding r eigenvectors $\boldsymbol{v}_i$ are the right singular vectors of $\boldsymbol{A}$. The remaining $m - r$ eigenvalues of $\boldsymbol{A}^H\boldsymbol{A}$ are zero. On the other hand, if we premultiply both sides of Equation 2-17 by $\boldsymbol{A}$ and use Equation 2-15, we obtain

$$\boldsymbol{A}\boldsymbol{A}^H\boldsymbol{u}_i = \sigma_i^2 \boldsymbol{v}_i,\ i = 1, 2, \cdots, r \tag{2-19}$$

This demonstrates that the r nonzero eigenvalues of $\boldsymbol{A}\boldsymbol{A}^H$ are the squares of the singular values of $\boldsymbol{A}$, and the corresponding r eigenvectors $\boldsymbol{u}_i$ are the left singular vectors of $\boldsymbol{A}$. The remaining $n - r$ eigenvalues of $\boldsymbol{A}\boldsymbol{A}^H$ are zero. Hence, $\boldsymbol{A}\boldsymbol{A}^H$ and $\boldsymbol{A}^H\boldsymbol{A}$ have the same set of nonzero eigenvalues.

2.4 Matrix Norm and Condition Number

Recall that the Euclidean norm (L_2 norm) of a vector $\boldsymbol{v}$, denoted as $\|\boldsymbol{v}\|$, is defined as

$$\|\boldsymbol{v}\| = \left(\boldsymbol{v}^H \boldsymbol{v}\right)^{1/2} \tag{2-20}$$

The Euclidean norm of a matrix $\boldsymbol{A}$, denoted as $\|\boldsymbol{A}\|$, is defined as

$$\|\boldsymbol{A}\| = \max \frac{\|\boldsymbol{A}\boldsymbol{v}\|}{\|\boldsymbol{v}\|} \tag{2-21}$$

for any vector $\boldsymbol{v}$. It is easy to verify that the norm of a Hermitian matrix is equal to the largest eigenvalue.

Another useful quantity associated with a matrix $\boldsymbol{A}$ is the nonzero minimum value of $\|\boldsymbol{A}\boldsymbol{v}\|/\|\boldsymbol{v}\|$. When $\boldsymbol{A}$ is a nonsingular Hermitian matrix, this minimum value is equal to the smallest eigenvalue.

The squared Frobenius norm of an $n \times m$ matrix $\boldsymbol{A}$ is defined as

$$\|\boldsymbol{A}\|_F^2 = \mathrm{tr}\left(\boldsymbol{A}\boldsymbol{A}^H\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} |a_{ij}|^2 \tag{2-22}$$

From the SVD of the matrix $\boldsymbol{A}$, it follows that

$$\|\boldsymbol{A}\|_F^2 = \sum_{i=1}^{n} \lambda_i \tag{2-23}$$

where $\{\lambda_i\}$ are the eigenvalues of $\boldsymbol{A}\boldsymbol{A}^H$.

The following are bounds on matrix norms:

$$\begin{array}{rcl} \|\boldsymbol{A}\| & > & 0,\ \boldsymbol{A} \neq \boldsymbol{0} \\ \|\boldsymbol{A}+\boldsymbol{B}\| & \leq & \|\boldsymbol{A}\| + \|\boldsymbol{B}\| \\ \|\boldsymbol{A}\boldsymbol{B}\| & \leq & \|\boldsymbol{A}\|\|\boldsymbol{B}\| \end{array} \tag{2-24}$$

The condition number of a matrix $\boldsymbol{A}$ is defined as the ratio of the maximum value to the minimum value of $\|\boldsymbol{A}\boldsymbol{v}\|/\|\boldsymbol{v}\|$. When $\boldsymbol{A}$ is Hermitian, the condition number is $\lambda_{\max}/\lambda_{\min}$, where $\lambda_{\max}$ is the largest eigenvalue and $\lambda_{\min}$ is the smallest eigenvalue of $\boldsymbol{A}$.

2.5 The Moore-Penrose Pseudoinverse

Let us consider a rectangular $n \times m$ matrix $\boldsymbol{A}$ of rank r, having an SVD as $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^H$. The Moore-Penrose pseudoinverse, denoted by $\boldsymbol{A}^+$, is an $m \times n$ matrix defined as

$$\boldsymbol{A}^+ = \boldsymbol{V}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^H \tag{2-25}$$

where $\boldsymbol{\Sigma}^{-1}$ is an $r \times r$ diagonal matrix with diagonal elements $1/\sigma_i$, $i = 1, 2, \cdots, r$. We may also express $\boldsymbol{A}^+$ as

$$\boldsymbol{A}^+ = \sum_{i=1}^{r} \frac{1}{\sigma_i} \boldsymbol{v}_i \boldsymbol{u}_i^H \tag{2-26}$$

We observe that the rank of $\boldsymbol{A}^+$ is equal to the rank of $\boldsymbol{A}$.

When the rank $r = m$ or $r = n$, the pseudoinverse $\boldsymbol{A}^+$ can be expressed as

$$\begin{array}{lcll} \boldsymbol{A}^+ & = & \boldsymbol{A}^H \left(\boldsymbol{A}\boldsymbol{A}^H\right)^{-1} & r = n \\ \boldsymbol{A}^+ & = & \left(\boldsymbol{A}^H\boldsymbol{A}\right)^{-1} \boldsymbol{A}^H & r = m \\ \boldsymbol{A}^+ & = & \boldsymbol{A}^{-1} & r = m = n \end{array} \tag{2-27}$$

These relations are equivalent to $\boldsymbol{A}\boldsymbol{A}^+ = \boldsymbol{I}_n$ and $\boldsymbol{A}^+\boldsymbol{A} = \boldsymbol{I}_m$.

2.6 Some Useful Random Variables

In subsequent chapters, we shall encounter several different types of random variables. In section we list these frequently encountered random variables, their probability density functions (PDFs), their cumulative distribution functions (CDFs), and their moments. Our main emphasis will be on the Gaussian random variable and many random variables that are derived from the Gaussian random variable.

2.6.1 The Bernoulli Random Variable

The Bernoulli random variable is a discrete binary-valued random variable taking values 1 and 0 with probabilities p and $1-p$, respectively. Therefore the probability mass function (PMF) for this random variable is given by

$$\begin{cases} \Pr\{X=1\} &= p \\ \Pr\{X=0\} &= 1-p \end{cases} \tag{2-28}$$

The mean and variance of this random variable are given by

$$\begin{cases} E\{X\} &= p \\ \mathrm{Var}\{X\} &= p(1-p) \end{cases} \tag{2-29}$$

2.6.2 The Binomial Random Variable

The binomial random variable models the sum of n independent Bernoulli random variables with common parameter p. The PMF of this random variable is given by

$$\Pr\{X=k\} = {}_kC_n p^k (1-p)^{n-k}, \quad k=0,1,\cdots,n \tag{2-30}$$

For this random variable we have

$$\begin{cases} E\{X\} &= np \\ \mathrm{Var}\{X\} &= np(1-p) \end{cases} \tag{2-31}$$

This random variable models, for instance, the number of errors when n bits are transmitted over a communication channel and the probability of error for each bit is p.

2.6.3 The Uniform Random Variable

The uniform random variable is a continuous random variable with PDF

$$p(x) = \begin{cases} \frac{1}{b-a} &= a \le x \le b \\ 0 &= \text{otherwise} \end{cases} \tag{2-32}$$

where $b > a$ and the interval $[a,b]$ is the range of the random variable. Here we have

$$\begin{cases} E\{X\} &= \frac{b-a}{2} \\ \mathrm{Var}\{X\} &= \frac{(b-a)^2}{12} \end{cases} \tag{2-33}$$

2.6.4 The Gaussian (Normal) Random Variable

The Gaussian random variable is described in terms of two parameters $\mu \in \mathbb{R}$ and $\sigma > 0$ by the PDF

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{2-34}$$

We usually use the shorthand form $\mathcal{N}(\mu, \sigma^2)$ to denote the PDF of Gaussian random variables and write $X \sim \mathcal{N}(\mu, \sigma^2)$. For this random variable

$$\begin{cases} E\{X\} &= \mu \\ \text{Var}\{X\} &= \sigma^2 \end{cases} \tag{2-35}$$

A Gaussian random variable with $\mu = 0$ and $\sigma = 1$ is called a *standard normal*. A function closely related to the Gaussian random variable is the Q function defined as

$$Q(x) = \Pr\{\mathcal{N}(0,1) > x\} = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-\frac{t^2}{2}} dt \tag{2-36}$$

The CDF of a Gaussian random variable is given by

$$\begin{aligned} F(x) &= \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt \\ &= 1 - \int_{x}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt \\ &= 1 - \int_{\frac{x-\mu}{\sigma}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du \\ &= 1 - Q\left(\frac{x-\mu}{\sigma}\right) \end{aligned} \tag{2-37}$$

where we have introduced the change of variable $u = (t - \mu)/\sigma$. The PDF and the CDF of a Gaussian random variable are shown in Figure 2-1.

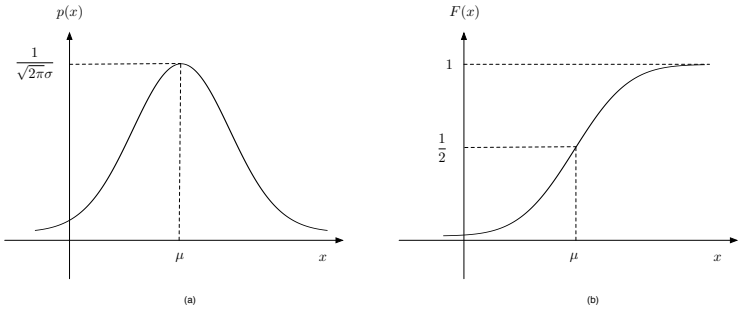

Figure 2-1: PDF and CDF of a Gaussian random variable.

In general if $X \sim \mathcal{N}(\mu, \sigma^2)$, then

$$\begin{aligned} \Pr\{X > \alpha\} &= Q\left(\frac{\alpha - \mu}{\sigma}\right) \\ \Pr\{X < \alpha\} &= Q\left(\frac{\mu - \alpha}{\sigma}\right) \end{aligned} \tag{2-38}$$

Following are some of the important properties of the Q function:

$$\begin{array}{rcl} Q(0) & = & \frac{1}{2} \\ Q(\infty) & = & 0 \\ Q(-\infty) & = & 1 \\ Q(-x) & = & 1 - Q(x) \end{array} \tag{2-39}$$

Some useful bounds for the Q function for $x > 0$ are

$$\begin{array}{rcl} Q(x) & \leq & \frac{1}{2}e^{-\frac{x^2}{2}} \\ Q(x) & \leq & \frac{1}{x\sqrt{2\pi}}e^{-\frac{x^2}{2}} \\ Q(x) & \leq & \frac{x}{(1+x^2)\sqrt{2\pi}}e^{-\frac{x^2}{2}} \end{array} \tag{2-40}$$

From the last two bounds we conclude that for large x we have

$$Q(x) \approx \frac{1}{x\sqrt{2\pi}}e^{-\frac{x^2}{2}} \tag{2-41}$$

A plot of the Q function bounds is given in Figure 2-2.

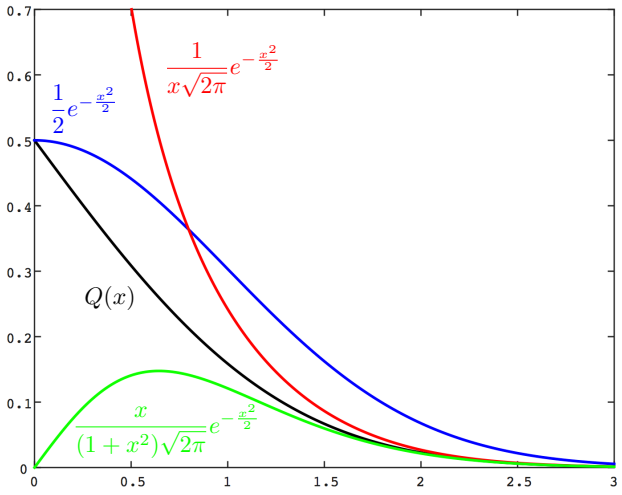

Figure 2-2: Plot of $Q(x)$ and its upper and lower bounds.

Tables 2.1 and 2.2 give values of the Q function.

Another function closely related to the Q function is the *complementary error function*, defined as

$$\text{erfc}(x) = \frac{2}{\sqrt{\pi}}\int_x^\infty e^{-t^2}dt \tag{2-42}$$

The complementary error function is related to the Q function as follows:

$$Q(x) = \frac{1}{2}\text{erfc}\left(\frac{x}{\sqrt{2}}\right)$$

Table 2.1: Table of Q Function Values

x	$Q(x)$	x	$Q(x)$	x	$Q(x)$	x	$Q(x)$
0	5.0000×10^{-1}	1.8	3.5930×10^{-2}	3.6	1.59×10^{-4}	5.4	3.3320×10^{-8}
0.1	4.6017×10^{-1}	1.9	2.8717×10^{-2}	3.7	1.08×10^{-4}	5.5	1.8990×10^{-8}
0.2	4.2074×10^{-1}	2	2.2750×10^{-2}	3.8	7.2348×10^{-5}	5.6	1.0718×10^{-8}
0.3	3.8209×10^{-1}	2.1	1.7864×10^{-2}	3.9	4.8096×10^{-5}	5.7	5.9904×10^{-9}
0.4	3.4458×10^{-1}	2.2	1.3903×10^{-2}	4	3.1671×10^{-5}	5.8	3.3157×10^{-9}
0.5	3.0854×10^{-1}	2.3	1.0724×10^{-2}	4.1	2.0658×10^{-5}	5.9	1.8175×10^{-9}
0.6	2.7425×10^{-1}	2.4	8.198×10^{-3}	4.2	1.3346×10^{-5}	6.0	9.8659×10^{-10}
0.7	2.4196×10^{-1}	2.5	6.210×10^{-3}	4.3	8.5399×10^{-6}	6.1	5.3034×10^{-10}
0.8	2.1186×10^{-1}	2.6	4.661×10^{-3}	4.4	5.4125×10^{-6}	6.2	2.8232×10^{-10}
0.9	1.8406×10^{-1}	2.7	3.467×10^{-3}	4.5	3.3977×10^{-6}	6.3	1.4882×10^{-10}
1.0	1.5866×10^{-1}	2.8	2.555×10^{-3}	4.6	2.1125×10^{-6}	6.4	7.7689×10^{-11}
1.1	1.3567×10^{-1}	2.9	1.866×10^{-3}	4.7	1.3008×10^{-6}	6.5	4.0160×10^{-11}
1.2	1.1507×10^{-1}	3.0	1.350×10^{-3}	4.8	7.9333×10^{-7}	6.6	2.0558×10^{-11}
1.3	9.6800×10^{-2}	3.1	9.68×10^{-4}	4.9	4.7918×10^{-7}	6.7	1.0421×10^{-11}
1.4	8.0757×10^{-2}	3.2	6.87×10^{-4}	5.0	2.8665×10^{-7}	6.8	5.2309×10^{-12}
1.5	6.6807×10^{-2}	3.3	4.83×10^{-4}	5.1	1.6983×10^{-7}	6.9	2.6001×10^{-12}
1.6	5.4799×10^{-2}	3.4	3.37×10^{-4}	5.2	9.9644×10^{-8}	7.0	1.2799×10^{-12}
1.7	4.4565×10^{-2}	3.5	2.33×10^{-4}	5.3	5.7901×10^{-8}	7.1	6.2378×10^{-13}

Table 2.2: Selected Q Function Values

$Q(x)$	x
10^{-1}	1.2816
10^{-2}	2.3263
10^{-3}	3.0902
10^{-4}	3.7190
10^{-5}	4.2649
10^{-6}	4.7534
10^{-7}	5.1993
0.5×10^{-5}	4.4172
0.25×10^{-5}	4.5648
0.667×10^{-5}	4.3545

$$\text{erfc}(x) \quad = \quad 2Q\left(\sqrt{2}x\right) \tag{2-43}$$

The characteristic function[1] of a Gaussian random variable is given by

$$\Phi_X(\omega) = e^{j\omega\mu - \frac{1}{2}\omega^2\sigma^2} \tag{2-44}$$

For an $\mathcal{N}(\mu, \sigma^2)$ random variable we have

$$E\left\{(X-\mu)^n\right\} = \begin{cases} 1 \times 3 \times 5 \times \cdots \times (2k-1)\sigma^{2k} = \frac{(2k)!\sigma^{2k}}{2^k k!} & \text{for } n = 2k \\ 0 & \text{for } n = 2k+1 \end{cases} \tag{2-45}$$

from which we can obtain moments of the Gaussian random variable.

The sum of n independent Gaussian random variables is a Gaussian random variable whose mean and variance are the sum of the means and the sum of the variances of the random variables, respectively.

[1]Recall that for any random variable X, the *characterisric function* is defined by $\Phi_X(\omega) = E\{e^{j\omega X}\}$. The *moment generating function* (MGF) is defined by $\Theta_X(t) = E\{e^{tX}\}$. Obviously, $\Theta(t) = \Phi(-jt)$ and $\Phi(\omega) = \Theta(j\omega)$.

2.6.5 The Rayleigh Random Variable

If X_1 and X_2 are two i.i.d. random variables each distributed according to $\mathcal{N}(0, \sigma^2)$, then

$$X = \sqrt{X_1^2 + X_2^2} \tag{2-46}$$

is a *Rayleigh random variable*. From our discussion of the χ^2 random variables, it is readily seen that a Rayleigh random variable the square root of a χ^2 random variable with two degrees of freedom. We can also conclude that the Rayleigh random variable is the square root of an exponential random variable as given by Equation 2.3-27. The PDF of a Rayleigh random variable is given by

$$p(x) = \begin{cases} \frac{x}{\sigma^2} e^{-\frac{x^2}{2\sigma^2}} & x > 0 \\ 0 & \text{otherwise} \end{cases} \tag{2-47}$$

and its mean and variance are

$$\begin{cases} E\{X\} &= \sigma\sqrt{\frac{\pi}{2}} \\ \text{Var}\{X\} &= \left(2 - \frac{\pi}{2}\right)\sigma^2 \end{cases} \tag{2-48}$$

2.6.6 Jointly Gaussian Random Variables

An $n \times 1$ column random vector $\boldsymbol{X}$ with components $\{X_i, 1 \leq i \leq n$ is called a *Gaussian vector*, and its components are called *jointly Gaussian random variables* or *multivariate Gaussian random variables* if the joint PDF of X_i's can be written as

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{n/2}\left(\det \boldsymbol{C}\right)^{1/2}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{C}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})} \tag{2-49}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{C}$ are the mean vector and covariance matrix, respectively, of $\boldsymbol{X}$ and are given by

$$\begin{cases} \boldsymbol{\mu} &= E\{\boldsymbol{X}\} \\ \boldsymbol{C} &= E\left\{(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^T\right\} \end{cases} \tag{2-50}$$

From this definition it is clear that

$$C_{ij} = \text{Cov}\{X_i, X_j\} \tag{2-51}$$

and therefore $\boldsymbol{C}$ is a symmetric matrix. From elementary probability it is also well known that $\boldsymbol{C}$ is nonnegative definite.

In the special case of $n = 2$, we have

$$\begin{cases} \boldsymbol{\mu} &= \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \\ \boldsymbol{C} &= \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} \end{cases} \tag{2-52}$$

where

$$\rho = \frac{\text{Cov}\{X_1, X_2\}}{\sigma_1\sigma_2}$$

is the correlation coefficient of the two random variables. In this case the PDF reduces to

$$p(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2+\left(\frac{x_2-\mu_2}{\sigma_2}\right)^2-2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right)}{2(1-\rho^2)}} \tag{2-53}$$

where μ_1, μ_2, σ_1^2 and, σ_2^2 are means and variances of the two random variables and ρ is their correlation coefficient. Note that in the special case when $\rho = 0$ (*i.e.*, when the two random variables are uncorrelated), we have

$$p(x_1, x_2) = \mathcal{N}(\mu_1, \sigma_1^2) \times \mathcal{N}(\mu_2, \sigma_2^2)$$

This means that the two random variables are independent, and therefore for this case independence and uncorrelatedness are equivalent. This property is true for general jointly Gaussian random variables.

Another important property of jointly Gaussian random variables is that linear combinations of jointly Gaussian random variables are also jointly Gaussian. In other words, if $\boldsymbol{X}$ is a Gaussian vector, the random vector $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X}$, where the invertible matrix $\boldsymbol{A}$ represents a linear transformation, is also a Gaussian vector whose mean and covariance matrix are given by

$$\begin{cases} \boldsymbol{\mu}_Y &= \boldsymbol{A}\boldsymbol{\mu}_X \\ \boldsymbol{C}_Y &= \boldsymbol{A}\boldsymbol{C}_X\boldsymbol{A}^T \end{cases} \tag{2-54}$$

In summary, jointly Gaussian random variables, have the following important properties:

1. For jointly Gaussian random variables, *uncorrelated* is equivalent to *independent*.
2. Linear combinations or jointly random variables are themselves jointly Gaussian.
3. The random variables in any subset of jointly Gaussian random variables are jointly Gaussian, and any subset of random variables conditioned on random variables in any other subset is also jointly Gaussian (all joint subsets and all conditional subsets are Gaussian).

We also emphasize that any set of independent Gaussian random variables is jointly Gaussian, but this is not necessarily true for a set of dependent Gaussian random variables.

2.7 Complex Random Variables

A complex random variable $Z = X + jY$ can be considered as a pair of real random variables X and Y. Therefore, we treat a complex random variable as a two-dimensional random vector with components X and Y. The PDF of a complex random variable is defined to be the joint PDF of its real and complex parts. If X and Y are jointly Gaussian random variables, then Z is a complex Gaussian random variable. The PDF of a zero-mean complex Gaussian random variable Z with i.i.d. real and imaginary parts is given by

$$\begin{aligned} p(z) &= \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} && (2\text{-}55) \\ &= \frac{1}{2\pi\sigma^2} e^{-\frac{|z|^2}{2\sigma^2}} && (2\text{-}56) \end{aligned}$$

For a complex random variable Z the mean and variance are defined by

$$\begin{aligned} E\{Z\} &= E\{X\} + jE\{Y\} && (2\text{-}57) \\ \mathrm{Var}\{Z\} &= E\left\{|Z|^2\right\} - |E\{Z\}|^2 = \mathrm{Var}\{X\} + \mathrm{Var}\{Y\} && (2\text{-}58) \end{aligned}$$

2.7.1 Complex Random Vectors

A complex random vector is defined as $\boldsymbol{Z} = \boldsymbol{X} + j\boldsymbol{Y}$, where $\boldsymbol{X}$ and $\boldsymbol{Y}$ are real-valued random vectors of size n. We define the following real-valued matrices for a complex random vector $\boldsymbol{Z}$.

$$\begin{aligned} \boldsymbol{C_X} &= E\left\{(\boldsymbol{X} - E\{\boldsymbol{X}\})(\boldsymbol{X} - E\{\boldsymbol{X}\})^T\right\} && (2\text{-}59) \\ \boldsymbol{C_Y} &= E\left\{(\boldsymbol{Y} - E\{\boldsymbol{Y}\})(\boldsymbol{Y} - E\{\boldsymbol{Y}\})^T\right\} && (2\text{-}60) \\ \boldsymbol{C_{XY}} &= E\left\{(\boldsymbol{X} - E\{\boldsymbol{X}\})(\boldsymbol{Y} - E\{\boldsymbol{Y}\})^T\right\} && (2\text{-}61) \\ \boldsymbol{C_{YX}} &= E\left\{(\boldsymbol{Y} - E\{\boldsymbol{Y}\})(\boldsymbol{X} - E\{\boldsymbol{X}\})^T\right\} && (2\text{-}62) \end{aligned}$$

Matrices $\boldsymbol{C_X}$ and $\boldsymbol{C_Y}$ are the *covariance matrices* of real random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively, and hence they are symmetric and nonnegative definite. It is clear from above that $\boldsymbol{C_{YX}} = \boldsymbol{C_{XY}}^T$.

The PDF of $\boldsymbol{Z}$ is the joint PDF of its real and imaginary parts. If we define the $2n$-dimensional real vector

$$\tilde{\boldsymbol{Z}} = \begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \end{bmatrix} \tag{2-63}$$

then the PDF of the complex vector $\boldsymbol{X}$ is the PDF of the real vector $\tilde{\boldsymbol{Z}}$. It is clear that $\boldsymbol{C}_{\tilde{\boldsymbol{Z}}}$, the covariance matrix of $\tilde{\boldsymbol{Z}}$, can be written as

$$\boldsymbol{C}_{\tilde{\boldsymbol{Z}}} = \begin{bmatrix} \boldsymbol{C_X} & \boldsymbol{C_{XY}} \\ \boldsymbol{C_{YX}} & \boldsymbol{C_Y} \end{bmatrix} \tag{2-64}$$

We also define the following two, in general complex-valued, matrices

$$\begin{aligned} \boldsymbol{C_Z} &= E\left\{(\boldsymbol{Z} - E\{\boldsymbol{Z}\})(\boldsymbol{Z} - E\{\boldsymbol{Z}\})^H\right\} && (2\text{-}65) \\ \tilde{\boldsymbol{C}}_{\boldsymbol{Z}} &= E\left\{(\boldsymbol{Z} - E\{\boldsymbol{Z}\})(\boldsymbol{Z} - E\{\boldsymbol{Z}\})^T\right\} && (2\text{-}66) \end{aligned}$$

where $\boldsymbol{A}^T$ denotes the transpose and $\boldsymbol{A}^H$ denotes the Hermitian transpose of $\boldsymbol{A}$ ($\boldsymbol{A}$ is transposed and each element of it is conjugated). $\boldsymbol{C_Z}$ and $\tilde{\boldsymbol{C}}_{\boldsymbol{Z}}$ are called the *covariance* and the *pseudocovariance* of the complex random vector $\boldsymbol{Z}$, respectively. It is easy to verify that for any $\boldsymbol{Z}$, the covariance matrix is Hermitian[2] and nonnegative definite. The pseudocovariance is skew-Hermitian.

From these definitions it is easy to verify the following relations.

$$\boldsymbol{C_Z} = \boldsymbol{C_X} + \boldsymbol{C_Y} + j(\boldsymbol{C_{YX}} - \boldsymbol{C_{XY}}) \tag{2-67}$$

$$\tilde{\boldsymbol{C}}_{\boldsymbol{Z}} = \boldsymbol{C_X} - \boldsymbol{C_Y} + j(\boldsymbol{C_{YX}} + \boldsymbol{C_{XY}}) \tag{2-68}$$

$$\boldsymbol{C_X} = \frac{1}{2}\Re\left\{\boldsymbol{C_Z} + \tilde{\boldsymbol{C}}_{\boldsymbol{Z}}\right\} \tag{2-69}$$

$$\boldsymbol{C_Y} = \frac{1}{2}\Re\left\{\boldsymbol{C_Z} - \tilde{\boldsymbol{C}}_{\boldsymbol{Z}}\right\} \tag{2-70}$$

$$\boldsymbol{C_{YX}} = \frac{1}{2}\Im\left\{\boldsymbol{C_Z} + \tilde{\boldsymbol{C}}_{\boldsymbol{Z}}\right\} \tag{2-71}$$

$$\boldsymbol{C_{XY}} = \frac{1}{2}\Im\left\{\tilde{\boldsymbol{C}}_{\boldsymbol{Z}} - \boldsymbol{C_Z}\right\} \tag{2-72}$$

2.7.2 Proper and Circularly Symmetric Random Vectors

A complex random vector $\boldsymbol{Z}$ is called *proper* if its pseudocovariance is zero, *i.e.*, if $\tilde{\boldsymbol{C}}_{\boldsymbol{Z}} = 0$. From Equation 2-68 it is clear that for a proper random vector we have

$$\boldsymbol{C_X} = \boldsymbol{C_Y} \tag{2-73}$$

$$\boldsymbol{C_{XY}} = -\boldsymbol{C_{YX}} \tag{2-74}$$

Substituting these results into Equations 2-67 to 2-72 and 2-64, we conclude that for proper random vectors

$$\boldsymbol{C_Z} = 2\boldsymbol{C_X} + 2j\boldsymbol{C_{YX}} \tag{2-75}$$

$$\boldsymbol{C_X} = \boldsymbol{C_Y} = \frac{1}{2}\Re\{\boldsymbol{C_Z}\} \tag{2-76}$$

$$\boldsymbol{C_{YX}} = -\boldsymbol{C_{XY}} = \frac{1}{2}\Im\{\boldsymbol{C_Z}\} \tag{2-77}$$

$$\boldsymbol{C}_{\tilde{\boldsymbol{Z}}} = \begin{bmatrix} \boldsymbol{C_X} & \boldsymbol{C_{XY}} \\ -\boldsymbol{C_{XY}} & \boldsymbol{C_X} \end{bmatrix} \tag{2-78}$$

For the special case of $n = 1$, *i.e.*, when we are dealing with a single complex random variable $Z = X + jY$, the conditions for being proper become

$$\mathrm{Var}\{X\} = \mathrm{Var}\{Y\} \tag{2-79}$$

$$\mathrm{Cov}\{X, Y\} = -\mathrm{Cov}\{Y, X\} \tag{2-80}$$

which means that Z is proper if X and Y have equal variances and are uncorrelated. In this case $\mathrm{Var}\{Z\} = 2\mathrm{Var}\{X\}$. Since in the case of jointly Gaussian random variables uncorrelated is equivalent to independent, we conclude that a complex Gaussian random variable Z is proper if and only if its real and complex parts independent with equal variance. For a zero-mean proper complex Gaussian random variable, the PDF is given by Equation 2-56.

[2]Matrix $\boldsymbol{A}$ is Hermitian if $\boldsymbol{A} = \boldsymbol{A}^H$. It is skew-Hermitian if $\boldsymbol{A}^H = -\boldsymbol{A}$.

If the complex random vector $\boldsymbol{Z} = \boldsymbol{X} + j\boldsymbol{Y}$ is Gaussian, meaning that $\boldsymbol{X}$ and $\boldsymbol{Y}$ are jointly Gaussian, then we have

$$p(\boldsymbol{z}) = p(\tilde{\boldsymbol{z}}) = \frac{1}{(2\pi)^n (\det \boldsymbol{C}_{\tilde{\boldsymbol{Z}}})^{\frac{1}{2}}} e^{-\frac{1}{2}(\tilde{\boldsymbol{z}}-\tilde{\boldsymbol{\mu}})^T \boldsymbol{C}_{\tilde{\boldsymbol{Z}}}^{-1} (\tilde{\boldsymbol{z}}-\tilde{\boldsymbol{\mu}})} \tag{2-81}$$

where

$$\tilde{\boldsymbol{\mu}} = E\left\{\tilde{\boldsymbol{Z}}\right\} \tag{2-82}$$

It can be shown that in the special case where $\boldsymbol{Z}$ is a proper n-dimensional complex Gaussian random vector, with mean $\boldsymbol{\mu} = E\{\boldsymbol{Z}\}$ and nonsingular covariance matrix $\boldsymbol{C}_{\boldsymbol{Z}}$, its PDF can be written as

$$p(\boldsymbol{z}) = \frac{1}{\pi^n \det \boldsymbol{C}_{\tilde{\boldsymbol{Z}}}} e^{-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})^T \boldsymbol{C}_{\boldsymbol{Z}}^{-1} (\boldsymbol{z}-\boldsymbol{\mu})} \tag{2-83}$$

A complex random vector $\boldsymbol{Z}$ is called *circularly symmetric* or *circular* if rotating the vector by any angle does not change its PDF. In other words, a complex random vector $\boldsymbol{Z}$ is circularly symmetric if $\boldsymbol{Z}$ and $e^{j\theta}\boldsymbol{Z}$ have the same PDF for all θ. If $\boldsymbol{Z}$ is circular, then it is zero-mean and proper, *i.e.*, $E\{\boldsymbol{Z}\} = \boldsymbol{0}$ and $E\left\{\boldsymbol{Z}\boldsymbol{Z}^T\right\} = \boldsymbol{0}$. If $\boldsymbol{Z}$ is a zero-mean proper Gaussian complex vector, then $\boldsymbol{Z}$ is circular. In other words, *for complex Gaussian random vectors being zero-mean and proper is equivalent to being circular.*

If $\boldsymbol{Z}$ is a proper complex vector, then any *affine transformation* of it, *i.e.*, any transform of the form $\boldsymbol{W} = \boldsymbol{A}\boldsymbol{Z} + \boldsymbol{b}$, is also a proper complex vector. Since we know that if $\boldsymbol{Z}$ is Gaussian, so is $\boldsymbol{W}$, we conclude that if $\boldsymbol{Z}$ is a proper Gaussian vector, so is $\boldsymbol{W}$.

2.8 Random Processes

Random processes, stochastic processes, or random signals are fundamental in the study of systems. Modeling information sources and communication channels requires a good understanding of random processes and techniques for analyzing them. We assume that the reader has a knowledge of the basic concepts of random processes including definitions of mean, autocorrelation, cross-correlation. stationarity, and ergodicity as given in standard texts such as Leon-Garcia (1994). Papoulis and Pillai (2002), Stark and Woods (2002). In the following paragraphs we present a brief review of the most important properties of random processes.

The mean $\mu_X(t)$ and the *autocorrelation function* of a random process $X(t)$ are defined as

$$\mu_X(t) = E\{X(t)\} \tag{2-84}$$

$$R_X(t_1, t_2) = E\{X(t_1)X^*(t_2)\} \tag{2-85}$$

The *cross-correlation function* of two random processes $X(t)$ and $Y(t)$ is defined by

$$R_{XY}(t_1, t_2) = E\{X(t_1)Y^*(t_2)\} \tag{2-86}$$

Note that $R_X(t_2, t_1) = R_X^*(t_1, t_2)$, *i.e.*, $R_X(t_1, t_2)$ is Hermitian. For the cross-correlation we have $R_{YX}(t_2, t_1) = R_{XY}^*(t_1, t_2)$.

2.8.1 Wide-Sense Stationary Random Processes

Random process $X(t)$ is *wide-sense stationary* (WSS) if its mean is constant and $R_X(t_1, t_2) = R_X(\tau)$, where $\tau = t_1 - t_2$. For WSS processes $R_X(-\tau) = R_X^*(\tau)$. Two processes $X(t)$ and $Y(t)$ are *jointly wide-sense stationary* if both $X(t)$ and $Y(t)$ are WSS and $R_{XY}(t_1, t_2) = R_{XY}(\tau)$. For jointly WSS processes $R_{YX}(-\tau) = R_{XY}^*(\tau)$. A complex process is WSS if its real and imaginary parts are jointly WSS.

The *power spectral density* (PSD) or *power spectrum* of a WSS random process $X(t)$ is a function $\mathcal{S}_X(f)$ describing the distribution o power as a function of frequency. The unit for power spectral density is watts per hertz. The *Wiener-Khinchin theorem* states that for a WSS process, the power spectrum is the Fourier transform of the autocorrelation function $R_X(\tau)$, *i.e.*,

$$\mathcal{S}_X(f) = \mathcal{F}\{R_X(\tau)\} \tag{2-87}$$

Similarly, *the cross spectral density* (CSD) of two jointly WSS processes is defined as the Fourier transform of their cross-correlation function.

$$\mathcal{S}_{XY}(f) = \mathcal{F}\{R_{XY}(\tau)\} \tag{2-88}$$

The cross spectral density satisfies the following symmetry property:

$$\mathcal{S}_{XY}(f) = \mathcal{F}\{R_{YX}^*(\tau)\} \tag{2-89}$$

From properties of the autocorrelation it is easy verify that the power spectral density of any real WSS process $X(t)$ is real, nonnegative, and even function of f. For complex processes, power spectrum is real but not necessarily even. The cross spectral density can be a complex function, even when both $X(t)$ and $Y(t)$ are real processes.

If $X(t)$ and $Y(t)$ are jointly WSS processes, then $Z(t) = aX(t) + bY(t)$ is a WSS random process with autocorrelation and power spectral density given by

$$R_Z(\tau) = |a|^2 R_X(\tau) + |b|^2 R_Y(\tau) + ab^* R_{XY}(\tau) + ba^* R_{YX}(\tau) \tag{2-90}$$

$$\mathcal{S}_Z(\tau) = |a|^2 \mathcal{S}_X(f) + |b|^2 \mathcal{S}_Y(f) + 2\Re\{ab^* \mathcal{S}_{XY}(f)\} \tag{2-91}$$

In the special case where $a = b = 1$, we have $Z(t) = X(t) + Y(t)$, which results in

$$\begin{aligned} R_Z(\tau) &= R_X(\tau) + R_Y(\tau) + R_{XY}(\tau) + R_{YX}(\tau) && (2\text{-}92) \\ \mathcal{S}_Z(\tau) &= \mathcal{S}_X(f) + \mathcal{S}_Y(f) + 2\Re\{\mathcal{S}_{XY}(f)\} && (2\text{-}93) \end{aligned}$$

and when $a = 1$ and $b = j$, we have $Z(t) = X(t) + jY(t)$ and

$$\begin{aligned} R_Z(\tau) &= R_X(\tau) + R_Y(\tau) + j(R_{YX}(\tau) + R_{XY}(\tau)) && (2\text{-}94) \\ \mathcal{S}_Z(\tau) &= \mathcal{S}_X(f) + \mathcal{S}_Y(f) + 2\Im\{\mathcal{S}_{XY}(f)\} && (2\text{-}95) \end{aligned}$$

When a WSS process $X(t)$ passes through an LTl system with impulse response $h(t)$ and transfer function $H(f) = \mathcal{F}\{h(t)\}$, the output process $Y(t)$ and $X(t)$ are jointly WSS and the following relations hold:

$$\begin{aligned} \mu_Y &= \mu_X \int_{-\infty}^{\infty} h(t)dt && (2\text{-}96) \\ R_{XY}(\tau) &= R_X(\tau) * h^*(-\tau) && (2\text{-}97) \\ R_Y(\tau) &= R_X(\tau) * h(\tau) * h^*(-\tau) && (2\text{-}98) \\ \mu_Y &= \mu_X H(0) && (2\text{-}99) \\ \mathcal{S}_{XY}(f) &= \mathcal{S}_X(f)H^*(f) && (2\text{-}100) \\ \mathcal{S}_Y(f) &= \mathcal{S}_X(f)|H(f)|^2 && (2\text{-}101) \end{aligned}$$

The power in a WSS process $X(t)$ is the sum of the powers at all frequencies, and therefore it is the integral of the power spectrum over all frequencies. We can write

$$P_X = E\left\{|X(t)|^2\right\} = R_X(0) = \int_{-\infty}^{\infty} \mathcal{S}_X(f)df \qquad (2\text{-}102)$$

2.8.2 Gaussian Random Processes

A real random process $X(t)$ is Gaussian if for all positive integers n and for all $(t_1, t_2, \cdots, t_n)$, the random vector $(X(t_1), X(t_2), \cdots, X(t_n))^T$ is a Gaussian random vector; *i.e.*, random variables $\{X(t_i)\}_{i=1}^n$ are jointly Gaussian random variables. Similar to Jointly Gaussian random variables, linear filtering of Gaussian random processes results in a Gaussian random process, even when the filtering is time-varying.

Two real random processes $X(t)$ and $Y(t)$ are jointly Gaussian if for all positive integers n, m and all $(t_1, t_2, \cdots, t_n)$ and $(t'_1, t'_2, \cdots, t'_n)$ the random vector

$$(X(t_1), X(t_2), \cdots, X(t_n), Y(t'_1), Y(t'_2), \cdots, Y(t'_n))^T$$

is a Gaussian vector. For two jointly Gaussian random processes $X(t)$ and $Y(t)$, being uncorrelated, *i.e.*, having

$$R_{XY}(t + \tau, t) = E\{X(t + \tau)\}E\{Y(t)\} \quad \text{for all } t \text{ and } \tau \qquad (2\text{-}103)$$

is equivalent to being independent.

A complex process $Z(t) = X(t) + jY(t)$ is Gaussian if $X(t)$ and $Y(t)$ are jointly Gaussian processes.

2.8.3 White Processes*

A process is called a *white process* if its power spectral density is constant for all frequencies; this constant value is usually denoted by $\frac{N_0}{2}$.

$$\mathcal{S}_X(f) = \frac{N_0}{2} \tag{2-104}$$

Using Equation 2-102, we see that the power in a white process is infinite, indicating that white processes cannot exist as a physical process. Although white processes are not physically realizable processes, they are very useful, closely modeling some important physical phenomenon including the *thermal noise*.

Thermal noise is the noise generated in electric devices by thermal agitation of electrons. Thermal noise can be closely modeled by a random process $N(t)$ having the following properties:

1. $N(t)$ is a stationary process.
2. $N(t)$ is a zero-mean process.
3. $N(t)$ is a Gaussian process.
4. $N(t)$ is a white process whose power spectral density is given by

$$\mathcal{S}_N(f) = \frac{N_0}{2} = \frac{kT}{2} \tag{2-105}$$

where T is the ambient temperature in kelvins and k is *Boltzmann's constant*, equal to 38×10^{-23} J/K.

Discrete-Time Random Processes

Discrete-time random processes have similar properties to continuous time processes. In particular the PSD of a WSS discrete-time random process is defined as the discrete-time Fourier transform of its autocorrelation function

$$\mathcal{S}_X(f) = \sum_{m=-\infty}^{\infty} R_X(m) e^{-j2\pi fm} \tag{2-106}$$

and the autocorrelation function can be obtained as the inverse Fourier transform of the power spectral density as

$$R_X(m) = \int_{-1/2}^{1/2} \mathcal{S}_X(f) e^{j2\pi fm} df \tag{2-107}$$

The power in a discrete-time random process is given by

$$P = E\left\{|X(n)|^2\right\} = R_X(0) = \int_{-1/2}^{1/2} \mathcal{S}_X(f) df \tag{2-108}$$

Problems

Problem 2.8.1. Proof Theorem 1.

Problem 2.8.2. State some matrix decomposition methods and the conditions to apply for each method.

Problem 2.8.3. Let X_I and X_Q be statistically independent zero-mean Gaussian random variables with identical variance. Show that a (rotational) transformation of the form

$$Y_I + jY_Q = (X_I + jX_Q)e^{j\phi}$$

results in another pair (Y_I, Y_Q) of Gaussian random variables that have the same joint PDF as the pair (X_I, X_Q).

Problem 2.8.4. Show that if $X(t)$ is a real, wide-sense stationary process, the following properties of the autocorrelation function and power spectral density hold:

1. $R_X(\tau) = R_X(-\tau)$
2. The power spectrum $\mathcal{S}_X(f)$ is real and even.

Chapter 3

Optimum Receivers for AWGN Channels

In Chapter 1 we have seen that communication channels can suffer from a variety of impairments that contribute to errors. These impairments include noise, attenuation, distortion, fading, and interference. Characteristics of a communication channel deter mine which impairments apply to that particular channel and which are the determining factors in the performance of the channel. Noise is present in all communication channels and is the major impairment in many communication systems.

In particular, this chapter deals with the design and performance characteristics of optimum receivers for the various modulation methods when the channel corrupts the transmitted signal by the addition of white Gaussian noise.

3.1 Waveform and Vector Channel Models

The additive white Gaussian noise (AWGN) channel model is a channel whose sole esffect is addition of a white Gaussian noise process to the transmitted signal. This channel is mathematically described by the relation

$$r(t) = s_m(t) + n(t) \tag{3-1}$$

where $s_m(t)$ is the transmitted signal; $n(t)$ is a sample waveform of a zero-mean white Gaussian noise process with power spectral density of $N_0/2$; and $r(t)$ is the received waveform. This channel model is shown in Figure 3-1.

The receiver observes the received signal $r(t)$ and, based on this observation, makes the optimal decision about which message m, $1 \leq m \leq M$, was transmitted. By an optimal decision we mean a decision rule which results in minimum error probability, *i.e.*, the decision rule that minimizes the probability of disagreement between the transmitted message m and the detected message $\hat{m}$ given by

$$P_e = \Pr\{\hat{m} \neq m\} \tag{3-2}$$

Although the AWGN channel model seems very limiting, its study is beneficial from two points of view.

1. First, noise is the major type of corruption introduced by many channels. Therefore isolating it from other channel impairments and studying its effect results in better understanding of its effect on all communication systems.

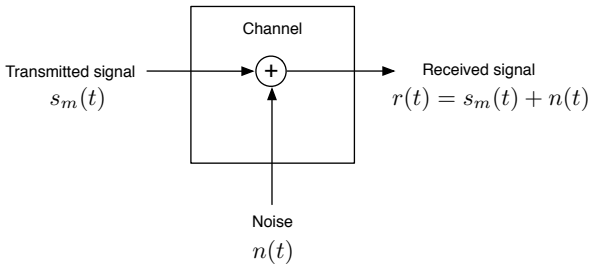

Figure 3-1: Model for received signal passed through an AWGN channel.

2. Second the AWGN channel, although very simple, is a good model for studying deep space communication channels which were historically one of the first challenges encountered by communication engineers.

By using an orthonormal basis $\{\phi_j(t), 1 \leq j \leq N\}$, each signal $s_m(t)$ can be represented by a vector $\boldsymbol{s}_m \in \mathbb{R}^N$. Any orthonormal basis can be used for expansion of a zero-mean white Gaussian process, and the resulting coefficients of expansion will be *i.i.d.* zero-mean Gaussian random variables with variance $N_0/2$. Therefore, $\{\phi_j(t), 1 \leq j \leq N\}$, when extended appropriately, can be used for expansion of the noise process $n(t)$. This observation prompts us to view the waveform channel $r(t) = s_m(t) + n(t)$ in the vector form $\boldsymbol{r} = \boldsymbol{s}_m + \boldsymbol{n}$ where all vectors are N-dimensional and components of $\boldsymbol{n}$ are *i.i.d.* zero-mean Gaussian random variables with variance $N_0/2$. We will give a rigorous proof of this equivalence in Section 3.2. We continue our analysis with the study of the vector channel introduced above.

3.1.1 Optimal Detection for a General Vector Channel

The mathematical model for the AWGN vector channel is given by

$$\boldsymbol{r} = \boldsymbol{s}_m + \boldsymbol{n} \tag{3-3}$$

where all vectors are N-dimensional real vectors. The message m is chosen according to probabilities P_m, from the set of possible messages $\{1, 2, \cdots, M\}$. The noise components n_j, $1 \leq j \leq N$, are i.i.d., zero-mean, Gaussian random variables each distributed according to $\mathcal{N}(0, N_0/2)$. Therefore, the PDF of the noise vector $\boldsymbol{n}$ is given by

$$p(\boldsymbol{n}) = \left(\frac{1}{\sqrt{\pi N_0}}\right)^N e^{-\frac{\sum_{j=1}^N n_j^2}{2\sigma^2}} = \left(\frac{1}{\sqrt{\pi N_0}}\right)^N e^{-\frac{\|\boldsymbol{n}\|^2}{N_0}} \tag{3-4}$$

We, however, study a more general vector channel model in this section which is not limited to the AWGN channel model. This model will later be specialized to an AWGN channel model in Section 3.2. In our model, vectors $\boldsymbol{s}_m$ are selected from a set of possible signal vectors $\{\boldsymbol{s}_m, 1 \leq m \leq M\}$ according to prior or a priori probabilities P_m and transmitted over the channel. The received vector $\boldsymbol{r}$ depends statistically on the transmitted vector through the conditional probability density function $p(\boldsymbol{r}|\boldsymbol{s}_m)$. The channel model is shown in Figure 3-2.

The receiver observes $\boldsymbol{r}$ and based on this observation decides which message was transmitted. Let us denote the decision function employed at the receiver by $g(\boldsymbol{r})$, which is a function from $\mathbb{R}^N$ into the set of messages $\{1, 2, \cdots, M\}$. Now if $g(\boldsymbol{r}) = \hat{m}$, *i.e.*, the receiver decides that $\hat{m}$

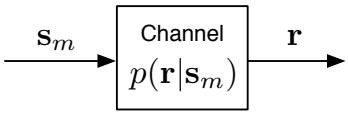

Figure 3-2: A general vector channel.

was transmitted, then the probability that this decision is correct is the probability that $\hat{m}$ was in fact the transmitted message. In other words, the probability of a correct decision, given that $\boldsymbol{r}$ is received, is given by

$$\Pr\{\text{correct decision}|\boldsymbol{r}\} = \Pr\{\hat{m} \text{ sent}|\boldsymbol{r}\} \tag{3-5}$$

and therefore the probability of a correct decision is

$$\begin{aligned} \Pr\{\text{correct decision}\} &= \int \Pr\{\text{correct decision}|\boldsymbol{r}\}\, p(\boldsymbol{r}) d\boldsymbol{r} \\ &= \int \Pr\{\hat{m} \text{ sent}|\boldsymbol{r}\}\, p(\boldsymbol{r}) d\boldsymbol{r} \end{aligned} \tag{3-6}$$

Our goal is to design an optimal detector that minimizes the error probability or, equivalently, maximizes $\Pr\{\text{correct decision}\}$. Since $p(\boldsymbol{r})$ is nonnegative for all $\boldsymbol{r}$, the right-hand side of Equation 3-6 is maximized if for each $\boldsymbol{r}$ the quantity $\Pr\{\hat{m}|\boldsymbol{r}\}$ is maximized. This means that the optimal detection rule is the one that upon observing $\boldsymbol{r}$ decides in favor of the message m that maximizes $\Pr\{m|\boldsymbol{r}\}$. In other words,

$$\hat{m} = g_{\text{opt}}(\boldsymbol{r}) = \arg\max_{1 \le m \le M} \Pr\{m|\boldsymbol{r}\} \tag{3-7}$$

The optimal detection scheme described in Equation 3-7 simply looks among all $\Pr\{m|\boldsymbol{r}\}$ for $1 \le m \le M$ and selects the m that maximizes $\Pr\{m|\boldsymbol{r}\}$. The detector then declares this maximizing m as its best decision. Note that since transmitting message m is equivalent to transmitting $\boldsymbol{s}_m$, the optimal decision rule can be written as

$$\hat{m} = g_{\text{opt}}(\boldsymbol{r}) = \arg\max_{1 \le m \le M} \Pr\{\boldsymbol{s}_m|\boldsymbol{r}\} \tag{3-8}$$

MAP and ML Receivers

The optimal decision rule given by Equations 3-7 and 3-8 is known as the maximum a posteriori probability rule, or MAP rule. Note that the MAP receiver can be simplified to

$$\hat{m} = \arg\max_{1 \le m \le M} \frac{P_m p(\boldsymbol{r}|\boldsymbol{s}_m)}{p(\boldsymbol{r})} \tag{3-9}$$

and since $p(\boldsymbol{r})$ is independent of m and for all m remains the same, this is equivalent to

$$\hat{m} = \arg\max_{1 \le m \le M} P_m p(\boldsymbol{r}|\boldsymbol{s}_m) \tag{3-10}$$

Equation 3-10 is easier to use than Equation 3-7 since it is given in terms of the prior probabilities P_m and the probabilistic description of the channel $p(\boldsymbol{r}|\boldsymbol{s}_m)$, both directly known.

In the case where the messages are equiprobable a priori, *i.e.*, when $P_m = \frac{1}{M}$ for all $1 \leq m \leq M$, the optimal detection rule reduces to

$$\hat{m} = \arg \max_{1 \leq m \leq M} p(\boldsymbol{r}|\boldsymbol{s}_m) \tag{3-11}$$

The term $p(\boldsymbol{r}|\boldsymbol{s}_m)$ is called the *likelihood of message* m, and the receiver given by Equation 3-11 is called the *maximum-likelihood receiver*, or ML receiver. It is important to note that the ML detector is not an optimal detector unless the messages are equiprobable. The ML detector, however, is a very popular detector since in many cases having exact information about message probabilities is difficult.

The Decision Regions

Any detector-including MAP and ML detectors-partitions the output space $\mathbb{R}^N$ into M regions denoted by $\mathbb{D}_1, \mathbb{D}_2, \cdots, \mathbb{D}_M$, such that if $\boldsymbol{r} \in \mathbb{D}_m$, then $\hat{m} = g(\boldsymbol{r}) = m$, *i.e.*, the detector makes a decision in favor of m. The region $\mathbb{D}_m$, $1 \leq m \leq M$ is called the decision region for message m; and $\mathbb{D}_m$ is the set of all outputs of the channel that are mapped into message m by the detector. If a MAP detector is employed, then the $\mathbb{D}_m$'s constitute the optimal decision regions resulting in the minimum possible error probability. For a MAP detector we have

$$\mathbb{D}_m = \left\{ \boldsymbol{r} \in \mathbb{R}^N | \Pr\{m|\boldsymbol{r}\} > \Pr\left\{m'|\boldsymbol{r}\right\}, \text{ for all } 1 \leq m' \leq M \text{ and } m' \neq m \right\} \tag{3-12}$$

Note that if for some given $\boldsymbol{r}$ two or more messages achieve the maximum a posteriori probability, we can arbitrarily assign $\boldsymbol{r}$ to one of the corresponding decision regions.

The Error Probability

To determine the error probability of a detection scheme, we note that when $\boldsymbol{s}_m$ is transmitted, an error occurs when the received $\boldsymbol{r}$ is not in $\mathbb{D}_m$. The symbol error probability of a receiver with decision regions $\{\mathbb{D}_m, 1 \leq m \leq M\}$ is therefore given by

$$\begin{aligned} P_e &= \sum_{m=1}^{M} P_m \Pr\{\boldsymbol{r} \notin \mathbb{D}_m | \boldsymbol{s}_m \text{ sent}\} \\ &= \sum_{m=1}^{M} P_m P_{e|m} \end{aligned} \tag{3-13}$$

where $P_{e|m}$ denotes the error probability when messege m is transmitted and is given by

$$\begin{aligned} P_{e|m} &= \int_{\mathbb{D}_m^c} p(\boldsymbol{r}|\boldsymbol{s}_m) d\boldsymbol{r} \\ &= \sum_{1 \leq m' \leq M, m' \neq m} \int_{\mathbb{D}_{m'}} p(\boldsymbol{r}|\boldsymbol{s}_m) d\boldsymbol{r} \end{aligned} \tag{3-14}$$

Using Equation 3-14 in Equation 3-13 gives

$$P_e = \sum_{m=1}^{M} P_m \sum_{1 \leq m' \leq M, m' \neq m} \int_{\mathbb{D}_{m'}} p(\boldsymbol{r}|\boldsymbol{s}_m) d\boldsymbol{r} \tag{3-15}$$

Equation 3-15 gives the probability that an error occurs in transmission of a symbol or a message and is called *symbol error probability* or *message error probability*. Another type of error probability is the *bit error probability*. This error probability is denoted by P_b and is the error probability in transmission of a single bit. Determining the bit error probability in general requires detailed know ledge of how different bit sequences are mapped to signal points. Therefore, in general finding the bit error probabiility is not easy unless the constellation exhibits certain symmetry properties to make the derivation of the bit error probability easy. We will see later in this chapter that orthogonal signaling exhibits the required symmetry for calculation or the bit error probability. In other cases we can bound the bit error probability by noting that a symbol error occurs when at least one bit is in error, and the event of a symbol error is the union of the events of the errors in the $k = \log_2 M$ bits representing that symbol. Therefore we can write

$$P_b \leq P_e \leq kP_b \tag{3-16}$$

or

$$\frac{P_e}{\log_2 M} \leq P_e \leq P_e \tag{3-17}$$

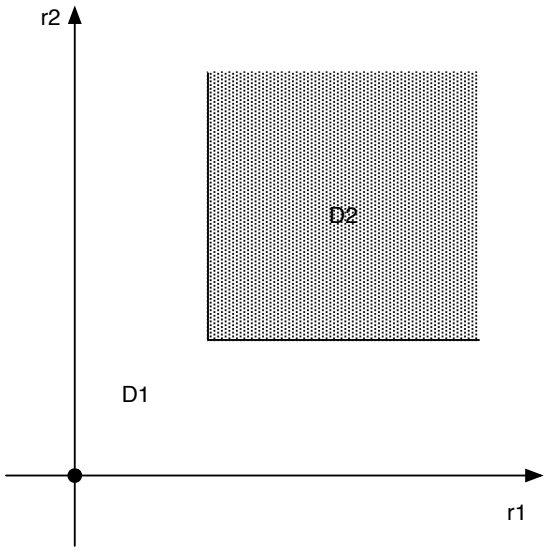

Figure 3-3: Decision regions $\mathbb{D}_1$ and $\mathbb{D}_2$.

Example 1. Consider two equiprobable message signals $\boldsymbol{s}_1 = (0, 0)$ and $\boldsymbol{s}_2 = (1, 1)$. The channel adds *i.i.d.* noise components n_1 and n_2 to the transmitted vector each with an exponential PDF of the form

$$p(n) = \begin{cases} e^{-n} & n \geq 0 \\ 0 & n < 0 \end{cases}$$

Since the messages are equiprobable, the MAP detector is equi valent to the ML detector, and the decision region $\mathbb{D}_1$ is given by

$$\mathbb{D}_1 = \{\boldsymbol{r} \in \mathbb{R}^2 : p(\boldsymbol{r}|\boldsymbol{s}_1) > p(\boldsymbol{r}|\boldsymbol{s}_2)\}$$

Noting that $p(\boldsymbol{r}|\boldsymbol{s} = (s_1, s_2)) = p(\boldsymbol{n} = \boldsymbol{r} - \boldsymbol{s})$, we have

$$\mathbb{D}_1 = \{\boldsymbol{r} \in \mathbb{R}^2; p_{\boldsymbol{n}}(r_1, r_2) > p_{\boldsymbol{n}}(r_1 - 1, r_2 - 1)\}$$

where

$$p_{\boldsymbol{n}}(n_1, n_2) = \begin{cases} e^{-n_1 - n_2} & n_1, n_2 \geq 0 \\ 0 & n < 0 \end{cases}$$

From this relation we conclude that if either r_1 or r_2 is less than 1, then the point $\boldsymbol{r}$ belongs to $\mathbb{D}_1$, and if both r_1 and r_2 are greater than 1, we have $e^{-r_1 - r_2} < e^{-(r_1 - 1) - (r_2 - 1)}$ and $\boldsymbol{r}$ belongs to $\mathbb{D}_2$.

Note that in this channel neither r_1 nor r_2 can be negative, because signal and noise are always nonnegative. Therefore,

$$\mathbb{D}_2 = \{\boldsymbol{r} \in \mathbb{R}^2 : r_1 \geq 1, r_2 \geq 1\}$$

and

$$\mathbb{D}_1 = \{\boldsymbol{r} \in \mathbb{R}^2 : r_1, r_2 \geq 0, \text{ either } 0 \leq r_1 \leq 1 \text{ or } 0 \leq r_2 < 1\}$$

The decision regions are shown in Figure 3-3. For this channel, when $\boldsymbol{s}_2$ is transmitted, regardless of the value of noise components, $\boldsymbol{r}$ will always be in $\mathbb{D}_2$ and no error will occur.

Errors will occur only when $\boldsymbol{s}_1 = (0, 0)$ is transmitted and the received vector $\boldsymbol{r}$ belongs to $\mathbb{D}_2$, *i.e.*, when both noise components exceed 1. Therefore, the error probability is given by

$$\begin{aligned} P_e &= \frac{1}{2} \Pr\{\boldsymbol{r} \in \mathbb{D}_2 | \boldsymbol{s}_1 = (0, 0) \text{ sent}\} \\ &= \frac{1}{2} \int_1^\infty e^{-n_1} dn_1 \int_1^\infty e^{-n_2} dn_2 \\ &= \frac{1}{2} e^{-2} \approx 0.0068 \end{aligned}$$

Sufficient Statistics

Let us assume that at the receiver we have access to a vector $\boldsymbol{r}$ that can be written in terms of two vectors $\boldsymbol{r}_1$ and $\boldsymbol{r}_2$, *i.e.*, $\boldsymbol{r} = (\boldsymbol{r}_1, \boldsymbol{r}_2)$. We further assume that $\boldsymbol{s}_m$, $\boldsymbol{r}_1$ and $\boldsymbol{r}_2$ constitute a Markov chain in the given order, *i.e.*,

$$p(\boldsymbol{r}_1, \boldsymbol{r}_2 | \boldsymbol{s}_m) = p(\boldsymbol{r}_1 | \boldsymbol{s}_m) p(\boldsymbol{r}_2 | \boldsymbol{r}_1) \tag{3-18}$$

Under these assumptions $\boldsymbol{r}_2$ can he ignored in the dctcction of $\boldsymbol{s}_m$, and the detection can be based only on $\boldsymbol{r}_1$. The reason is that by Equation 3-10

$$\begin{aligned}\hat{m} &= \arg\max_{1\le m\le M} P_m p(\boldsymbol{r}|\boldsymbol{s}_m) \\ &= \arg\max_{1\le m\le M} P_m p(\boldsymbol{r}_1,\boldsymbol{r}_2|\boldsymbol{s}_m) \\ &= \arg\max_{1\le m\le M} P_m p(\boldsymbol{r}_1|\boldsymbol{s}_m)p(\boldsymbol{r}_2|\boldsymbol{r}_1) \\ &= \arg\max_{1\le m\le M} P_m p(\boldsymbol{r}_1|\boldsymbol{s}_m) \end{aligned} \tag{3-19}$$

where in the last step we have ignored the positive factor $p(\boldsymbol{r}_2|\boldsymbol{r}_1)$ since it does not depend on m. This shows that the optimal detection can be based only on $\boldsymbol{r}_1$.

When the Markov chain relation among $\boldsymbol{s}_m$, $\boldsymbol{r}_1$ and $\boldsymbol{r}_2$ as given in Equation 3-18 is satisfied, it is said that $\boldsymbol{r}_1$ is a *sufficient statistic* for detection of $\boldsymbol{s}_m$. In such a case, when $\boldsymbol{r}_2$ can be ignored without sacrificing the optimality of the receiver, $\boldsymbol{r}_2$ is called *irrelevant data* or *irrelevant information*. Recognizing sufticient statistics helps to reduce the complexity of the detection process through ignoring a usually large amount of irrelevant data at the receiver.

Example 2. Let us assume that in Example 1, in addition to $\boldsymbol{r}$, the receiver can observe n_1 as well. Therefore, we can assume that $\boldsymbol{r} = (\boldsymbol{r}_1, \boldsymbol{r}_2)$ is available at the receiver, where $\boldsymbol{r}_1 = (r_1, n_1)$ and $\boldsymbol{r}_2 = r_2$. To design the optimal detector, we notice that having access to both r_1 and n_1 uniquely determines $\boldsymbol{s}_{m1}$ at the receiver; and since $s_{11} = 0$ and $s_{21} = 1$, this uniquely determines the message m, thus making $\boldsymbol{r}_2 = r_2$ irrelevant. The optimal decision rule in this case becomes

$$\hat{m} = \begin{cases} 1 & \text{if } r_1 - n_1 = 0 \\ 2 & \text{if } r_1 - n_1 = 1 \end{cases} \tag{3-20}$$

and the resulting error probability is zero.

Preprocessing at the Receiver

Let us assume that the receiver applies an invertible operation $\mathscr{G}(\cdot)$(denoted as $\boldsymbol{G}(\cdot)$ in Figure 3-4) on the received vector $\boldsymbol{r}$. In other words instead of supplying $\boldsymbol{r}$ to the detector, the receiver passes $\boldsymbol{r}$ through $\mathscr{G}$ and supplies the detector with $\boldsymbol{\rho} = \mathscr{G}(\boldsymbol{r})$, as shown in Figure 3-4.

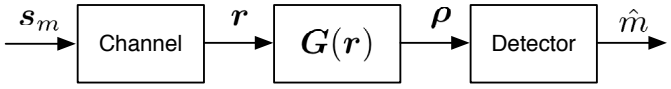

Figure 3-4: Preprocessing at the receiver.

Since $\mathscr{G}$ is invertible and the detector has access to $\boldsymbol{\rho}$, it can apply $\mathscr{G}^{-1}$ to $\boldsymbol{\rho}$ to obtain $\mathscr{G}^{-1} = \mathscr{G}^{-1}(\mathscr{G}(\boldsymbol{r})) = \boldsymbol{r}$. The detector now has access to both $\boldsymbol{\rho}$ and $\boldsymbol{r}$; therefore the optimal

detection rule is

$$\begin{aligned} \hat{m} &= \arg\max_{1\le m\le M} P_m p(\boldsymbol{r},\boldsymbol{\rho}|\boldsymbol{s}_m) \\ &= \arg\max_{1\le m\le M} P_m p(\boldsymbol{r}|\boldsymbol{s}_m) p(\boldsymbol{\rho}|\boldsymbol{r}) \\ &= \arg\max_{1\le m\le M} P_m p(\boldsymbol{r}|\boldsymbol{s}_m) \end{aligned} \tag{3-21}$$

where we have used the fact that $\boldsymbol{\rho}$ is a function of $\boldsymbol{r}$ and hence, when $\boldsymbol{r}$ is given, $\boldsymbol{\rho}$ does not depend on $\boldsymbol{s}_m$. From Equation 3-22 it is clear that the optimal detector based on the observation of $\boldsymbol{\rho}$ makes the same decision as the optimal detector based on the observation of $\boldsymbol{r}$. In other words. an invertible preprocessing of the received information does not change the optimality of the receiver.

Example 3. Let us assume the received vector is of the form

$$\boldsymbol{r} = \boldsymbol{s}_m + \boldsymbol{n}$$

where $\boldsymbol{n}$ is a nonwhite (colored) noise. Let us further assume that there exists an invertible whitening operator denoted by matrix $\boldsymbol{W}$ such that $\boldsymbol{\nu} = \boldsymbol{W}\boldsymbol{n}^T$ is a white vector. Then we can consider

$$\boldsymbol{\rho} = \boldsymbol{W}\boldsymbol{r}^T = \boldsymbol{W}\boldsymbol{s}_m^T + \boldsymbol{\nu}$$

which is equivalent to a channel with white noise for detection without degrading the performance. The linear operation denoted by $\boldsymbol{W}$ is called a *whitening filter*.

3.2 Waveform and Vector AWGN Channels

The waveform AWGN channel is described by the input-output relation

$$r(t) = s_m(t) + n(t) \tag{3-22}$$

where $s_m(t)$ is one of the possible M signals $\{s_1(t), s_2(t), \cdots, s_M(t)\}$, each selected with prior probability P_m and $n(t)$ is a zero-mean white Gaussian process with power spectral density $\frac{N_0}{2}$. Let us assume that using the Gram-Schmidt procedure, we have derived an orthonormal basis $\{\phi_j(t), 1 \leq j \leq N\}$ for representation of the signals and, using this set, the vector representation of the signals is given by $\{\boldsymbol{s}_m, 1 \leq m \leq M\}$. The noise process cannot be completely expanded in terms of the basis $\{\phi_j(t)\}_{j=1}^N$. We decompose the noise process $n(t)$ into two components. One component, denoted by $n_1(t)$ is part of the noise process that can be expanded in terms of $\{\phi_j(t)\}_{j=1}^N$, *i.e.*, the projection of the noise onto the space spanned by these basis functions; and the other part, denoted by $n_2(t)$, is the part that cannot be expressed in terms of this basis function. With this definition we have

$$n_1(t) = \sum_{j=1}^{N} n_j \phi_j(t), \quad n_j = \langle n(t), \phi_j(t) \rangle \tag{3-23}$$

and

$$n_2(t) = n(t) - n_1(t) \tag{3-24}$$

Noting that

$$s_m(t) = \sum_{j=1}^{N} s_{m,j} \phi_j(t), \quad s_{mj} = \langle s_m(t), \phi_j(t) \rangle \tag{3-25}$$

and using Equations 3-23 and 3-24, we can write Equation 3-22 as

$$r(t) = \sum_{j=1}^{N} (s_{m,j} + n_j)\phi_j(t) + n_2(t) \tag{3-26}$$

By defining

$$r_j = s_{m,j} + n_j \tag{3-27}$$

where

$$r_j = \langle s_m(t), \phi_j(t) \rangle + \langle n(t), \phi_j(t) \rangle = \langle s_m(t) + n(t), \phi_j(t) \rangle = \langle r(t), \phi_j(t) \rangle \tag{3-28}$$

we have

$$r(t) = \sum_{j=1}^{N} r_j \phi_j(t) + n_2(t), \quad r_j = \langle r(t), \phi_j(t) \rangle \tag{3-29}$$

We know that n_j's are *i.i.d.* zero-mean Gaussian random variables each with variance $\frac{N_0}{2}$. This result can also be directly shown, by noting that the n_j's defined by

$$n_j = \int_{-\infty}^{\infty} n(t)\phi_j(t)dt \tag{3-30}$$

are linear combinations of the Gaussian random process $n(t)$, and therefore they are Gaussian. Their mean is given by

$$\begin{aligned} E\{n_j\} &= E\left\{\int_{-\infty}^{\infty} n(t)\phi_j(t)dt\right\} \\ &= \int_{-\infty}^{\infty} E\{n(t)\}\,\phi_j(t)dt \\ &= 0 \end{aligned} \tag{3-31}$$

where the last equality holds since $n(t)$ is zero-mean, *i.e.*, $E\{n(t)\} = 0$.

We can also find the covariance of n_i and n_j as

$$\begin{aligned} \mathrm{Cov}\{n_i n_j\} &= E\{n_i n_j\} - E\{n_i\}E\{n_j\} \\ &= E\left\{\int_{-\infty}^{\infty} n(t)\phi_i(t)dt \int_{-\infty}^{\infty} n(t)\phi_j(s)ds\right\} \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} E\{n(t)n(s)\}\,\phi_i(t)\phi_j(s)dtds \\ &= \frac{N_0}{2}\int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty}\delta(s-t)\phi_i(t)dt\right]\phi_j(s)ds \\ &= \frac{N_0}{2}\int_{-\infty}^{\infty}\phi_i(s)\phi_j(s)ds \\ &= \begin{cases} \frac{N_0}{2} & i = j \\ 0 & i \neq j \end{cases} \end{aligned} \tag{3-32}$$

where we have used the facts that n_i and n_j are zero-mean, and since $n(t)$ is white, its autocorrelation function is $\frac{N_0}{2}\delta(\tau)$. In the last step we applied the orthonormality of $\{\phi_j(t)\}$. Equation 3-32 shows that for $i \neq j$, n_i and n_j are uncorrelated and since they are Gaussian, they are independent as well. It also shows that each n_j has a variance equal to $\frac{N_0}{2}$.

Now we study the properties of $n_2(t)$. We first observe that since the n_j's are jointly Gaussian random variables, the process $n_1(t)$ is a Gaussian process and thus $n_2(t) = n(t) - n_1(t)$, which is a linear combination of two jointly Gaussian processes, is itself a Gaussian process. At any given t we have

$$\begin{aligned} \mathrm{Cov}\{n_j n_2\} &= E\{n_j n_2\} \\ &= E\{n_j n\} - E\{n_j n_1\} \\ &= E\left\{n(t)\int_{-\infty}^{\infty} n(s)\phi_j(s)ds\right\} - E\left\{n_j\sum_{i=1}^{N} n_i\phi_i(t)\right\} \\ &= \frac{N_0}{2}\int_{-\infty}^{\infty}\delta(t-s)\phi_j(s)ds - \frac{N_0}{2}\phi_j(t) \\ &= \frac{N_0}{2}\phi_j(t) - \frac{N_0}{2}\phi_j(t) \\ &= 0 \end{aligned} \tag{3-33}$$

where we have used the fact that $E\{n_j n_i\} = 0$, except when $i = j$, in which case $E\{n_j n_j\} = \frac{N_0}{2}$.

Equation 3-33 shows that $n_2(t)$ is uncorrelated with all n_j's, and since they are jointly Gaussian, $n_2(t)$ is independent of all n_j's, and therefore it is independent of $n_1(t)$.

Since $n_2(t)$ is independent of $s_m(t)$ and $n_1(t)$, we conclude that in Equation 3-29, the two components of $r(t)$, namely, $\sum_j r_j\phi_j(t)$ and $n_2(t)$, are independent. Since the first component

is the only component that carries the transmitted signal, and the second component is independent of the first component, the second component cannot provide any information about the transmitted signal and therefore has no effect in the detection process and can be ignored without sacrificing the optimality of the detector. In other words $n_2(t)$ is irrelevant information for optimal detection.

From the above discussion it is clear that for the design of the optimal detector, the AWGN waveform channel of the form

$$r(t) = s_m(t) + n(t), \quad 1 \leq m \leq M \tag{3-34}$$

is equivalent to the N-dimensional vector channel

$$\boldsymbol{r} = \boldsymbol{s}_m + \boldsymbol{n}, \quad 1 \leq m \leq M \tag{3-35}$$

3.2.1 Optimal Detection for the Vector AWGN Channel

The additive AWGN vector channel is the vector equivalent channel to the waveform AWGN channel and is described by Equation 3-1 in which the components of the noise vector are i.i.d. zero-mean Gaussian random variables with variance $N_0/2$. The joint PDF of the noise vector is given by Equation 3-4. The MAP detector for this channel is given by

$$\begin{aligned}
\hat{m} &= \underset{1\leq m\leq M}{\arg\max} \{P_m p(\boldsymbol{r}|\boldsymbol{s}_m)\} \\
&= \underset{1\leq m\leq M}{\arg\max} \{P_m p_{\boldsymbol{n}}(\boldsymbol{r}-\boldsymbol{s}_m)\} \\
&= \underset{1\leq m\leq M}{\arg\max} \left\{ P_m \left(\frac{1}{\sqrt{\pi N_0}}\right)^N e^{-\frac{\|\boldsymbol{r}-\boldsymbol{s}_m\|^2}{N_0}} \right\} \\
&\overset{(a)}{=} \underset{1\leq m\leq M}{\arg\max} \left\{ P_m e^{-\frac{\|\boldsymbol{r}-\boldsymbol{s}_m\|^2}{N_0}} \right\} \\
&\overset{(b)}{=} \underset{1\leq m\leq M}{\arg\max} \left\{ \ln P_m - \frac{\|\boldsymbol{r}-\boldsymbol{s}_m\|^2}{N_0} \right\} \\
&\overset{(c)}{=} \underset{1\leq m\leq M}{\arg\max} \left\{ \frac{N_0}{2}\ln P_m - \frac{1}{2}\|\boldsymbol{r}-\boldsymbol{s}_m\|^2 \right\} \\
&= \underset{1\leq m\leq M}{\arg\max} \left\{ \frac{N_0}{2}\ln P_m - \frac{1}{2}\left(\|\boldsymbol{r}\|^2 + \|\boldsymbol{s}_m\|^2 - 2\boldsymbol{r}\cdot\boldsymbol{s}_m^T\right) \right\} \\
&\overset{(d)}{=} \underset{1\leq m\leq M}{\arg\max} \left\{ \frac{N_0}{2}\ln P_m - \frac{1}{2}\mathcal{E}_m + \boldsymbol{r}\cdot\boldsymbol{s}_m^T) \right\} \\
&\overset{(e)}{=} \underset{1\leq m\leq M}{\arg\max} \left\{ \eta_m + \boldsymbol{r}\cdot\boldsymbol{s}_m^T \right\}
\end{aligned} \tag{3-36}$$

where we have used the following steps in simplifying the expression:

(a) $\left(\frac{1}{\sqrt{\pi N_0}}\right)^N$ is a positive constant and can be dropped.

(b) $\ln(\cdot)$ is an increasing function.

(c) $N_0/2$ is positive and multiplying by a positive number docs not affect the result of arg max.

(d) $\|\boldsymbol{r}\|^2$ was dropped since it docs not depend on m and $\|\boldsymbol{s}_m\|^2 = \mathcal{E}_m$.

(e) We have defined

$$\eta_m = \frac{N_0}{2} \ln P_m - \frac{1}{2}\mathcal{E}_m \tag{3-37}$$

as the *basis term*.

From Equation 3-36, it is clear that the optimal (MAP) decision rule for an AWGN vector channel is given by

$$\begin{aligned} \hat{m} &= \arg\max_{1\le m\le M} \left\{\eta_m + \boldsymbol{r}\cdot\boldsymbol{s}_m^T\right\} \\ \eta_m &= \frac{N_0}{2}\ln P_m - \frac{1}{2}\mathcal{E}_m \end{aligned} \tag{3-38}$$

In the special case where the signal $\boldsymbol{s}$ are equiprobable, *i.e.*, $P_m = 1/M$ for all m, this relation becomes somewhat simpler. In this case Equation 3-15 at step (c) can be written as

$$\begin{aligned} \hat{m} &= \arg\max_{1\le m\le M} \left\{\frac{N_0}{2}\ln P_m - \frac{1}{2}\|\boldsymbol{r}-\boldsymbol{s}_m\|^2\right\} \\ &= \arg\max_{1\le m\le M} \left\{-\|\boldsymbol{r}-\boldsymbol{s}_m\|^2\right\} \\ &= \arg\min_{1\le m\le M} \left\{\|\boldsymbol{r}-\boldsymbol{s}_m\|\right\} \end{aligned} \tag{3-39}$$

where we have used the fact that maximizing $-\|\boldsymbol{r}-\boldsymbol{s}_m\|^2$ is equivalent to minimizing its negative, *i.e.*, $\|\boldsymbol{r}-\boldsymbol{s}_m\|^2$, which is equivalent to minimizing its square root $\|\boldsymbol{r}-\boldsymbol{s}_m\|$.

A geometric interpretation of Equation 3-39 is particularly convenient. The receiver receives $\boldsymbol{r}$ and looks among all $\boldsymbol{s}_m$ to find the one that is closest to $\boldsymbol{r}$ using standard Euclidean distance. Such a detector is called a *nearest-neighbor* or *minimum-distance*, detector. Also note that in this case, since the signals are equiprobable, the MAP and the ML detector coincide, and both are equivalent to the minimum-distance detector. In this case the boundaries of decisions $\mathbb{D}_m$ and $\mathbb{D}_{m'}$ are the set of points that are equidistant from $\boldsymbol{s}_m$ and $\boldsymbol{s}_{m'}$, which is the perpendicular bisector of the line connecting these two signal points. This boundary in general is a hyperplane. For the case of $N = 2$ the boundary is a line, and for $N = 3$ it is a plane. These hyperplanes completely determine the decision regions. An example of a two-dimensional constellation ($N = 2$) with four signal points ($M = 4$) is show n in Figure 3-5. The solid lines denote the boundaries of the decision regions which are the perpendicular bisectors of the dashed lines connecting the signal points.

When the signals are both equiprobable and have equal energy, the bias terms defined as $\eta_m = \frac{N_0}{2}\ln P_m - \frac{1}{2}\mathcal{E}_m$ are independent of m and can be dropped from Equation 3-38. The optimal detection rule in this case reduces to

$$\hat{m} = \arg\max_{1\le m\le M} \left\{\boldsymbol{r}\cdot\boldsymbol{s}_m^T\right\} \tag{3-40}$$

In general, the decision region $\mathbb{D}_m$ is given as

$$\mathbb{D}_m = \left\{\boldsymbol{r}\in\mathbb{R}^N | \boldsymbol{r}\cdot\boldsymbol{s}_m^T + \eta_m > \boldsymbol{r}\cdot\boldsymbol{s}_{m'}^T + \eta_{m'},\ \forall\ 1\le m'\le M, m'\ne m\right\} \tag{3-41}$$

Note that each decision region is described in terms of at most $M-1$ inequalities. In some cases some of these inequalities are dominated by the others and are redundant. Also note that each boundary is of the general form of

$$\boldsymbol{r}\cdot(\boldsymbol{s}_m - \boldsymbol{s}_{m'})^T > \eta_{m'} - \eta_m \tag{3-42}$$

which is the equation of a hyperplane. Therefore the boundaries of the decision regions in general are hyperplanes.

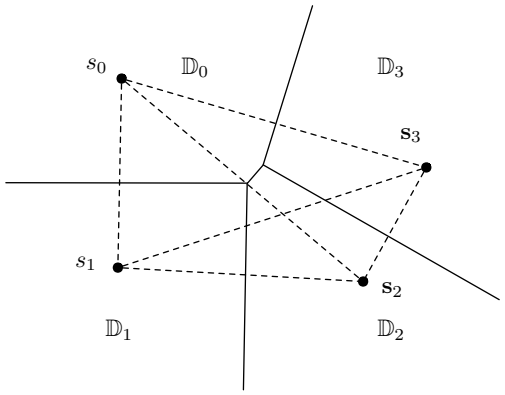

Figure 3-5: The decision regions for equiprobable signaling.

Optimal Detection for Binary Antipodal Signaling

In a binary antipodal signaling scheme $s_1(t) = s(t)$ and $s_2(t) = -s(t)$. The probabilities of messages 1 and 2 are p and $1-p$, respectively. This is obviously a case with $N = 1$, and the vector representations of the two signals are just scalars with $s_1 = \sqrt{\mathcal{E}_s}$ and $s_2 = -\sqrt{\mathcal{E}_s}$, where $\mathcal{E}_s$ is energy in each signal and is equal to $\mathcal{E}_b$. Following Equation 3-41, the decision region $\mathbb{D}_1$ is given as

$$\begin{aligned}
\mathbb{D}_1 &= \left\{ r \,\middle|\, r\sqrt{\mathcal{E}_b} + \frac{N_0}{2}\ln p - \frac{1}{2}\mathcal{E}_b > -r\sqrt{\mathcal{E}_b} + \frac{N_0}{2}\ln(1-p) - \frac{1}{2}\mathcal{E}_b \right\} \\
&= \left\{ r \,\middle|\, r > \frac{N_0}{4\sqrt{\mathcal{E}_b}} \ln\frac{1-p}{p} \right\} \\
&= \{ r \,|\, r > r_{\text{th}} \}
\end{aligned} \tag{3-43}$$

where the threshold r_{th} is defined as

$$r_{\text{th}} = \frac{N_0}{4\sqrt{\mathcal{E}_b}} \ln\frac{1-p}{p} \tag{3-44}$$

The constellation and teh decision regions are shown in Figure 3-6.

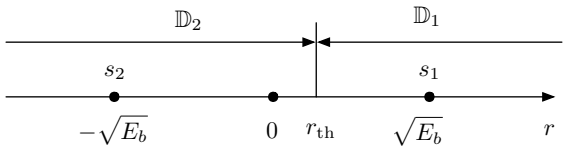

Figure 3-6: The decision regions for antipodal signaling.

Note that as $p \to 0$, we have $r_{\text{th}} \to \infty$ and the entire real line becomes $\mathbb{D}_2$; and when $p \to 1$, the entire line becomes $\mathbb{D}_1$, as expected. Also note that when $p = \frac{1}{2}$, *i.e.*, when the messages are equiprobable, $r_{\text{th}} = 0$ and the decvision rule reduces to a minimum-distance rule. To derive the

error probability for this system, we use Equation 3-15. This yields

$$\begin{aligned} P_e &= \sum_{m=1}^{2} P_m \sum_{1\le m'\le 2 | m'\neq m} \int_{\mathbb{D}_{m'}} p(\boldsymbol{r}|\boldsymbol{s}_m) d\boldsymbol{r} \\ &= p\int_{\mathbb{D}_2} p\left(r \left| s=\sqrt{\mathcal{E}_b}\right.\right) dr + (1-p)\int_{\mathbb{D}_1} p\left(r\left|s=-\sqrt{\mathcal{E}_b}\right.\right) dr \\ &= p\int_{-\infty}^{r_{\text{th}}} p\left(r \left| s=\sqrt{\mathcal{E}_b}\right.\right) dr + (1-p)\int_{r_{\text{th}}}^{\infty} p\left(r\left|s=-\sqrt{\mathcal{E}_b}\right.\right) dr \\ &= p\Pr\left\{\mathcal{N}\left(\sqrt{\mathcal{E}_b}, \frac{N_0}{2}\right) < r_{\text{th}}\right\} + (1-p)\Pr\left\{\mathcal{N}\left(-\sqrt{\mathcal{E}_b}, \frac{N_0}{2}\right) > r_{\text{th}}\right\} \\ &= pQ\left(\frac{\sqrt{\mathcal{E}_b}-r_{\text{th}}}{\sqrt{\frac{N_0}{2}}}\right) + (1-p)Q\left(\frac{\sqrt{\mathcal{E}_b}+r_{\text{th}}}{\sqrt{\frac{N_0}{2}}}\right) \end{aligned} \tag{3-45}$$

where in the last step we have used Equation 2-38. In the special case where $p = \frac{1}{2}$, we have $r_{\text{th}} = 0$ and the error probability simplifies to

$$P_e = Q\left(\sqrt{\frac{2\mathcal{E}_b}{N_0}}\right) \tag{3-46}$$

Also note that since the system is binary, the error probability for each message is equal to the bit error probability, *i.e.*, $P_b = P_e$.

Error Probability for Equiprobable Binary Signaling Scheme

In this case the transmitter transmits one or the two equiprobable signals $s_1(t)$ and $s_2(t)$ over the AWGN channel. Since the signals are equiprobable, the two decision regions are separated by the perpendicular bisector of the line connecting $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$. By symmetry, error probabilities when $\boldsymbol{s}_1$ or $\boldsymbol{s}_2$ is transmitted are equal, therefore $P_b = \Pr\{\text{error}|\boldsymbol{s}_1 \text{ sent}\}$. The decision regions and the perpendicular bisector of the line connecting $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$ are shown in Figure 3-7.

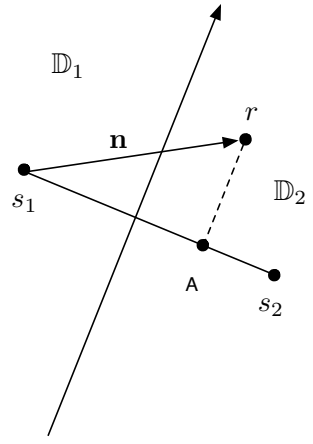

Figure 3-7: Decision regions for binary equiprobable signals.

Since we are assuming that $\boldsymbol{s}_1$ is sent, an error occurs if $\boldsymbol{r}$ is in $\mathbb{D}_2$, which means the distance between the projection of $\boldsymbol{r} - \boldsymbol{s}_1$ on $\boldsymbol{s}_2 - \boldsymbol{s}_1$, *i.e.*, point A, from $\boldsymbol{s}_1$ is larger than $\frac{d_{12}}{2}$, where

$d_{12} = \|\boldsymbol{s}_2 - \boldsymbol{s}_1\|$. Note that since $\boldsymbol{s}_1$ is sent, $\boldsymbol{n} = \boldsymbol{r} - \boldsymbol{s}_1$, and the projection of $\boldsymbol{r} - \boldsymbol{s}_1$ on $\boldsymbol{s}_2 - \boldsymbol{s}_1$ becomes equal to $\frac{\boldsymbol{n}\cdot(\boldsymbol{s}_2-\boldsymbol{s}_1)^T}{d_{12}}$. Therefore, the error probability is given by

$$P_b = \Pr\left\{\frac{\boldsymbol{n}\cdot(\boldsymbol{s}_2-\boldsymbol{s}_1)^T}{d_{12}} > \frac{d_{12}}{2}\right\} \tag{3-47}$$

or

$$P_b = \Pr\left\{\boldsymbol{n}\cdot(\boldsymbol{s}_2-\boldsymbol{s}_1)^T > \frac{d_{12}^2}{2}\right\} \tag{3-48}$$

We note that $\boldsymbol{n}\cdot(\boldsymbol{s}_2-\boldsymbol{s}_1)^T$ is a zero-mean Gaussaan random variable with rariance $\frac{d_{12}^2 N_0}{2}$; therefore, using Equation 2-38, we obtain

$$\begin{aligned} P_b &= Q\left(\frac{\frac{d_{12}^2}{2}}{d_{12}\sqrt{\frac{N_0}{2}}}\right) \\ &= Q\left(\sqrt{\frac{d_{12}^2}{2N_0}}\right) \end{aligned} \tag{3-49}$$

Equation 3-49 is very general and applies to all binary equiprobable signaling systems regardless of the shape of the signals. Since $Q(\cdot)$ is a decreasing function, in order to minimize the error probability, the distance between signal points has to be maximized. The distance d_{12} is obtained from

$$d_{12}^2 = \int_{-\infty}^{\infty}(s_1(t) - s_2(t))^2 dt \tag{3-50}$$

In the special case that the binary signals are equiprobable and have equal energy, *i.e.*, when $\mathcal{E}_{s_1} = \mathcal{E}_{s_2} = \mathcal{E}$, we can expand Equation 3-50 and get

$$d_{12}^2 = \mathcal{E}_{s_1} + \mathcal{E}_{s_2} - 2\langle s_1(t), s_2(t)\rangle = 2\mathcal{E}(1-\rho) \tag{3-51}$$

where ρ is the cross-correlation coefficient between $s_1(t)$ and $s_2(t)$. Since $-1 \le \rho \le 1$, we observe from Equation 3-51 that the binary signals are maximally separated when $\rho = -1$, *i.e.*, when the signals are antipodal. In this case the error prohahility of the system is minimized.

Optimal Detection for Binary Orthogonal Signaling

For binary orthogonal signals we have

$$\int_{-\infty}^{\infty} s_i(t)s_j(t)dt = \begin{cases} \mathcal{E}; & i = j \\ 0; & i \ne j \end{cases} \quad 1 \le i, j \le 2 \tag{3-52}$$

Note that since the system is binary, $\mathcal{E}_b = E$. Here we choose $\phi_j(t) = \frac{s_j(t)}{\sqrt{\mathcal{E}_b}}$ for $j = 1, 2$, and the vector representations of the signal set become

$$\begin{aligned} \boldsymbol{s}_1 &= (\sqrt{\mathcal{E}_b}, 0) \\ \boldsymbol{s}_2 &= (0, \sqrt{\mathcal{E}_b}) \end{aligned} \tag{3-53}$$

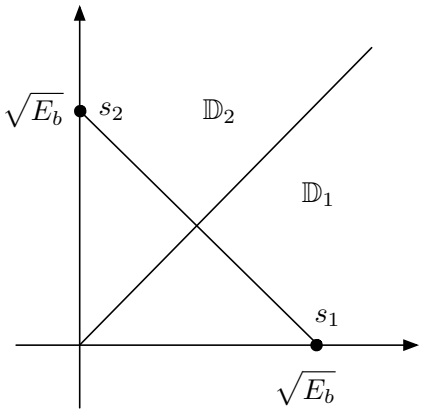

Figure 3-8: Signal constellation and decision regions for equiprobable binary orthogonal signaling.

The constellation and the optimal decision regions for the case of equiprohahlc signals are shown in Figure 3-8.

For this signaling scheme it is clear that $d = \sqrt{2\mathcal{E}_b}$ and

$$P_b = Q\left(\sqrt{\frac{d^2}{2N_0}}\right) = Q\left(\sqrt{\frac{\mathcal{E}_b}{N_0}}\right) \tag{3-54}$$

Comparing this result with the error probability of binary antipodal signaling given in Equation 3-46, we see that a binary orthogonal signaling requires twice the energy per bit of a binary antipodal signaling system to provide the same error probability. Therefore in terms of power efficiency, binary orthogonal signaling underperforms binary antipodal signaling by a factor of 2, or equivalently by 3 dB.

The term

$$\gamma_b = \frac{\mathcal{E}_b}{N_0} \tag{3-55}$$

which appears in the expression for error probability of many signaling systems is called the *signal-to-noise ratio per bit*, or *SNR per bit*, or simply the *SNR* of the communication system. Plots of error probability as a function of SNR/bit for binary antipodal and binary orthogonal signaling are shown in Figure 3-9. It is clear from this figure that the plot for orthogonal signaling is the result of a 3-dB shift of the plot for antipodal signaling.

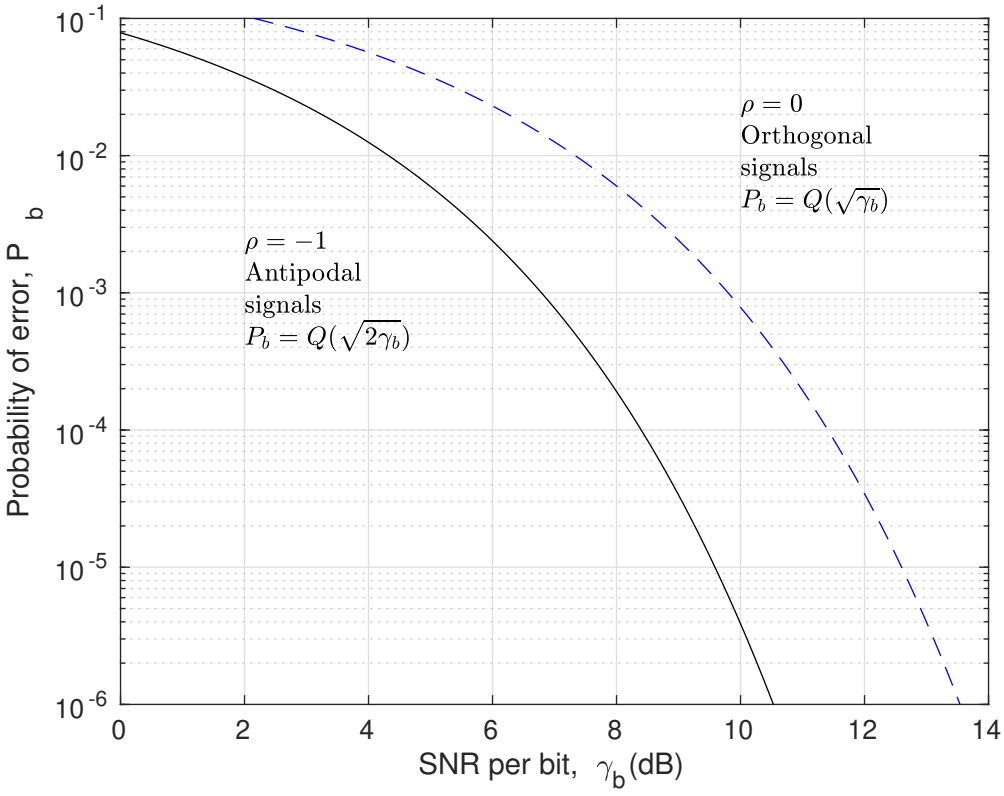

Figure 3-9: Error probability for binary antipodal and binary orthogonal signaling.

Chapter 4

Digital Communication Through Band-Limited Channels

In this chapter, we consider the problem of signal design when the channel is band-limited to some specified bandwidth of W Hz. Under this condition, the channel may be modeled as a linear filter having an equivalent lowpass[1] frequency response $C(f)$ that is zero for $|f| > W$.

The first topic that is treated is the design of the signal pulse $g(t)$ in a linearly modulated signal, represented as

$$v(t) = \sum_n I_n g(t - nT)$$

that efficiently utilizes the total available channel bandwidth W. We shall see that when the channel is ideal for $|f| \leq W$, a signal pulse can be designed that allows us to transmit at symbol rates comparable to or exceeding the channel bandwidth W. On the other hand, when the channel is not ideal, signal transmission at a symbol rate equal to or exceeding W results in intersymbol interference (ISI) among a number of adjacent symbols.

The second topic that we consider is the design of the receiver in the presence of intersymbol interference and AWGN. The solution to the ISI problem is to design a receiver that employs a means for compensating or reducing the ISI in the received signal. The compensator for the ISI is called an equalizer.

4.1 Characterization of Band-limited Channels

For our purposes, a band-limited channel such as a telephone channel will be characterized as a linear filter having an equivalent lowpass frequency-response characteristic $C(f)$. Its equivalent lowpass impulse response is denoted by $c(t)$. Then, if a signal of the form

$$s(t) = \Re\left[v(t)e^{j2\pi f_c t}\right] \tag{4-1}$$

is transmitted over a bandpass telephone channel, the equivalent low-pass received signal is

$$r(t) = \int_{-\infty}^{\infty} v(\tau)c(t - \tau)d\tau + z(t) \tag{4-2}$$

where the integral represents the convolution of $c(t)$ with $v(t)$, and $z(t)$ denotes the additive noise. Alternatively, the signal term can be represented in the frequency domain as $V(f)C(f)$, where $V(f)$ is the Fourier transform of $v(t)$.

[1]For convenience, the subscript on lowpass equivalent signals is omitted throughout this chapter.

If the channel is band-limited to W Hz, then $C(f) = 0$ for $|f| > W$. As a consequence, any frequency components in $V(f)$ above $|f| = W$ will not be passed by the channel. For this reason, we limit the bandwidth of the transmitted signal to W Hz also.

Within the bandwidth of the channel, we may express the frequency response $C(f)$ as

$$C(f) = |C(f)|e^{j\theta(f)} \tag{4-3}$$

where $|C(f)|$ is the amplitude-response characteristic and $\theta(f)$ is the phase-response characteristic. Furthermore, the envelope delay characteristic is defined as

$$\tau(f) = -\frac{1}{2\pi}\frac{d\theta(f)}{df} \tag{4-4}$$

A channel is said to be *nondistorting* or *ideal* if the amplitude response $|C(f)|$ is constant for all $|f| \leq W$ and $\theta(f)$ is a linear function of frequency, *i.e.*, $\tau(f)$ is a constant for all $|f| \leq W$. On the other hand, if $|C(f)|$ is not constant for all $|f| \leq W$, we say that the channel *distorts the transmitted signal* $V(f)$ *in amplitude*, and, if $\tau(f)$ is not constant for all $|f| \leq W$, we say that the channel *distorts the signal* $V(f)$ *in delay.*

As a result of the amplitude and delay distortion caused by the nonideal channel frequency-response characteristic $C(f)$, a succession of pulses transmitted through the channel at rates comparable to the bandwidth W are smeared to the point that they are no longer distinguishable as well-defined pulses at the receiving terminal. Instead, they overlap, and, thus, we have intersymbol interference. As an example of the effect of delay distortion on a transmitted pulse,

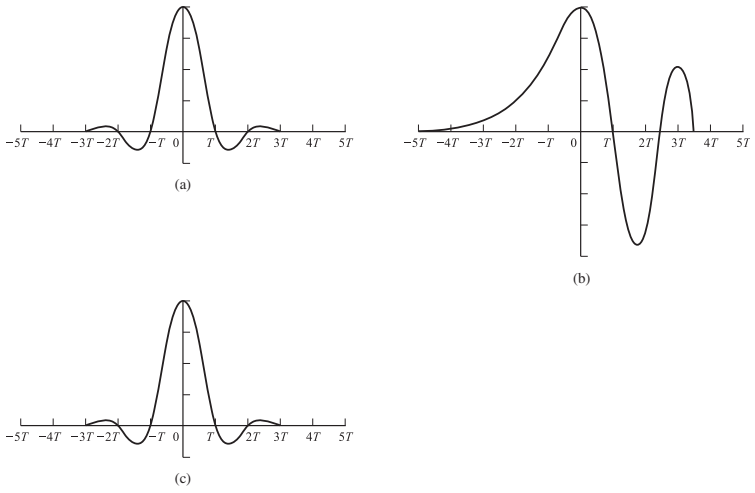

Figure 4-1: Effect of channel distortion: (a) channel input; (b) channel output; (c) equalizer output.

Figure 4-1a illustrates a band-limited pulse having zeros periodically spaced in time at points labeled $\pm T, \pm 2T$, etc. If information is conveyed by the pulse amplitude, as in PAM, for example, then one can transmit a sequence of pulses, each of which has a peak at the periodic zeros of the other pulses. However, transmission of the pulse through a channel modeled as having a linear envelope delay characteristic $\tau(f)$ (quadratic phase $\theta(f)$) results in the received pulse shown in Figure 4-1b having zero-crossings that are no longer periodically spaced. Consequently, a sequence of successive pulses would be smeared into one another and the peaks of the pulses

would no longer be distinguishable. Thus, the channel delay distortion results in intesymbol interference. As will be discussed in this chapter, it is possible to compensate for the nonideal frequency-response characteristic of the channel by use of a filter or equalizer at the demodulator. Figure 4-1c illustrates the output of a linear equalizer that compensates for the linear distortion in the channel.

The extent of the intersymbol interference on a telephone channel can be appreciated by observing a frequency-response characteristic of the channel. Figure 4-2 illustrates the measured

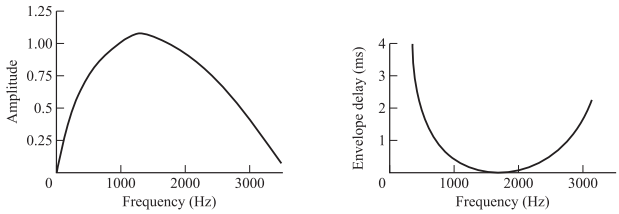

Figure 4-2: Average amplitude and delay characteristics of medium-range telephone channel.

average amplitude and delay as functions of frequency for a medium-range (180-725 mi) telephone channel of the switched telecommunications network as given by Duffy and Tratcher (1971). We observe that the usable band of the channel extends from about 300 Hz to about 3000 Hz. The

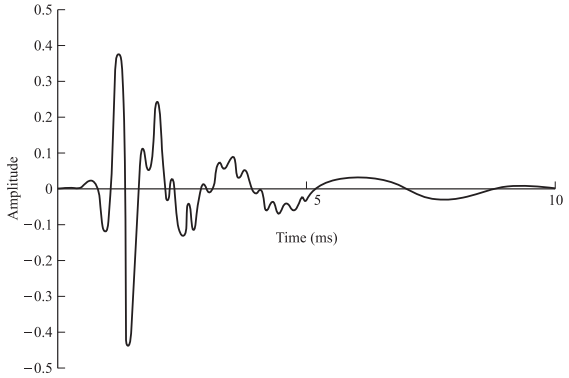

Figure 4-3: Average amplitude and delay characteristics of medium-range telephone channel.

corresponding impulse response of this average channel is shown in Figure 4-3. Its duration is about 10 ms. In comparison, the transmitted symbol rates on such a channel may be of the order of 2500 pulses or symbols per second. Hence, intersymbol interference might extend over 20-30 symbols.

4.2 Signal Design for Band-limited Channels

The equivalent lowpass transmitted signal for several different types of digital modulation techniques has the common form

$$v(t) = \sum_{n=0}^{\infty} I_n g(t - nT) \tag{4-5}$$

where I_n represents the discrete information-bearing sequence of symbols and $g(t)$ is a pulse that, for the purposes of this discussion, is assumed to have a band-limited frequency-response characteristic $G(f)$, *i.e.*, $G(f) = 0$ for $|f| > W$. This signal is transmitted over a channel having a frequency response $C(f)$, also limited to $|f| \leq W$. Consequently, the received signal can be represented as

$$r(t) = \sum_{n=0}^{\infty} I_n h(t - nT) + z(t) \tag{4-6}$$

where

$$h(t) = \int_{-\infty}^{\infty} g(\tau) c(t - \tau) d\tau \tag{4-7}$$

and $z(t)$ represents the additive white Gaussian noise.

Let us suppose that the received signal is passed first through a filter and then sampled at a rate $1/T$ samples/s. We shall show in a subsequent section that the optimum filter from the point of view of signal detection is one matched to the received pulse. That is, the frequency response of the receiving filter is $H^*(f)$. We denote the output of the receiving filter as

$$y(t) = \sum_{n=0}^{\infty} I_n x(t - nT) + \nu(t) \tag{4-8}$$

where $x(t)$ is the pulse representing the response of the receiving filter to the input pulse $h(t)$ and $\nu(t)$ is the response of the receiving filter to the noise $z(t)$.

Now, if $y(t)$ is sampled at times $t = kT + \tau_0$, $k = 0, 1, \cdots$, we have

$$y(kT + \tau_0) \equiv y_k = \sum_{n=0}^{\infty} I_n x(kT - nT + \tau_0) + \nu(kT + \tau_0) \tag{4-9}$$

or, equivalently,

$$y_k = \sum_{n=0}^{\infty} I_n x_{k-n} + \nu_k, \quad k = 0, 1, \cdots \tag{4-10}$$

where τ_0 is the transmission delay through the channel. The sample values can be expressed as

$$y_k = x_0 \left(I_k + \frac{1}{x_0} \sum_{n=0, n\neq k}^{\infty} I_n x_{k-n} \right) + \nu_k, \; k = 0, 1, \cdots \tag{4-11}$$

We regard x_0 as an arbitrary scale factor, which we arbitrarily set equal to unity for convenience. Then

$$y_k = I_k + \sum_{n=0, n\neq k}^{\infty} I_n x_{k-n} + \nu_k \tag{4-12}$$

The term I_k represents the desired information symbol at the kth sampling instant, the term

$$\sum_{n=0,n\neq k}^{\infty} I_n x_{k-n}$$

represents the ISI, and $\nu(t)$ is the additive Gaussian noise variable at the kth sampling instant.

Below, we consider the problem of signal design under the condition that there is no intersymbol interference at the sampling instants.

4.2.1 Design of Band-Limited Signals for No Intersymbol Interference-The Nyquist Criterion

For the discussion in this section and in Section 4.2.2, we assume that the band-limited channel has ideal frequency-response characteristics, *i.e.*, $C(f) = 1$ for $|f| \leq W$. Then the pulse $x(t)$ has a spectral characteristic $X(f) = |G(f)|^2$, where

$$x(t) = \int_{-W}^{W} X(f) e^{j2\pi ft} df \tag{4-13}$$

We are interested in determining the spectral properties of the pulse $x(t)$ and, hence, the transmitted pulse $g(t)$, that results in no intersymbol interference. From 4-12, the condition for no intersymbol interference is

$$x(t = kT) \equiv x_k = \begin{cases} 1 & k = 0 \\ 0 & k \neq 0 \end{cases} \tag{4-14}$$

Below, we derive the necessary and sufficient condition on $X(f)$ in order for $x(t)$ to satisfy the above relation. This condition is known as the *Nyquist pulse-shaping criterion* or *Nyquist condition for zero ISI* and is stated in the following theorem.

Theorem 2. The necessary and sufficient condition for $x(t)$ to satisfy

$$x_{nT} = \begin{cases} 1 & n = 0 \\ 0 & n \neq 0 \end{cases} \tag{4-15}$$

is that its Fourier transform $X(f)$ satisfy

$$\sum_{m=-\infty}^{\infty} X(f + m/T) = T \tag{4-16}$$

Proof. In general, $x(t)$ is the inverse Fourier transform of $X(f)$. Hence,

$$x(t) = \int_{-\infty}^{\infty} X(f) e^{j2\pi ft} df \tag{4-17}$$

At the sampling instants $t = nT$, this relation becomes

$$x(nT) = \int_{-\infty}^{\infty} X(f) e^{j2\pi fnT} df \tag{4-18}$$

Let us break up the integral in Equation (4-18) into integrals covering the finite range of $1/T$. Thus, we obtain

$$\begin{aligned} x(nT) &= \sum_{m=-\infty}^{\infty} \int_{(2m-1)/2T}^{(2m+1)/2T} X(f) e^{j2\pi fnT} df \\ &= \sum_{m=-\infty}^{\infty} \int_{-1/2T}^{1/2T} X(f+m/T) e^{j2\pi fnT} df \\ &= \int_{-1/2T}^{1/2T} \left[\sum_{m=-\infty}^{\infty} X(f+m/T) \right] e^{j2\pi fnT} df \\ &= \int_{-1/2T}^{1/2T} B(f) e^{j2\pi fnT} df \end{aligned} \tag{4-19}$$

where we have defined $B(f)$ as

$$B(f) = \sum_{m=-\infty}^{\infty} X(f+m/T) \tag{4-20}$$

Obviously $B(f)$ is a periodic function with period $1/T$, and, therefore, it can be expanded in terms of its Fourier series coefficients $\{b_n\}$ as

$$B(f) = \sum_{n=-\infty}^{\infty} b_n e^{j2\pi nfT} \tag{4-21}$$

where

$$b_n = T \int_{-1/2T}^{1/2T} B(f) e^{-j2\pi nfT} df \tag{4-22}$$

Comparing Equations (4-22) and (4-19), we obtain

$$b_n = Tx(-nT) \tag{4-23}$$

Therefore, the necessary and sufficient condition for (4-14) to be satisfied is that

$$b_n = \begin{cases} T & n = 0 \\ 0 & n \neq 0 \end{cases} \tag{4-24}$$

which, when substituted into (4-21), yields

$$B(f) = T \tag{4-25}$$

or, equivalently,

$$\sum_{m=-\infty}^{\infty} X(f+m/T) = T \tag{4-26}$$

This concludes the proof of the theorem. □

Now suppose that the channel has a bandwidth of W. Then $C(f) = 0$ for $|f| > W$ and, consequently, $X(f) = 0$ for $|f| > W$. We distinguish three cases.

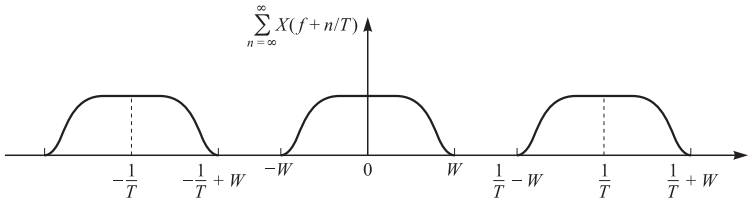

Figure 4-4: Plot of $B(f)$ for the case $T < 1/2W$.

1. When $T < 1/2W$, or, equivalently, $1/T > 2W$, since $B(f) = \sum_{m=-\infty}^{\infty} X(f + n/T)$ consists of non overlapping replicas of $X(f)$, separated by $1/T$ as shown in Figure 4-4, there is no choice for $X(f)$ to ensure $B(f) \equiv T$ in this case and there is no way that we can design a system with no ISI.

2. When $T = 1/2W$, or, equivalently, $1/T = 2W$ (the Nyquist rate), the replications of $X(f)$, separated by $1/T$, are as shown in Figure 4-5. It is clear that in this case there exists only one $X(f)$ that results in $B(f) = T$, namely,

$$X(f) = \begin{cases} T & |f| < W \\ 0 & \text{otherwise} \end{cases} \tag{4-27}$$

which corresponds to the pulse

$$x(t) = \frac{\sin(\pi t/T)}{\pi t/T} \equiv \text{sinc}\left(\frac{\pi t}{T}\right) \tag{4-28}$$

This means that the smallest value of T for which transmission with zero ISI is possible is $T = 1/2W$, and for this value, $x(t)$ has to be a sinc function. The difficulty with this choice of $x(t)$ is that it is noncausal and, therefore, nonrealizable. To make it realizable, usually a delayed version of it, *i.e.*, $\text{sinc}[\pi(t - t_0)/T]$ is used and t_0 is chosen such that for $t < 0$, we have $\text{sinc}[\pi(t - t_0)/T] \approx 0$. Of course, with this choice of $x(t)$, the sampling time must also be shifted to $mT + t_0$. A second difficulty with this pulse shape is that its rate of convergence to zero is slow. The tails of $x(t)$ decay as $1/t$; consequently, a small mistiming error in sampling the output of the matched filter at the demodulator results in an infinite series of ISI components. Such a series is not absolutely summable because of the $1/t$ rate of decay of the pulse, and, hence, the sum of the resulting ISI does not converge.

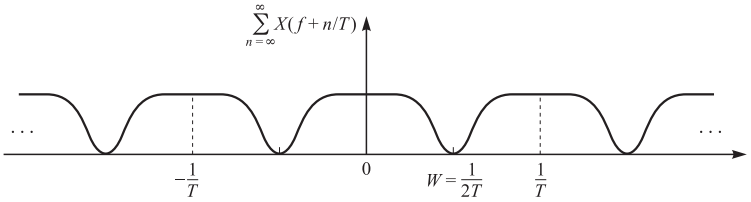

Figure 4-5: Plot of $B(f)$ for the case $T = 1/2W$.

3. When $T > 1/2W$, $B(f)$ consists of overlapping replications of $X(f)$ separated by $1/T$, as shown in Figure 4-6. In this case, there exist numerous choices for $X(f)$ such that $B(f) \equiv T$.

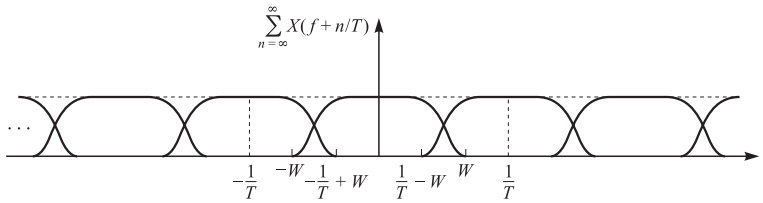

Figure 4-6: Plot of $B(f)$ for the case $T > 1/2W$.

Raised cosine spectrum

A particular pulse spectrum, for the $T > 1/2W$ case, that has desirable spectral properties and has been widely used in practice is the raised cosine spectrum. The raised cosine frequency characteristic is given as

$$X_{rc}(f) = \begin{cases} T & 0 \le |f| \le \dfrac{1-\beta}{2T} \\ \dfrac{T}{2}\left\{1+\cos\left[\dfrac{\pi T}{\beta}\left(|f|-\dfrac{1-\beta}{2T}\right)\right]\right\} & \dfrac{1-\beta}{2T} \le |f| \le \dfrac{1+\beta}{2T} \\ 0 & |f| > \dfrac{1+\beta}{2T} \end{cases} \tag{4-29}$$

where β is called the *roll-off factor* and takes values in the range $0 \le f \le 1$. The bandwidth occupied by the signal beyond the Nyquist frequency $T/2$ is called the *excess bandwidth* and is usually expressed as a percentage of the Nyquist frequency. For example, when $\beta = 1/2$, the excess bandwidth is 50 percent and when $\beta = 1$, the excess bandwidth is 100 percent. The pulse $x(t)$, having the raised cosine spectrum, is

$$\begin{aligned} x(t) &= \frac{\sin(\pi t/T)}{\pi t/T}\frac{\cos(\pi\beta t/T)}{1-4\beta^2t^2/T^2} \\ &= \mathrm{sinc}(\pi t/T)\frac{\cos(\pi\beta t/T)}{1-4\beta^2t^2/T^2} \end{aligned} \tag{4-30}$$

Note that $x(t)$ is normalized so that $x(0) = 1$. Figure 4-7 illustrates the raised cosine spectral characteristics and the corresponding pulses for $\beta = 0, 1/2$, and 1. Note that for $\beta = 0$, the pulse reduces to $x(t) = \mathrm{sinc}(\pi t/T)$, and the symbol rate $1/T = 2W$. When $\beta = 1$, the symbol rate is $1/T = W$. In general, the tails of $x(t)$ decay as $1/t^3$ for $\beta > 0$. Consequently, a mistiming error in sampling leads to a series of ISI components that converges to a finite value.

Because of the smooth characteristics of the raised cosine spectrum, it is possible to design practical filters for the transmitter and the receiver that approximate the overall desired frequency response. In the special case where the channel is ideal, *i.e.*, $C(f) = 1$, $|f| \le W$, we have

$$X_{rc}(f) = G_T(f)G_R(f) \tag{4-31}$$

where $G_T(f)$ and $G_R(f)$ are the frequency responses of the two filters. In this case, if the receiver filter is matched to the transmitter filter, we have $X_{rc}(f) = G_T(f)G_R(f) = |G_T(f)|^2$. Ideally,

$$G_T(f) = \sqrt{|X_{rc}(f)|}e^{-j2\pi f t_0} \tag{4-32}$$

and $G_R(f) = G_T^*(f)$, where t_0 is some nominal delay that is required to ensure physical realizability of the filter. Thus, the overall raised cosine spectral characteristic is split evenly between the transmitting filter and the receiving filter. Note also that an additional delay is necessary to ensure the physical realizability of the receiving filter.

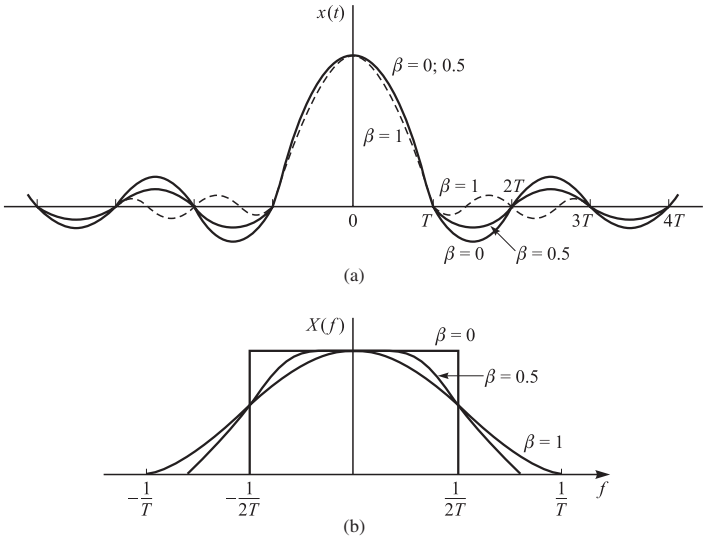

Figure 4-7: Pulses having a raised cosine spectrum.

4.2.2 Design of Band-Limited Signals with Controlled ISI: Partial-Response Signals

As we have observed from our discussion of signal design for zero ISI, it is necessary to reduce the symbol rate $1/T$ below the Nyquist rate of $2W$ symbols/s to realize practical transmitting and receiving filters. On the other hand, suppose we choose to relax the condition of zero ISI and, thus, achieve a symbol transmission rate of $2W$ symbols/s. By allowing for a controlled amount of ISI, we can achieve this symbol rate.

We have already seen that the condition for zero ISI is $x(nT) = 0$ for $n \neq 0$. However, suppose that we design the band-limited signal to have controlled ISI at one time instant. This means that we allow one additional nonzero value in the samples $\{x(nT)\}$. The ISI that we introduce is deterministic or "controlled" and, hence, it can be taken into account at the receiver, as discussed below.

One special case that leads to (approximately) physically realizable transmitting and receiving filters is specified by the samples[2]

$$x(nT) = \begin{cases} 1 & n = 0, 1 \\ 0 & \text{otherwise} \end{cases} \tag{4-33}$$

Now, using Equation (4-23), we obtain

$$b_n = \begin{cases} T & n = 0, -1 \\ 0 & \text{otherwise} \end{cases} \tag{4-34}$$

which, when substituted into Equation (4-21), yields

$$B(f) = T + Te^{-j2\pi fT} \tag{4-35}$$

[2]It is convenient to deal with samples of $x(t)$ that are normalized to unity for $n = 0, 1$.

As in the preceding section, it is impossible to satisfy the above equation for $T < 1/2W$. However, for $T = 1/2W$, we obtain

$$\begin{aligned} X(f) &= \begin{cases} \frac{1}{2W}(1 + e^{-j\pi f/W}) & |f| < W \\ 0 & \text{otherwise} \end{cases} \\ &= \begin{cases} \frac{1}{W} e^{-j\pi f/W} \cos \frac{\pi f}{2W} & |f| < W \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{4-36}$$

Therefore, $x(t)$ is given by

$$x(t) = \text{sinc}(2\pi W t) + \text{sinc}\left[2\pi\left(Wt - \frac{1}{2}\right)\right] \tag{4-37}$$

This pulse is called a duobinary signal pulse. It is illustrated along with its magnitude spectrum in Figure 4-8. Note that the spectrum decays to zero smoothly, which means that physically realizable filters can be designed that approximate this spectrum very closely. Thus, a symbol rate of $2W$ is achieved.

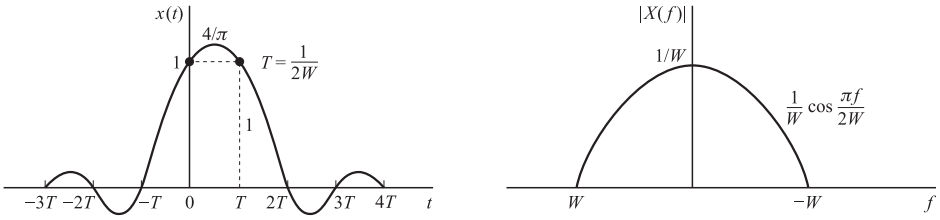

Figure 4-8: Time-domain and frequency-domain characteristics of a duobinary signal

Another special case that leads to (approximately) physically realizable transmitting and receiving filters is specified by the samples

$$X\left(\frac{n}{2W}\right) = x(nT) = \begin{cases} 1 & n = -1 \\ -1 & n = 1 \\ 0 & \text{otherwise} \end{cases} \tag{4-38}$$

The corresponding pulse $x(t)$ is given as

$$x(t) = \text{sinc}\,\frac{\pi(t+T)}{T} - \text{sinc}\,\frac{\pi(t-T)}{T} \tag{4-39}$$

and its spectrum is

$$X(f) = \begin{cases} \frac{1}{2W}(e^{j\pi f/W} - e^{-j\pi f/W}) = \frac{1}{W}\sin\frac{\pi f}{W} & |f| \le W \\ 0 & \text{otherwise} \end{cases} \tag{4-40}$$

This pulse and its magnitude spectrum are illustrated in Figure 4-9. It is called a modified duobinary signal pulse. It is interesting to note that the spectrum of this signal has a zero at $f = 0$, making it suitable for transmission over a channel that does not pass DC.

One can obtain other interesting and physically realizable filter characteristics, as shown by Kretzmer (1966) and Lucky et al. (1968), by selecting different values for the samples $\{x(n/2W)\}$

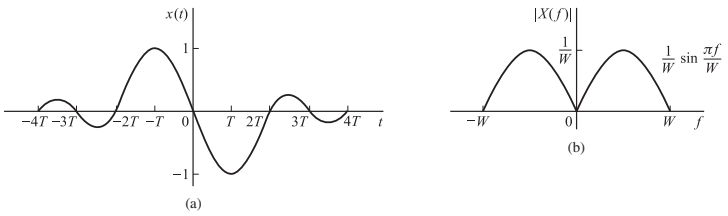

Figure 4-9: Time-domain and frequency-domain characteristics of a modified duobinary signal.

and more than two nonzero samples. However, as we select more nonzero samples, the problem of unraveling the controlled ISI becomes more cumbersome and impractical.

In general, the class of band-limited signal pulses that have the form

$$x(t) = \sum_{n=-\infty}^{\infty} x\left(\frac{n}{2W}\right) \text{sinc}\left[2\pi W\left(t - \frac{n}{2W}\right)\right] \tag{4-41}$$

and their corresponding spectra

$$X(f) \quad = \quad \begin{cases} \frac{1}{2W}\sum_{n=-\infty}^{\infty} x\left(\frac{n}{2W}\right) - e^{-j\pi f/W} & |f| \le W \\ 0 & \text{otherwise} \end{cases} \tag{4-42}$$

are called partial-response signals when controlled ISI is a purposely introduced by selecting two or more nonzero samples from the set $\{x(n/2W)\}$. The resulting signal pulses allow us to transmit information symbols at the Nyquist rate of $2W$ symbols/s. The detection of the received symbols in the presence of controlled ISI is described below.

Alternative characterization of partial-response signals

We conclude this subsection by presenting another interpretation of a partial-response signal. Suppose that the partial-response signal is generated, as shown in Figure 4-10, by passing the discrete-time sequence $\{I_n\}$ through a discrete-time filter with coefficients $x_n \equiv x(n/2W), n = 0, 1, \cdots, N-1$, and using the output sequence $\{B_n\}$ from this filter to excite periodically with an input $B_n\delta(t-nT)$ an analog filter having an impulse response $\text{sinc}(2\pi Wt)$. The resulting output signal is identical to the partial-response signal given by Equation (4-41).

Since

$$B_n = \sum_{k=0}^{N-1} x_k I_{n-k} \tag{4-43}$$

the sequence of symbols $\{B_n\}$ is correlated as a consequence of the filtering performed on the sequence $\{I_n\}$. In fact, the autocorrelation function of the sequence $\{B_n\}$ is

$$\begin{aligned} R(m) &= E(B_n B_{n+m}) \\ &= \sum_{k=0}^{N-1}\sum_{l=0}^{N-1} x_k x_l E(I_{n-k} I_{n+m-l}) \end{aligned} \tag{4-44}$$

When the input sequence is zero-mean and white,

$$E(I_{n-k} I_{n+m-l}) = \delta_{m+k-l} \tag{4-45}$$

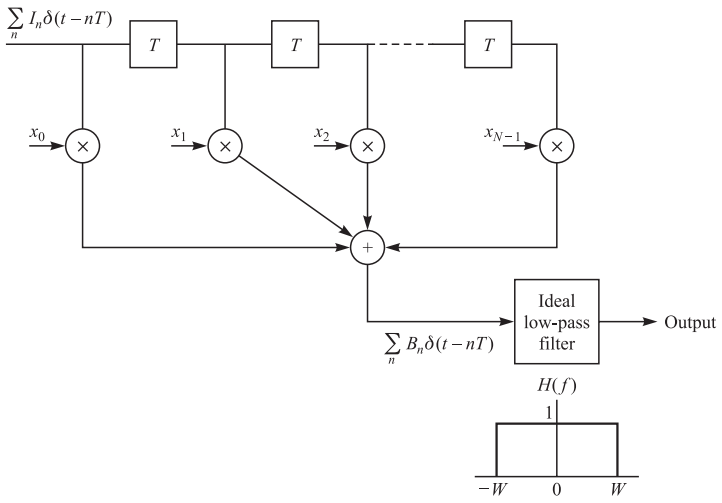

Figure 4-10: An alternative method for generating a partial-response signal.

where we have used the normalization $E(I_n^2) = 1$. Substitution of Equation 4-45, into Equation 4-44 yields the desired autocorrelation function for $\{B_n\}$ in the form

$$R(m) = \sum_{k=0}^{N-1-|m|} x_k x_{k+|m|}, \; m = 0, \pm 1, \cdots, \pm(N-1) \tag{4-46}$$

The corresponding power spectral density is

$$\begin{aligned} \mathcal{S}(f) &= \sum_{m=-(N-1)}^{N-1} R(m) e^{-j2\pi fmT} \\ &= \left| \sum_{m=0}^{N-1} x_m e^{-j2\pi fmT} \right|^2 \end{aligned} \tag{4-47}$$

where $T = 1/2W$ and $|f| \leq 1/2T = W$. Thus, the partial-response signal designs provide spectral shaping of the signal transmitted through the channel.

4.2.3 Data Detection for Controlled ISI

In this section, we describe two methods for detecting the information symbols at the receiver when the received signal contains controlled ISI. One is a symbol-by-symbol detection method that is relatively easy to implement. The second method is based on the maximum-likelihood criterion for detecting a sequence of symbols. The latter method minimizes the probability of error but is a little more complex to implement. In particular, we consider the detection of the duobinary and the modified duobinary partial-response signals. In both cases, we assume that the desired spectral characteristic $X(f)$ for the partial-response signal is split evenly between the transmitting and receiving filters, *i.e.*, $|G_T(f)| = |G_R(f)| = |X(f)|^{1/2}$. This treatment is based on PAM signals, but it is easily generalized to QAM and PSK.

Symbol-by-symbol suboptimum detection

For the duobinary signal pulse, $x(nT) = 1$, for $n = 0, 1$, and is zero otherwise. Hence, the samples at the output of the receiving filter (demodulator) have the form

$$y_m = B_m + v_m = I_m + I_{m-1} + v_m \tag{4-48}$$

where $\{I_m\}$ is the transmitted sequence of amplitudes and $\{v_m\}$ is a sequence of additive Gaussian noise samples. Let us ignore the noise for the moment and consider the binary case where $I_m = \pm 1$ with equal probability. Then B_m takes on one of three possible values, namely, $B_m = -2, 0, 2$ with corresponding probabilities 1/4, 1/2, 1/4. I f I_{m-1} is the detected symbol from the $(m-1)$th signaling interval, its effect on B_m, the received signal in the mth signaling interval, can be eliminated by subtraction, thus allowing I_m to be detected. This process can be repeated sequentially for every received symbol.

The major problem with this procedure is that errors arising from the additive noise tend to propagate. For example, if I_{m-1} is in error, its effect on B_m is not eliminated but, in fact, is reinforced by the incorrect subtraction. Consequently, the detection of I_m is also likely to be in error.

Error propagation can be avoided by precoding the data at the transmitter instead of eliminating the controlled ISI by subtraction at the receiver. The precoding is performed on the binary data sequence prior to modulation. From the data sequence $\{D_n\}$ of 1s and 0s that is to be transmitted, a new sequence $\{P_n\}$, called the precoded sequence, is generated. For the duobinary signal, the precoded sequence is defined as

$$P_m = D_m \ominus P_{m-1} \tag{4-49}$$

where $\ominus$ denotes modulo-2 subtraction[3]. Then we set $I_m = -1$ if $P_m = 0$ and $I_m = 1$ if $P_m = 1$, *i.e.*, $I_m = 2P_m - 1$.

The noise-free samples at the output of the receiving tilter are given by

$$\begin{aligned} B_m &= I_m + I_{m-1} \\ &= (2P_m - 1) + (2P_{m-1} - 1) \\ &= 2(P_m + P_{m-1} - 1) \end{aligned} \tag{4-50}$$

Consequently,

$$P_m + P_{m-1} = \frac{1}{2} B_m + 1 \tag{4-51}$$

Since $D_m = P_m \oplus P_{m-1}$, it follows that the data sequence D_m is obtained from B_m using the relation

$$D_m = \frac{1}{2} B_m + 1 \pmod 2 \tag{4-52}$$

Consequently, if $B_m = \pm 2$, then $D_m = 0$, and if $B_m = 0$, then $D_m = -1$. An example that illustrates the precoding and decoding operations is given in Table 4.1. In the presence of additive noise, the sampled outputs from the receiving filter are given by Equation 4-48. In this

[3] Although this is identical to modulo-2 addition. it is convenient to view the precoding operation for duobinary in terms of modulo-2 subtraction.

Table 4.1: Binary Signaling with Duobinary Pulses

Data sequence D_n		1	1	1	0	1	0	0	1	0	0	0	1	1	0	1
Precoded sequence P_n	0	1	0	1	1	0	0	0	1	1	1	1	0	1	1	0
Transmitted sequence I_n	-1	1	-1	1	1	-1	-1	-1	1	1	1	1	-1	1	1	-1
Received sequence B_n		0	0	0	2	0	-2	-2	0	2	2	2	0	0	2	0
Decoded sequence D_n		1	1	1	0	1	0	0	1	0	0	0	1	1	0	1

case $y_m = B_m + v_m$ is compared with the two thresholds set at $+1$ and -1. The data sequence $\{D_m\}$ is obtained according to the detection rule

$$D_m = \begin{cases} 1 & (|y_m| < 1) \\ 0 & (|y_m| \geq 1) \end{cases} \tag{4-53}$$

The extension from binary PAM to multilevel PAM signaling using the duobinary pulses is straightforward. In this case the M-level amplitude sequence $\{I_m\}$ results in a (noise-free) sequence

$$B_m = I_m + I_{m-1},\ m = 1, 2, \cdots \tag{4-54}$$

which has $2M - 1$ possible equally spaced levels. The amplitude levels are determined from the relation

$$I_m = 2P_m - (M - 1) \tag{4-55}$$

where $\{P_m\}$ is the precoded sequence that is obtained from an M-level data sequence $\{D_m\}$ according to the relation

$$P_m = [D_m \ominus P_{m-1}]_M \tag{4-56}$$

where the possible values of the sequence $\{D_m\}$ are $0, l, 2, \cdots, M - 1$.

In the absence of noise, the samples at the output of the receiving filter may be expressed as

$$B_m = I_m + I_{m-1} = 2\left[P_m + P_{m-1} - (M - 1)\right] \tag{4-57}$$

Hence,

$$P_m + P_{m-1} = \frac{1}{2}B_m + (M - 1) \tag{4-58}$$

Since $D_m = [P_m + P_{m-1}]_M$, it follows that

$$D_m = \left[\frac{1}{2}B_m + (M - 1)\right]_M \tag{4-59}$$

An example illustrating multilevel preceding and decoding is given in Table 4.2.

In the presence of noise, the received signal-plus-noise is quantized to the nearest of the possible signal levels and the rule given above is used on the quantized values to recover the data sequence.

Table 4.2: Four-Level Signal Transmission with Duobinary Pulses

Data sequence D_n		0	0	1	3	1	2	0	3	3	2	0	1	0
Precoded sequence P_n	0	0	0	1	2	3	3	1	2	1	1	3	2	2
Transmitted sequence I_n	-3	-3	-3	-1	1	3	3	-1	1	-1	-1	3	1	1
Received sequence B_n		-6	-6	-4	0	4	6	2	0	0	-2	2	4	2
Decoded sequence D_n		0	0	1	3	1	2	0	3	3	2	0	1	0

In the case of the modified duobinary pulse, the controlled ISI is specified by the values $x(n/2W) = -1$, for $n = 1$, $x(n/2W) = 1$ for $n = -1$, and zero otherwise. Consequently, the noise-free sampled output from the receiving filter is given as

$$B_m = I_m - I_{m-2} \tag{4-60}$$

where the M-level sequence $\{I_m\}$ is obtained by mapping a precoded sequence according to the Equation 4-55 and

$$P_m = [D_m \oplus P_{m-2}]_M \tag{4-61}$$

From these relations, it is easy to show that the detection rule for recovering the data sequence $\{D_m\}$ from $\{B_m\}$ in the absence of noise is

$$D_m = \left[\frac{1}{2}B_m\right]_M \tag{4-62}$$

As demonstrated above, the precoding of the data at the transmitter makes it possible to detect the received data on a symbol-by-symbol basis without having to look back at previously detected symbols. Thus, error propagation is avoided.

The symbol-by-symbol detection rule described above is not the optimum detection scheme for partial-response signals due to the memory inherent in the received signal. Nevertheless, symbol-by-symbol detection is relatively simple to implement and is used in many practical applications involving duobinary and modified duobinary pulse signals.

Let us determine the probability of error for detection of digital M-ary PAM signaling using duobinary and modified duobinary pulses. The channel is assumed to be an ideal band-limited channel with additive white Gaussian noise. The model for the communication system is shown in Figure 4-11.

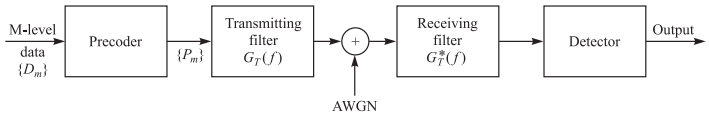

Figure 4-11: Block diagram of modulator and demodulator for partial-response signals.

At the transmitter, the M-level data sequence $\{D_m\}$ is precoded as described previously. The precoder output is mapped into one of M possible amplitude levels. Then the transmitting filter

with frequency response $G_T(f)$ has an output

$$v(t) = \sum_{n=-\infty}^{\infty} I_n g_T(t - nT) \tag{4-63}$$

The partial-response function $X(f)$ is divided equally between the transmitting and receiving filters. Hence, the receiving filter is matched to the transmitted pulse, and the cascade of the two filters results in the frequency characteristic

$$|G_T(f)G_R(f)| = |X(f)| \tag{4-64}$$

The matched filter output is sampled at $t = nT = n/2W$ and the samples are fed to the decoder. For the duobinary signal, the output of the matched filter at the sampling instant may be expressed as

$$y_m = I_m + I_{m-1} + v_m = B_m + v_m \tag{4-65}$$

where v_m is the additive noise component. Similarly, the output of the matched filter for the modified duobinary signal is

$$y_m = I_m - I_{m-2} + v_m = B_m + v_m \tag{4-66}$$

For binary transmission. let $I_m = \pm d$, where $2d$ is the distance between signal levels. Then, the corresponding values of B_m are $(2d, 0, -2d)$. For M-ary PAM signal transmission, where $I_m = \pm d, \pm 3d, \cdots, \pm(M-1)d$, the received signal levels are $B_m = 0, \pm 2d, \pm 4d, \cdots, \pm 2(M-l)d$. Hence, the number of received levels is $2M - 1$, and the scale factor d is equivalent to $x_0 = \mathcal{E}_g$.

The input transmitted symbols $\{I_m\}$ are assumed to be equally probable. Then, for duobinary and modified duobinary signals, it is easily demonstrated that, in the absence of noise, the received output levels have a (triangular) probability distribution of the form

$$P(B = 2md) = \frac{M - |m|}{M^2},\ m = 0, \pm 1, \pm 2, \cdots, \pm(M-1) \tag{4-67}$$

where B denotes the noise-free received level and $2d$ is the distance between any two adjacent received signal levels.

The channel corrupts the signal transmitted through it by the addition of white Gaussian noise with zero-mean and power spectral density $\frac{1}{2}N_0$.

We assume that a symbol error occurs whenever the magnitude of the additive noise exceeds the distance d. This assumption neglects the rare event that a large noise component with magnitude exceeding d may result in a received signal level that yields a correct symbol decision. The noise component v_m is zero-mean Gaussian with variance

$$\begin{aligned} \sigma_v^2 &= \frac{1}{2}N_0 \int_{-W}^{W} |G_R(f)|^2 df \\ &= \frac{1}{2}N_0 \int_{-W}^{W} |X(f)|^2 df \\ &= \frac{2N_0}{\pi} \end{aligned} \tag{4-68}$$

for both the duobinary and the modified duobinary signals. Hence, an upper bound on the symbol probability of error is

$$P_e$$

$$
\begin{aligned}
< & \sum_{m=-(M-2)}^{M-2} P(|y-2md|>d|B=2md)P(B=2md) \\
& +2P(y+2(M-2)d>d|B=-2(M-2)d)P[B=-2(M-1)d] \\
= & P(|y|>d|B=0) \\
& \times\left\{2\sum_{m=0}^{M-1}P(B=2md)-P(B=0)-P[B=-2(M-1)d]\right\} \\
= & (1-M^{-2})P(|y|>d|B=0) \qquad (4\text{-}69)
\end{aligned}
$$

But

$$
\begin{aligned}
P(|y|>d|B=0) &= \frac{2}{\sqrt{2\pi}\sigma_v}\int_d^\infty e^{-x^2/2\sigma_v^2}dx \\
&= 2Q\left(\sqrt{\frac{\pi d^2}{2N_0}}\right) \qquad (4\text{-}70)
\end{aligned}
$$

Therefore, the average probability of a symbol error is upper-hounded as

$$P_e < 2(1-M^{-2})Q\left(\sqrt{\frac{\pi d^2}{2N_0}}\right) \qquad (4\text{-}71)$$

The scale factor d in Equation 4-71 can be eliminated by expressing it in terms of the average power transmitted into the channel. For the M-ary PAM signal in which the transmitted levels are equally probable, the average power at the output of the transmitting filter is

$$P_{\text{av}} = \frac{E(I_m^2)}{T}\int_{-W}^{W}|G_T(f)|^2df = \frac{E(I_m^2)}{T}\int_{-W}^{W}|X(f)|df = \frac{4}{\pi T}E(I_m^2) \qquad (4\text{-}72)$$

where $E(I_m^2)$ is the mean square value of the M signal levels, which is

$$E(I_m^2) = \frac{1}{3}d^2(M^2-1) \qquad (4\text{-}73)$$

Therefore,

$$d^2 = \frac{3\pi P_{av}T}{4(M^2-1)} \qquad (4\text{-}74)$$

By substituting the value of d^2 from Equation 4-74 into Equation 4-71, we obtain the upper bound on the symbol error probability as

$$P_e < 2\left(1-\frac{1}{M^2}\right)Q\left(\sqrt{\left(\frac{\pi}{4}\right)^2\frac{6}{M^2-1}\frac{\mathcal{E}_{av}}{N_0}}\right) \qquad (4\text{-}75)$$

where $\mathcal{E}_{av}$ is the average energy per transmitted symbol, which can be also expressed in terms of the average bit energy as $\mathcal{E}_{av} = k\mathcal{E}_{bav} = (\log_2 M)\mathcal{E}_{bav}$.

The expression in Equation 4-75 for the probability of error of M-ary PAM holds for both duobinary and modified duobinary partial-response signals. If we compare this result with the error probability of M-ary PAM with zero ISI, which can be obtained by using a signal pulse with a raised cosine spectrum, we note that the performance of partial-response duobinary or modified duobinary has a loss of $\left(\frac{1}{4}\pi\right)^2$. This loss in SNR is due to the fact that the detector for the partial-response signals makes decisions on a symbol-by-symbol basis, and ignores the inherent memory contained in the received signal at its input.

Maximum-likelihood sequence detection

It is clear from the above discussion that partial-response waveforms are signal waveforms with memory. This memory is conveniently represented by a trellis. For example, the trellis for the duobinary partial-response signal for binary data transmission is illustrated in Figure 4-12. For

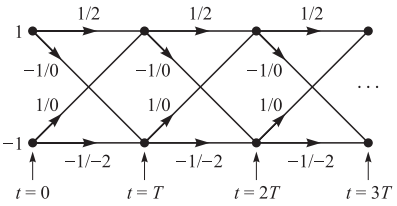

Figure 4-12: Trellis for duobinary partial-response signal.

binary modulation, this trellis contains two states, corresponding to the two possible input values of I_m, *i.e.*, $I_m = \pm 1$. Each branch in the trellis is labeled by two numbers. The first number on the left is the new data bit, *i.e.*, $I_{m+1} = \pm 1$. This number determines the transition to the new state. The number on the right is the received signal level.

The duobinary signal has a memory of length $L = 1$. Hence, for binary modulation the trellis has $S_t = 2$ states. In general, for M-ary modulation, the number of trellis states is M^L.

The optimum maximum-likelihood (ML) sequence detector selects the most probable path through the trellis upon observing the received data sequence $\{y_m\}$ at the sampling instants $t = mT$, $m = I, 2, \cdots$. In general, each node in the trellis will have M incoming paths and M corresponding metrics. One out of the M incoming paths is selected as the most probable, based on the values of the metrics and the other $M - 1$ paths and their metrics are discarded. The surviving path at each node is then extended to M new paths, one for each of the M possible input symbols, and the search process continues. This is basically the Viterbi algorithm for performing the trellis search.

4.2.4 Signal Design for Channels with Distortion

In Sections 4.2.1 and 4.2.2, we described signal design criteria for the modulation filter at the transmitter and the demodulation filter at the receiver when the channel is ideal. In this section, we perform the signal design under the condition that the channel distorts the transmitted signal. We assume that the channel frequency-response $C(f)$ is known for $|f| \leq W$ and that $C(f) = 0$ for $|f| > W$. The filter responses $G_T(f)$ and $G_R(f)$ may be selected to minimize the error probability at the detector. The additive channel noise is assumed to be Gaussian with power spectral density $\mathcal{S}_{nn}(f)$. Figure 4-13 illustrates the overall system under consideration.

For the signal component at the output of the demodulator, we must satisfy the condition

$$G_T(f)C(f)G_R(f) = X_d(f)e^{-j2\pi f t_0}, \quad |f| \leq W \tag{4-76}$$

where $X_d(f)$ is the desired frequency response of the cascade of the modulator, channel, and demodulator, and t_0 is a time delay that is necessary to ensure the physical realizability of the modulation and demodulation filters. The desired frequency response $X_d(f)$ may be selected to yield either zero ISI or controlled ISI at the sampling instants. We shall consider the case of zero ISI by selecting $X_d(f) = X_{rc}(f)$, where $X_{rc}(f)$ is the raised cosine spectrum with an arbitrary roll-off factor.

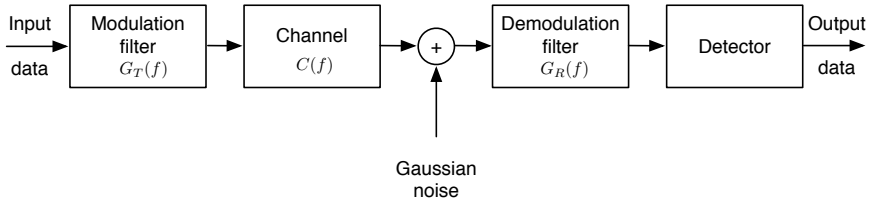

Figure 4-13: System model for the design of the modulation and demodulation filters.

The noise at the output of the demodulation filter may be expressed as

$$v(t) = \int_{-\infty}^{\infty} n(t-\tau) g_R(\tau) d\tau \tag{4-77}$$

where $n(t)$ is the input to the filter. Since $n(t)$ is zero-mean Gaussian, $v(t)$ is zero-mean Gaussian, with a power spectral density

$$\mathcal{S}_{vv}(f) = \mathcal{S}_{nn}(f)|G_R(f)|^2 \tag{4-78}$$

For simplicity, we consider binary PAM transmission. Then, the sampled output of the matched filter is

$$y_m = x_0 I_m + v_m = I_m + v_m \tag{4-79}$$

where x_0 is normalized[4] to unity, $I_m = \pm d$, and v_m represents the noise term, which is zero-mean Gaussian with variance

$$\sigma_v^2 = \int_{-\infty}^{\infty} \mathcal{S}_{nn}(f)|G_R(f)|^2 df \tag{4-80}$$

Consequently, the probability of error is

$$P_2 = \frac{1}{\sqrt{2\pi}} \int_{d/\sigma^2}^{\infty} e^{-y^2/2} dy = Q\left(\sqrt{\frac{d^2}{\sigma_v^2}}\right) \tag{4-81}$$

The probability of error is minimized by maximizing the ratio d^2/σ_v^2 or, equivalently, by minimizing the noise-to-signal ratio σ_v^2/d^2.

Let us consider two possible solutions for the case in which the additive Gaussian noise is white, so that $\mathcal{S}_{nn}(f) = N_0/2$. First, suppose that we precompensate for the total channel distortion at the transmitter, so that the filter at the receiver is matched to the received signal. In this case, the transmitter and receiver filters have the magnitude characteristics

$$\begin{cases} |G_T(f)| &= \frac{\sqrt{X_{rc}(f)}}{|C(f)|}, \quad |f| \le W \\ |G_R(f)| &= \sqrt{X_{rc}(f)}, \quad |f| \le W \end{cases} \tag{4-82}$$

The phase characteristic of the channel frequency response $C(f)$ may also be compensated at the transmitter filter. For these filter characteristics, the average transmitted power is

$$P_{av} = \frac{E\{I_m^2\}}{T} \int_{-\infty}^{\infty} g_T^2(t) dt = \frac{d^2}{T} \int_{-W}^{W} |G_T(f)|^2 df = \frac{d^2}{T} \int_{-W}^{W} \frac{X_{rc}(f)}{|C(f)|^2} df \tag{4-83}$$

[4]By setting $x_0 = 1$ and $I_m = \pm d$, the scaling by x_0 is incorporated into the parameter d.

and, hence,

$$d^2 = P_{av}T\left[\int_{-W}^{W}\frac{X_{rc}(f)}{|C(f)|^2}df\right]^{-1} \tag{4-84}$$

The noise variance at the output of the receiver filter is $\sigma_v^2 = N_0/2$ and, hence, the SNR at the detector is

$$\frac{d^2}{\sigma_v^2} = \frac{2P_{av}T}{N_0}\left[\int_{-W}^{W}\frac{X_{rc}(f)}{|C(f)|^2}df\right]^{-1} \tag{4-85}$$

As an alternative, suppose we split the channel compensation equally between the transmitter and receiver filters, *i.e.*,

$$\begin{cases} |G_T(f)| &= \frac{\sqrt{X_{rc}(f)}}{|C(f)|^{1/2}}, \quad |f| \le W \\ |G_R(f)| &= \frac{\sqrt{X_{rc}(f)}}{|C(f)|^{1/2}}, \quad |f| \le W \end{cases} \tag{4-86}$$

The phase characteristic of $C(f)$ may also be split equally between the transmitter and receiver filters. In this case, the average transmitter power is

$$P_{av} = \frac{d^2}{T}\int_{-W}^{W}\frac{X_{rc}(f)}{|C(f)|}df \tag{4-87}$$

and the noise variance at the output of the receiver filter is

$$\sigma_v^2 = \frac{N_0}{2}\int_{-W}^{W}\frac{X_{rc}(f)}{|C(f)|}df \tag{4-88}$$

Hence, the SNR at the detector is

$$\frac{d^2}{\sigma_v^2} = \frac{2P_{av}T}{N_0}\left[\int_{-W}^{W}\frac{X_{rc}(f)}{|C(f)|^2}df\right]^{-2} \tag{4-89}$$

From Equations 4-85 and 4-89, we observe that when we express the SNR d^2/σ_v^2 in terms of the average transmitter power P_{av}, there is a loss incurred due to channel distortion. In the case of the filters given by Equation 4-82, the loss is

$$10\log\int_{-W}^{W}\frac{X_{rc}(f)}{|C(f)|^2}df \tag{4-90}$$

and, in the case of the filters given by Equation 4-86, the loss is

$$10\log\left[\int_{-W}^{W}\frac{X_{rc}(f)}{|C(f)|^2}df\right]^2 \tag{4-91}$$

We observe that when $C(f) = 1$ for $|f| \le W$, the channel is ideal and

$$\int_{-W}^{W}X_{rc}(f)df = 1$$

so that no loss is incurred. On the other hand, when there is amplitude distortion, $|C(f)| < 1$ for some range of frequencies in the band $|f| < W$ and, hence, there is a loss in SNR as given by Equations 4-90 and 4-91. The interested reader may show (see Problem 9.30) that the filters given by Equation 4-86 result in the smaller SNR loss.

Example 4. Let us determine the transmitting and receiving filters given by Equation 4-86 for a binary communication system that transmits data at a rate of 4800 bits/s over a channel with frequency (magnitude) response

$$|C(f)| = \frac{1}{\sqrt{1+(f/W)^2}}, \quad |f| \leq w \tag{4-92}$$

where $W = 4800$ Hz. The additive noise is zero-mean white Gaussian with spectral density $\frac{1}{2}N_0 = 10^{-15}$ W/Hz.

Since $W = 1/T = 4800$, we use a signal pulse with a raised cosine spectrum and $\beta = 1$. Thus,

$$X_{rc}(f) = \frac{1}{2}T\left[1+\cos(\pi T|f|)\right] = T\cos^2\left(\frac{\pi|f|}{9600}\right) \tag{4-93}$$

Then,

$$|G_T(f)| = |G_R(f)| = \left[1+\left(\frac{f}{4800}\right)^2\right]^{1/4}\cos\left(\frac{\pi|f|}{9600}\right), \quad |f| \leq 4800 \tag{4-94}$$

and $|G_T(f)| = |G_R(f)| = 0$, otherwise. Figure 4-14 illustrates the filter characteristic $G_T(f)$.

One can now use these filters to determine the amount of transmitted energy E required to achieve a specified error probability. This problem is left as an exercise for the reader.

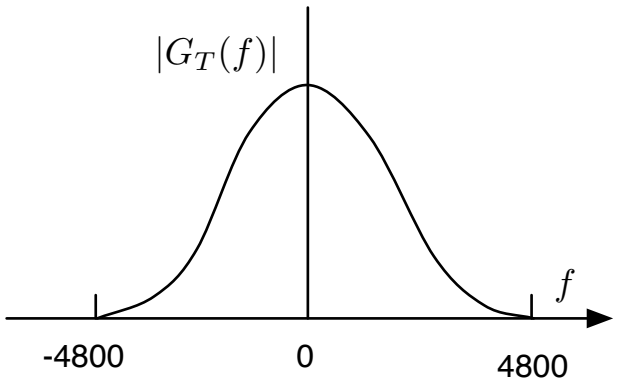

Figure 4-14: Frequency response of an optimum transmitter filter.

4.3 Optimum Receiver for Channels with ISI and AWGN

In this section, we derive the structure of the optimum demodulator and detector for digital transmission through a nonideal band-limited channel with additive Gaussian noise. We begin with the transmitted (equivalent lowpass) signal given by Equation 4-5. The received (equivalent lowpass) signal is expressed as

$$r(t) = \sum_n I_n h(t - nT) + z(t) \tag{4-95}$$

where $h(t)$ represents the response of the channel to the input signal pulse $g(t)$ and $z(t)$ represents the additive white Gaussian noise.

First we demonstrate that the optimum demodulator can be realized as a filter matched to $h(t)$, followed by a sampler operating at the symbol rate $1/T$ and a subsequent processing algorithm for estimating the information sequence $\{I_n\}$ from the sample values. Consequently, the samples at the output of the matched filter are sufficient for the estimation of the sequence $\{I_n\}$.

4.3.1 Optimum Maximum-Likelihood Receiver

Using the Karhunen-Loève expansion, we expand the received signal $r_l(t)$ in the series

$$r_l(t) = \lim_{N\to\infty} \sum_{k=1}^{N} r_k \phi_k(t) \tag{4-96}$$

where $\{\phi_k(t)\}$ is a complete set of orthonormal functions and $\{r_d\}$ are the observable random variables obtained by projecting $r_l(t)$ onto the set $\{\phi_k(t)\}$. It is easily shown that

$$r_k(t) = \sum_n I_n h_{kn} + z_k, \quad k = 1, 2, \cdots \tag{4-97}$$

where h_{kn} is the value obtained from projecting $h(t-nT)$ onto $\phi_k(t)$, and z_k is the value obtained from projecting $z(t)$ onto $\phi_k(t)$. The sequence $\{z_k\}$ is Gaussian with zero-mean and covariance

$$E\{z_k^* z_m\} = 2N_0 \delta(k - m) \tag{4-98}$$

The joint probability density function of the random variables $\boldsymbol{r}_N \equiv [r_1, r_2, \cdots, r_N]$ conditioned on the transmitted sequence $\boldsymbol{I}_p \equiv [I_1, I_2, \cdots, I_p]$, where $p \le N$, is

$$p(\boldsymbol{r}_N | \boldsymbol{I}_p) = \left(\frac{1}{2\pi N_0}\right)^N \exp\left(-\frac{1}{2N_0} \sum_{k=1}^{N} \left| r_k - \sum_n I_n h_{kn} \right|^2\right) \tag{4-99}$$

In the limit as the number N of observable random variables approaches infinity, the logarithm of $p(\boldsymbol{r}_N|\boldsymbol{I}_p)$ is proportional to the metrics $PM(\boldsymbol{I}_p)$ defined as

$$\begin{aligned} PM(\boldsymbol{I}_p) &= -\int_{-\infty}^{\infty} \left| r_l(t) - \sum_n I_n h(t - nT) \right|^2 dt \\ &= -\int_{-\infty}^{\infty} |r_l(t)|^2 \, dt - 2\Re \sum_n \left[I_n^* \int_{-\infty}^{\infty} r_l(t) h^*(t - nT) dt \right] \end{aligned}$$

$$-\sum_n \sum_m I_n^* I_m \int_{-\infty}^{\infty} h^*(t-nT)h(t-mT)dt \tag{4-100}$$

The maximum-likelihood estimates of the symbols $I_1, I_2, \cdots, I_p$ are those that maximize this quantity. Note, however, that the integral of $|r_l(t)|^2$ is common to all metrics, and, hence, it may be discarded. The other integral involving $r(t)$ gives rise to the variables

$$y_n \equiv y(nT) = \int_{-\infty}^{\infty} r_l(t)h^*(t-nT)dt \tag{4-101}$$

These variables can be generated by passing $r(t)$ through a filter matched to $h(t)$ and sampling the output at the symbol rate $1/T$. The samples $\{y_n\}$ form a set of sufficient statistics for the computation of $PM(\boldsymbol{I}_p)$ or, equivalently, of the correlation metrics

$$CM(\boldsymbol{I}_p) = 2\Re\left(\sum_n I_n^* y_n\right) - \sum_n \sum_m I_n^* I_m x_{n-m} \tag{4-102}$$

where, by definition, $x(t)$ is the response of the matched filter to $h(t)$ and

$$x_n \equiv x(nT) = \int_{-\infty}^{\infty} h^*(t)h(t+nT)dt \tag{4-103}$$

Hence, $x(t)$ represents the output of a filter having an impulse response $h^*(-t)$ and an excitation $h(t)$. In other words, $x(t)$ represents the autocorrelation function of $h(t)$. Consequently, $\{x_n\}$ represents the samples of the autocorrelation function of $h(t)$, taken periodically at $1/T$. We are not particularly concerned with the noncausal characteristic of the filter matched to $h(t)$, since, in practice, we can introduce a sufficiently large delay to ensure causality of the matched filter.

If we substitute for $r_l(t)$ in Equation 4-101 using Equation 4-95, we obtain

$$y_k = \sum_n I_n x_{k-n} + \nu_k \tag{4-104}$$

where ν_k denotes the additive noise sequence of the output of the matched filter, *i.e.*,

$$\nu_k = \int_{-\infty}^{\infty} z(t)h^*(t-kT)dt \tag{4-105}$$

The output of the demodulator (matched filter) at the sampling instants is corrupted by ISI as indicated by Equation 4-104. In any practical system, it is reasonable to assume that the ISI affects a finite number of symbols. Hence, we may assume that $x_n = 0$ for $|n| > L$. Consequently, the ISI observed at the output of the demodulator may be viewed as the output of a finite state machine. This implies that the channel output with ISI may be represented by a trellis diagram, and the maximum-likelihood estimate of the information sequence $(I_1, I_2, \cdots, I_p)$ is simply the most probable path through the trellis given the received demodulator output sequence $\{y_n\}$. Clearly, the Viterbi algorithm provides an efficient means for performing the trellis search.

The metrics that are computed for the MLSE of the sequence $\{I_k\}$ are given by Equation 4-102. It can be seen that these metrics can be computed recursively in the Viterbi algorithm, according to the relation

$$CM_n(\boldsymbol{I}_n) = CM_{n-1}(\boldsymbol{I}_{n-1}) + \Re\left[I_n^*\left(2y_n - x_0 I_n - 2\sum_{m=1}^{L} x_m I_{n-m}\right)\right] \tag{4-106}$$

Figure 4-15 illustrates the block diagram of the optimum receiver for an AWGN channel with ISI.

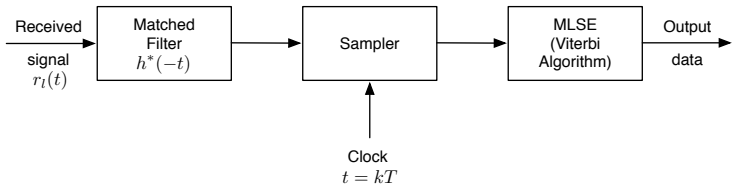

Figure 4-15: Optimum receiver for an AWGN channel with ISI.

4.3.2 A Discrete-Time Model for a Channel with ISI

In dealing with band-limited channels that result in ISI, it is convenient to develop an equivalent discrete-time model for the analog (continuous-time) system. Since the transmitter sends discrete-time symbols at a rate of $1/T$ symbols/s and the sampled output of the matched filter at the receiver is also a discrete-time signal with samples occurring at a rate of $1/T$ per second, it follows that the cascade of the analog filter at the transmitter with impulse response $g(t)$, the channel with impulse response $c(t)$, the matched filter at the receiver with impulse response $h^*(-t)$, and the sampler can be represented by an equivalent discrete-time tranversal filter having tap gain coefficients $\{x_k\}$. Consequently, we have an equivalent discrete-time transversal filter that spans a time interval of $2LT$ seconds. Its input is the sequence of information symbols $\{I_k\}$ and its output is the discrete-time sequence $\{y_k\}$ given by Equation 4-104. The equivalent discrete-time model is shown in Figure 4-16.

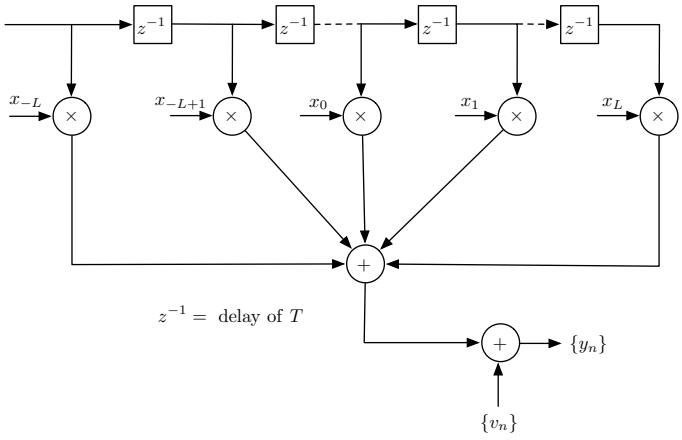

Figure 4-16: Equivalent discrete-time model of channel with intersymbol interference.

The major difficulty with this discrete-time model occurs in the evaluation of performance of the various equalization or estimation techniques that are discussed in the following sections. The difficulty is caused by the correlations in the noise sequence $\{v_k\}$ at the output of the matched filter. That is, the set of noise variables $\{v_k\}$ is a Gaussian-distributed sequence with zero-mean and autocorrelation function

$$E\{v_k^* v_j\} = \begin{cases} 2N_0 x_{j-k} & |k-j| \le L \\ 0 & \text{otherwise} \end{cases} \tag{4-107}$$

Hence, the noise sequence is correlated unless $x_k = 0$, $k \neq 0$. Since it is more convenient to deal with the white noise sequence when calculating the error rate performance, it is desirable to whiten the noise sequence by further filtering the sequence $\{y_k\}$. A discrete-time noise-whitening filter is determined as follows.

Let $X(z)$ denote the (two-sided) z transform of the sampled autocorrelation function $\{x_k\}$, *i.e.*,

$$X(z) = \sum_{k=-L}^{L} x_k z^{-k} \tag{4-108}$$

Since $x_k = x_{-k}^*$, it follows that $X(z) = X^*(1/z^*)$ and the $2L$ roots of $X(z)$ have the symmetry that if ρ is a root, $1/\rho^*$ is also a root. Hence, $X(z)$ can be factored and expressed as

$$X(z) = F(z)F^*\left(\frac{1}{z^*}\right) \tag{4-109}$$

where $F(z)$ is a polynomial of degree L having the roots $\rho_1, \rho_2, \cdots, \rho_L$ and $F^*(1/z^*)$ is a polynomial of degree L having the roots $1/\rho_1^*, 1/\rho_2^*, \cdots, 1/\rho_L^*$. Assuming that there are no roots on the unit circle, an appropriate noise-whitening filter has a z transform $1/F^*(1/z^*)$. Since there are $2L$ possible choices for the roots of $F^*(1/z^*)$, each choice resulting in a filter characteristic that is identical in magnitude but different in phase from other choices of the roots, we propose to choose the unique $F^*(1/z^*)$ that results in an anticausal impulse response with poles corresponding to the zeros of $X(z)$ that are outside the unit circle. Such an anticausal filter is stable. Selecting the noise-whitening filter in this manner ensures that the resulting channel response, characterized by $F(z)$, is minimum phase. Consequently, passage of the sequence $\{y_k\}$ through the digital filter $1/F^*(1/z^*)$ results in an output sequence $\{v_k\}$ that can be expressed as

$$v_k = \sum_{n=0}^{L} f_n I_{k-n} + \eta_k \tag{4-110}$$

where $\{\eta_k\}$ is a white Gaussian noise sequence and $\{f_k\}$ is a set of tap coefficients of an equivalent discrete-time transversal filter having a transfer function $F(z)$. The cascade of the matched filter, the sampler, and the noise-whitening filter is called the *whitened matched filter* (WMF).

It is convenient to normalize the energy of $F(z)$ to unity, *i.e.*,

$$\sum_{n=0}^{L} |f_n|^2 = 1$$

The minimum-phase condition on $F(z)$ implies that the energy in the first M values of the impulse response $\{f_0, f_1, \cdots, f_M\}$ is a maximum for every M.

In summary, the cascade of the transmitting filter $g(t)$, the channel $c(t)$, the matched filter $h^*(-t)$, the sampler, and the discrete-time noise-whitening filter $1/F^*(1/z^*)$ can be represented as an equivalent discrete-time transversal filter having the set $\{f_k\}$ as its tap coefficients. The additive noise sequence $\{\eta_k\}$ corrupting the output of the discrete-time transversal filter is a white Gaussian noise sequence having zero-mean and variance N_0. Figure 4-17 illustrates the model of the equivalent discrete-time system with white noise. We refer to this model as the *equivalent discrete-time white noise filter model*.

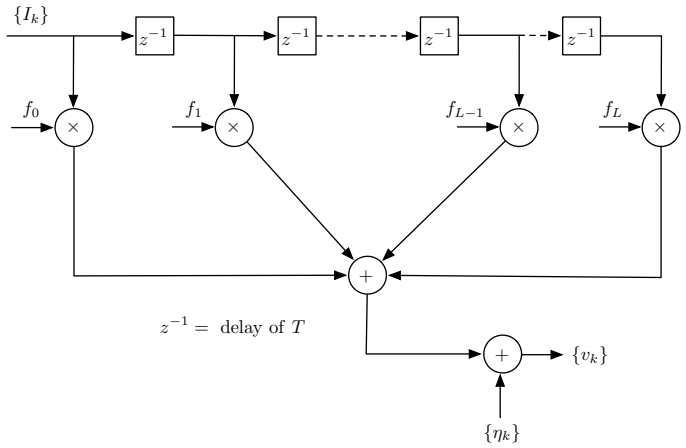

Figure 4-17: Equivalent discrete-time model of intersymbol interference channel with AWGN.

Example 5. Suppose that the transmitter signal pulse $g(t)$ has duration T and unit energy and the received signal pulse is $h(t) = g(t) + ag(t-T)$. Let us determine the equivalent discrete-time white noise filter model. The sampled autocorrelation function is given by

$$x_k = \begin{cases} a^* & k = -1 \\ 1 + |a|^2 & k = 0 \\ a & k = 1 \end{cases} \tag{4-111}$$

The z transform of x_k is

$$\begin{aligned} X(z) &= \sum_{k=-1}^{1} x_k z^{-k} \\ &= a^* z + (1 + |a|^2) + az^{-1} \\ &= (az^{-1} + 1)(a^* z + 1) \end{aligned} \tag{4-112}$$

Under the assumption that $|a| < 1$, one chooses $F(z) = az^{-1}+1$, so that the equivalent transversal filter consists of two taps having tap gain coefficients $f_0 = 1$, $f_1 = a$. Note that the correlation sequence $\{x_k\}$ may be expressed in terms of the $\{f_n\}$ as

$$x_k = \sum_{n=0}^{L-k} f_n^* f_{n+k}, \quad k = 0, 1, 2, \cdots, L \tag{4-113}$$

When the channel impulse response is changing slowly with time, the matched filter at the receiver becomes a time-variable filter. In this case, the time variations of the channel/matched-filter pair result in a discrete-time filter with time-variable coefficients. As a consequence, we

have time-variable intersymbol interference effects, which can be modeled by the filter illustrated in Figure 4-17, where the tap coefficients are slowly varying with time.

The discrete-time white noise linear filter model for the intersymbol interference effects that arise in high-speed digital transmission over nonideal band-limited channels will be used throughout the remainder of this chapter in our discussion of compensation techniques for the interference. In general, the compensation methods are called *equalization techniques* or *equalization algorithms.*

4.3.3 Maximum-Likelihood Sequence Estimation (MLSE) for the Discrete-Time White Noise Filter Model

In the presence of intersymbol interference that spans $L+1$ symbols (L interfering components), the MLSE criterion is equivalent to the problem of estimating the state of a discrete-time finite-state machine. The finite-state machine in this case is the equivalent discrete-time channel with coefficients $\{f_k\}$, and its state at any instant in time is given by the L most recent inputs, *i.e.*, the state at time k is

$$S_k = (I_{k-1}, I_{k-2}, \cdots, I_{k-L}) \tag{4-114}$$

where $I_k = 0$ for $k \leq 0$. Hence, if the information symbols are M-ary, the channel filter has M^L states. Consequently, the channel is described by an M^L-state trellis and the Viterbi algorithm may be used to determine the most probable path through the trellis.

The metrics used in the trellis search are akin to the metrics used in soft-decision decoding of convolutional codes. In brief, we begin with the samples $v_1, v_2, \cdots, v_{L+1}$, from which we compute the M^{L+1} metrics

$$\sum_{k=1}^{L+1} \ln p(v_k | I_k, I_{k-1}, \cdots, I_{k-L}) \tag{4-115}$$

The M^{L+1} possible sequences of $(I_{L+1}, I_L, \cdots, I_2, I_1)$ are subdivided into M^L groups corresponding to the M^L states $(I_{L+1}, I_L, \cdots, I_2)$. Note that the M sequences in each group (state) differ in I_1 and correspond to the paths through the trellis that merge at a single node. From the M sequences in each of the M^L states, we select the sequence with the largest probability (with respect to I_1) and assign to the surviving sequence the metric

$$\begin{aligned} PM_1(\boldsymbol{I}_{L+1}) &\equiv PM_1(I_{L+1}, I_L, \cdots, I_2) \\ &= \max_{I_1} \sum_{k=1}^{L+1} \ln p(v_k | I_k, I_{k-1}, \cdots, I_{k-L}) \end{aligned} \tag{4-116}$$

The $M-1$ remaining sequences from each of the M^L groups are discarded. Thus, we are left with M^L surviving sequences and their metrics.

Upon reception of v_{L+2}, the M^L surviving sequences are extended by one stage, and the corresponding M^{L+1} probabilities for the extended sequences are computed using the previous metrics and the new increment, which is $\ln p(v_{L+2} | I_{L+2}, I_{L+1}, \cdots, I_2)$. Again, the M^{L+1} sequences are subdivided into M^L groups corresponding to the M^L possible states $(I_{L+2}, \cdots, I_3)$ and the most probable sequence from each group is selected, while the other $M-1$ sequences are discarded.

The procedure described continues with the reception of subsequent signal samples. In general, upon reception of v_{L+k}, the metrics[5]

$$PM_k(\boldsymbol{I}_{L+k}) = \max_{I_k} [\ln p(v_{L+k}|I_{L+k}, \cdots, I_k) + PM_{k-1}(\boldsymbol{I}_{L+k-1})] \tag{4-117}$$

that are computed give the probabilities of the M^L surviving sequences. Thus, as each signal sample is received, the Viterbi algorithm involves first the computation of the M^{L+1} probabilities

$$\ln p(v_{L+k}|I_{L+k}, \cdots, I_k) + PM_{k-1}(\boldsymbol{I}_{L+k-1}) \tag{4-118}$$

corresponding to the M^{L+1} sequences that form the continuations of the M^L surviving sequences from the previous stage of the process. Then the M^{L+1} sequences are subdivided into M^L groups, with each group containing M sequences that terminate in the same set of symbols $I_{L+k}, \cdots, I_{k+1}$ and differ in the symbol I_k. From each group of M sequences, we select the one having the largest probability as indicated by Equation 4-117, while the remaining $M-1$ sequences are discarded. Thus, we are left again with M^L sequences having the metrics $PM_k(\boldsymbol{I}_{L+k})$.

As indicated previously, the delay in detecting each information symbol is variable. In practice, the variable delay is avoided by truncating the surviving sequences to the q most recent symbols, where $q \gg L$, thus achieving a fixed delay. In the case that the M^L surviving sequences at time k disagree on the symbol I_{k-q}, the symbol in the most probable sequence may be chosen. The loss of performance resulting from this suboptimum decision procedure is negligible if $q \geq 5L$.

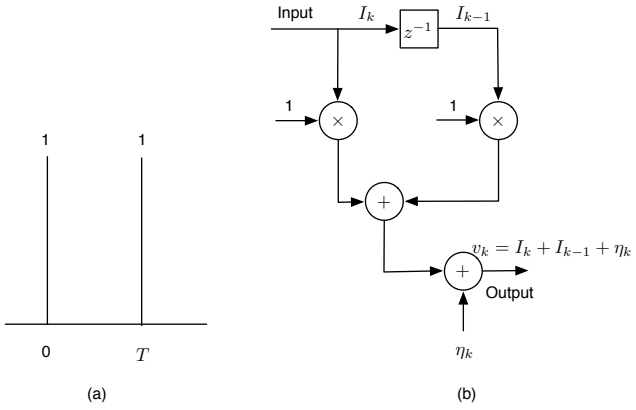

Figure 4-18: Equivalent discrete-time model for intersymbol interference resulting from a duobinary pulse.

Example 6. For illustrative purposes, suppose that a duobinary signal pulse is employed to transmit four-level ($M = 4$) PAM. Thus, each symbol is a number selected

[5]We observe that the metrics $PM_k(I)$ are simply related to the Euclidean distance metrics $DM_k(I)$ when the additive noise is Gaussian.

from the set $\{-3, -1, 1, 3\}$. The controlled intersymbol interference in this partial-response signal is represented by the equivalent discrete-time channel model shown in Figure 4-18. Suppose we have received v_1 and v_2, where

$$\begin{cases} v_1 &= I_1 + \eta_1 \\ v_2 &= I_2 + I_1 + \eta_2 \end{cases} \tag{4-119}$$

and $\{\eta_i\}$ is a sequence of statistically independent zero-mean Gaussian noise. We may now compute the 16 metrics

$$PM_1(I_2, I_1) = -\sum_{k=1}^{2} \left(v_k - \sum_{j=0}^{1} I_{k-j} \right)^2, \quad I_1, I_2 = \pm 1, \pm 3 \tag{4-120}$$

where $I_k = 0$ for $k \leq 0$.

Note that any subsequently received signals $\{v_i\}$ do not involve I_1. Hence, at this stage, we may discard 12 of the 16 possible pairs $\{I_1, I_2\}$. This step is illustrated by the tree diagram shown in Figure 4-19. In other words, after computing the 16 metrics corresponding to the 16 paths in the tree diagram, we discard three out of the four paths that terminate with $I_2 = 3$ and save the most probable of these four. Thus, the metric for the surviving path is

$$PM_1(I_2 = 3, I_1) = \max_{I_1} \left[-\sum_{k=1}^{2} \left(v_k - \sum_{j=0}^{1} I_{k-j} \right)^2 \right]$$

The process is repeated for each set of four paths terminating with $I_2 = 1, I_2 = -1$, and $I_2 = -3$. Thus four paths and their corresponding metrics survive after v_1 and v_2 are received.

When v_3 is received, the four paths are extended as shown in Figure 4-19 to yield 16 paths and 16 corresponding metrics given by

$$PM_2(I_3, I_2, I_1) = PM_1(I_2, I_1) - \left(v_3 - \sum_{j=0}^{1} I_{3-j} \right)^2 \tag{4-121}$$

Of the four paths terminating with the $l_3 = 3$, we save the most probable. This procedure is again repeated for $I_3 = 1$, $I_3 = -1$, and $I_3 = -3$. Consequently, only four paths survive at this stage. The procedure is then repeated for each subsequently received signal v_k for $k > 3$.

4.3.4 Performance of MLSE for Channels with ISI

We shall now determine the probability of error for the MLSE of the received information sequence when the information is transmitted via PAM and the additive noise is Gaussian. The similarity between a convolutional code and a finite-duration intersymbol interference channel implies that

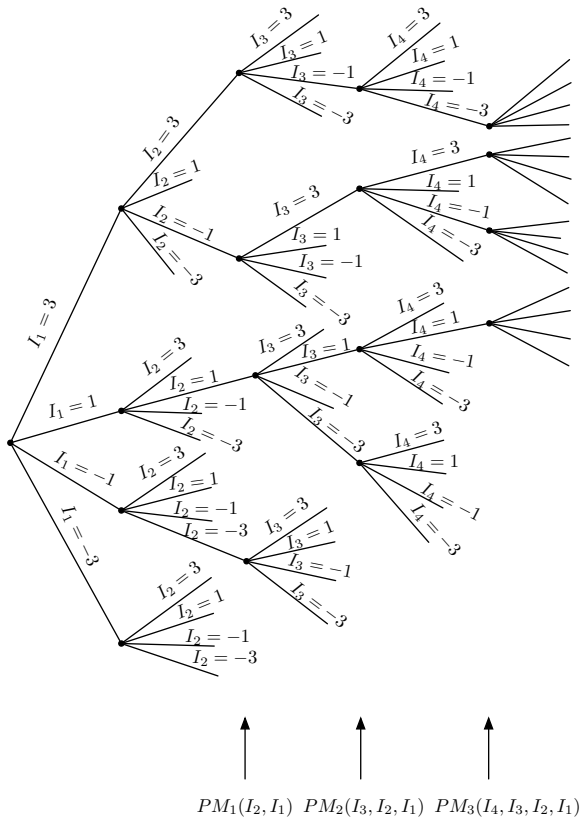

Figure 4-19: Tree diagram for Viterbi decoding of the duobinary pulse.

the method for computing the error probability for the latter carries over from the former. In particular, the method for computing the performance of soft-decision decoding of a convolutional code by means of the Viterbi algorithm applies with some modification.

In PAM signaling with the additive Gaussian noise and intersymbol interference, the metrics used in the Viterbi algorithm may be expressed as in Equation 4-117, or, equivalently, as

$$PM_{k-L}(\boldsymbol{I}_k) = PM_{k-L-1}(\boldsymbol{I}_{k-1}) - \left(v_k - \sum_{j=0}^{L} f_j I_{k-j} \right)^2 \tag{4-122}$$

where the symbols $\{I_n\}$ may take the values $\pm d, \pm 3d, \cdots, \pm(M-1)d$, and $2d$ is the distance between successive levels. The trellis has M^L states, defined at time k as

$$S_k = (I_{k-1}, I_{k-2}, \cdots, I_{k-L}) \tag{4-123}$$

Let the estimated symbols from the Viterbi algorithm be denoted by $\{\tilde{I}_n\}$ and the corresponding estimated state at time k by

$$\tilde{S}_k = (\tilde{I}_{k-1}, \tilde{I}_{k-2}, \cdots, \tilde{I}_{k-L}) \tag{4-124}$$

Now suppose that the estimated path through the trellis from the path at time k and remerges with the correct path at time $k+l$. Thus, $\tilde{S}_k = S_k$ and $\tilde{S}_{k+l} = \tilde{S}_{k+l}$, but $\tilde{S}_m \neq S_m$ for

$k < m < k + l$. As in a convolutional code, we call this an *error event*. Since the channel spans $L + 1$ symbols, it follows that $l \geq L + 1$.

For such an error event, we have $\tilde{I}_k \neq I_k$ hand $\tilde{I}_{k+l-L-1} \neq I_{k+l-L-1}$, but $\tilde{I}_m = I_m$ for $k - L \leq m \leq k - 1$ and $k + l - L \leq m \leq k + l - 1$. It is convenient to define an error vector $\boldsymbol{\varepsilon}$ corresponding to this error event as

$$\boldsymbol{\varepsilon} = [\varepsilon_k, \varepsilon_{k+1}, \cdots, \varepsilon_{k+l-L-1}] \tag{4-125}$$

where the components of $\boldsymbol{\varepsilon}$ are defined as

$$\varepsilon_j = \frac{1}{2d}\left(I_j - \hat{I}_j\right), \; j = k, k+1, \cdots, k+l-L-1 \tag{4-126}$$

The normalization factor of $2d$ in Equation 4-126 results in elements ε_j that take on the values 0, ± 1, ± 2, ± 3, $\cdots$, $\pm(M-1)$. Moreover, the error vector is characterized by the properties that $\varepsilon_k \neq 0$, $\varepsilon_{k+l-L-1} \neq 0$, and there is no sequence of L consecutive elements that are zero. Associated with the error vector in Equation 4-125 is the polynomial of degree $l - L - 1$,

$$\varepsilon(z) = \varepsilon_k + \varepsilon_{k+1} z^{-1} + \varepsilon_{k+2} z^{-2} + \cdots + \varepsilon_{k+l-L-1} z^{-(l-L-1)} \tag{4-127}$$

We wish to determine the probability of occurrence of the error event that begins at time k and is characterized by the error vector $\boldsymbol{\varepsilon}$ given in Equation 4-125 or, equivalently, by the polymonial given in Equation 4-127. To accomplish this, we follow the procedure developed by Forney (1972). Specifically, for the error event $\boldsymbol{\varepsilon}$ to occur, the following three subevents E_1, E_2, and E_3 must occur:

E_1**:** At time k, $\tilde{S}_k = S_k$.

E_2**:** The information symbols $I_k, I_{k+1}, \cdots, I_{k+l-L-1}$ when added to the scaled error sequence $2d\,(\varepsilon_k, \varepsilon_{k+1}, \cdots, \varepsilon_{k+l-L-1})$ must result in an allowable sequence, *i.e.*, the sequence $\tilde{I}_k, \tilde{I}_{k+1}, \cdots, \tilde{I}_{k+l-L-1}$ must have values selected from $\pm 1, \pm 2, \pm 3, \cdots, \pm(M-1)$.

E_3**:** For $k \leq k + l$, the sum of the branch metrics of the estimated path exceeds the sum of the branch metrics of the correct path.

The probability of occurrence of E_3 is

$$P(E_3) = P\left[\sum_{i=k}^{k+l-1}\left(v_i - \sum_{j=0}^{L} f_i \tilde{I}_{i-j}\right)^2 < \sum_{i=k}^{k+l-1}\left(v_i - \sum_{j=0}^{L} f_i I_{i-j}\right)^2\right] \tag{4-128}$$

But

$$v_i = \sum_{j=0}^{L} f_j I_{i-j} + \eta_j \tag{4-129}$$

where $\{\eta_i\}$ is a real-valued white Gaussian noise sequence. Substitution of Equation 4-129 into Equation 4-137 yields

$$P(E_3) \;=\; P\left[\sum_{i=k}^{k+l-1}\left(\eta_i + 2d\sum_{j=0}^{L} f_i \varepsilon_{i-j}\right)^2 < \sum_{i=k}^{k+l-1} \eta_i^2\right]$$

$$= \quad P\left[4d\sum_{i=k}^{k+l-1}\eta_i\left(\sum_{j=0}^{L}f_i\varepsilon_{i-j}\right) < -4d^2\sum_{i=k}^{k+l-1}\left(\sum_{j=0}^{L}f_i\varepsilon_{i-j}\right)^2\right] \tag{4-130}$$

where $\varepsilon_j = 0$ for $j < k$ and $j > k + l - L - 1$. If we define

$$\alpha_i = \sum_{j=0}^{L} f_j\varepsilon_{i-j} \tag{4-131}$$

the Equation 4-130 may be expressed as

$$P(E_3) \quad = \quad P\left(\sum_{i=k}^{k+l-1}\alpha_i\eta_i < -d\sum_{i=k}^{k+l-1}\alpha_i^2\right) \tag{4-132}$$

where the factor of $4d$ common to both terms has been dropped. Now Equation 4-132 is just the probability that a linear combination of statistically independent Gaussian random variables is less than some negative number. Thus

$$P(E_3) = Q\left(\sqrt{\frac{2d^2}{N_0}\sum_{i=k}^{k+l-1}\alpha_i^2}\right) \tag{4-133}$$

For convenience, we define

$$\delta^2(\boldsymbol{\varepsilon}) = \sum_{i=k}^{k+l-1}\alpha_i^2 = \sum_{i=k}^{k+l-1}\left(\sum_{j=0}^{L}f_i\varepsilon_{i-j}\right)^2 \tag{4-134}$$

where $\varepsilon_j = 0$ for $j < k$ and $j > k + l - L - 1$. Note that the $\{\alpha_i\}$ resulting from the convolution of $\{f_i\}$ with $\{\varepsilon_j\}$ are the coefficients of the polynomial

$$\begin{aligned}\alpha(z) &= F(z)\varepsilon(z) \\ &= \alpha_k + \alpha_{k+1}z^{-1} + \cdots + \alpha_{k+l-1}z^{-(l-1)}\end{aligned} \tag{4-135}$$

Furthermore, $\delta^2(\boldsymbol{\varepsilon})$ is simply equal to the coefficient of z^0 in the polynomial

$$\begin{aligned}\alpha(z)\alpha(z^{-1}) &= F(z)F(z^{-1})\varepsilon(z)\varepsilon(z^{-1}) \\ &= X(z)\varepsilon(z)\varepsilon(z^{-1})\end{aligned} \tag{4-136}$$

We call $\delta^2(\boldsymbol{\varepsilon})$ the *Euclidean weight* of the error event ε.

An alternative method for representing the result of convolving $\{f_i\}$ with $\{\varepsilon_i\}$ is the matrix form

$$\boldsymbol{\alpha} = \boldsymbol{e}\boldsymbol{f}$$

where $\boldsymbol{\alpha}$ is al l-dimentional vector, $\boldsymbol{f}$ is an $(L+1)$-dimentional vector, and $\boldsymbol{e}$ is an $l \times (L+1)$ matrix defined as

$$\boldsymbol{\alpha} = \begin{bmatrix}\alpha_k \\ \alpha_{k+1} \\ \vdots \\ \alpha_{k+l-1}\end{bmatrix},\ \boldsymbol{f} = \begin{bmatrix}f_0 \\ f_1 \\ \vdots \\ f_L\end{bmatrix},\ \boldsymbol{e} = \begin{bmatrix}\varepsilon_k & 0 & 0 & \cdots & 0 & \cdots & 0 \\ \varepsilon_{k+1} & \varepsilon_k & 0 & \cdots & 0 & \cdots & 0 \\ \varepsilon_{k+2} & \varepsilon_{k+1} & \varepsilon_k & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots\cdots & \vdots & \\ \varepsilon_{k+l-1} & \cdots & \cdots & \cdots & \cdots & \cdots & \varepsilon_{k+l-L-1}\end{bmatrix} \tag{4-137}$$

Then

$$\delta^2(\boldsymbol{\varepsilon}) \quad = \quad \boldsymbol{\alpha}^T\boldsymbol{\alpha} = \boldsymbol{f}^T\boldsymbol{e}^T\boldsymbol{e}\boldsymbol{f} = \boldsymbol{f}^T\boldsymbol{A}\boldsymbol{f} \tag{4-138}$$

where $\boldsymbol{A}$ is an $(L+1)\times(L+1)$ matrix of the form

$$\boldsymbol{A} = \boldsymbol{e}^T\boldsymbol{e} = \begin{bmatrix} \beta_0 & \beta_1 & \beta_2 & \cdots & \beta_L \\ \beta_1 & \beta_0 & \beta_1 & \cdots & \beta_{L-1} \\ \beta_2 & \beta_1 & \beta_0 & \cdots & \beta_{L-2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \beta_L & \cdots & \cdots & \cdots & \beta_0 \end{bmatrix} \tag{4-139}$$

and

$$\beta_m = \sum_{i=k}^{k+l-1-m} \varepsilon_i\varepsilon_{i+m} \tag{4-140}$$

We may use either Equations 4-134 and 4-135 or Equations 4-139 and 4-140 in evaluating the error rate performance. We consider these computations later. For now we conclude that the probability of the subevent E_3, given by Equations 4-133, may be expressed as

$$P(E_3) = Q\left(\sqrt{\frac{2d^2}{N_0}\delta^2(\boldsymbol{\varepsilon})}\right) = Q\left(\sqrt{\frac{6}{M^2-1}\gamma_{\text{av}}\delta^2(\boldsymbol{\varepsilon})}\right) \tag{4-141}$$

where we have used the relation

$$d^2 = \frac{3}{M^2-1}TP_{\text{av}} \tag{4-142}$$

to eliminate d^2 and $\gamma_{\text{av}} = TP_{\text{av}}/N_0$. Note that, in the absence of intersymbol interference, $\delta^2(\boldsymbol{\varepsilon}) = 1$ and $P(E_3)$ is proportional to the symbol error probability of M-ary PAM.

The probability of the subevent E_2 depends only on the statistical properties of the input sequence. We assume that the information symbols are equally probable and that the symbols in the transmitted sequence are statistically independent. Then, for an error of the form $|\varepsilon_i| = j$, $j = 1, 2, \cdots, M-1$, there are $M-j$ possible values of I_i such that

$$I_i = \tilde{I}_i + 2d\varepsilon_i$$

Hence

$$P(E_2) = \prod_{i=0}^{l-L-1} \frac{M-|\varepsilon_i|}{M} \tag{4-143}$$

The probability of the subevent E_1 is much more difficult to compute exactly because of its dependence on the subevent E_3. That is, we must compute $P(E_1|E_3)$. However, $P(E_1|E_3) = 1-P_e$, where P_e is the symbol error probability. Hence $P(E_1|E_3)$ is well approximated (and upper-bounded) by unity for reasonably low symbol error probabilities. Therefore, the probability of the error event $\boldsymbol{\varepsilon}$ is well approximated and upper-bounded as

$$P(\boldsymbol{\varepsilon}) = Q\left(\sqrt{\frac{6}{M^2-1}\gamma_{\text{av}}\delta^2(\boldsymbol{\varepsilon})}\right)\prod_{i=0}^{l-L-1}\frac{M-|\varepsilon_i|}{M} \tag{4-144}$$

Let E be the set of all error events $\boldsymbol{\varepsilon}$ starting at time k and let $w(\boldsymbol{\varepsilon})$ be the corresponding number of nonzero components (Hamming weight or number of symbol errors) in each error event $\boldsymbol{\varepsilon}$. Then the probability of a symbol error is upper-bounded (union bound) as

$$\begin{aligned} P_e &\leq \sum_{\boldsymbol{\varepsilon}\in E} w(\boldsymbol{\varepsilon}) P(\boldsymbol{\varepsilon}) \\ &\leq \sum_{\boldsymbol{\varepsilon}\in E} w(\boldsymbol{\varepsilon}) Q\left(\sqrt{\frac{6}{M^2-1}\gamma_{\text{av}}\delta^2(\boldsymbol{\varepsilon})}\right) \prod_{i=0}^{l-L-1} \frac{M-|\varepsilon_i|}{M} \end{aligned} \tag{4-145}$$

Now let D be the set of all $\delta(\boldsymbol{\varepsilon})$. For each $\delta \in D$, let E_δ be the subset of error events for which $\delta(\boldsymbol{\varepsilon}) = \delta$. Then Equation 4-145 may be expressed as

$$\begin{aligned} P_e &\leq \sum_{\delta\in D} Q\left(\sqrt{\frac{6}{M^2-1}\gamma_{\text{av}}\delta^2}\right) \left[\sum_{\boldsymbol{\varepsilon}\in E} w(\boldsymbol{\varepsilon}) \prod_{i=0}^{l-L-1} \frac{M-|\varepsilon_i|}{M}\right] \\ &\leq \sum_{\delta\in D} K_\delta Q\left(\sqrt{\frac{6}{M^2-1}\gamma_{\text{av}}\delta^2}\right) \end{aligned} \tag{4-146}$$

where

$$K_\delta = \sum_{\boldsymbol{\varepsilon}\in E} w(\boldsymbol{\varepsilon}) \prod_{i=0}^{l-L-1} \frac{M-|\varepsilon_i|}{M} \tag{4-147}$$

The weighting factors $\{K_\delta\}$ may be determined by means of the error state diagram, which is akin to the state diagram of a convolutional encoder. This approach has been illustrated by Forney (1972) and Viterbi and Omura (1979).

In general, however, the use of the error state diagram for computing P_e is tedious. Instead, we may simplify the computation of P_e by focusing on the dominant term in the summation of Equation 4-146. Because of the exponential dependence of each term in the sum, the expression P_e is dominated by the term corresponding to the minimum value of δ, denoted as $\delta_{\min}$. Hence the symbol error probability may be approximated as

$$P_e \approx K_{\delta_{\min}} Q\left(\sqrt{\frac{6}{M^2-1}\gamma_{\text{av}}\delta^2_{\min}}\right) \tag{4-148}$$

where

$$K_{\delta_{\min}} = \sum_{\boldsymbol{\varepsilon}\in E_{\delta_{\min}}} w(\boldsymbol{\varepsilon}) \prod_{i=0}^{l-L-1} \frac{M-|\varepsilon_i|}{M} \tag{4-149}$$

In general, $\delta^2_{\min} \leq 1$. Hence, $10\log\delta^2_{\min}$ represents the loss in SNR due to intersymbol interference.

The minimum value of δ may be determined either from Equation 4-134 or from evaluation of the quadratic form in Equation 4-138 for different error sequences. In the following two examples we use Equation 4-134.

Example 7. Consider a two path channel ($L = 1$) with arbitrary coefficients f_0 and f_1 satisfying the constraint $f_0^2 + f_1^2 = 1$. The channel characteristic is

$$F(z) = f_0 + f_1 z^{-1} \tag{4-150}$$

For an error event of length n,

$$\varepsilon(z) = \varepsilon_0 + \varepsilon_1 z^{-1} + \cdots + \varepsilon_{n-1} z^{-(n-1)},\ n \geq 1 \tag{4-151}$$

The product $\alpha(z) = F(z)\varepsilon(z)$ may be expressed as

$$\alpha(z) = \alpha_0 + \alpha_1 z^{-1} + \cdots + \alpha_n^{-n} \tag{4-152}$$

where $\alpha_0 = \varepsilon_0 f_0$ and $\alpha_n = f_1 \varepsilon_{n-1}$. Since $\varepsilon_0 \neq 0$, $\varepsilon_1 \neq 0$, and

$$\delta^2(\boldsymbol{\varepsilon}) = \sum_{k=0}^{n} \alpha_k^2 \tag{4-153}$$

it follows that

$$\delta_{\min}^2 \geq f_0^2 + f_0^2 = 1$$

Indeed, $\delta_{\min}^2 = 1$ when a single error occurs, *i.e.*, $\varepsilon(z) = \varepsilon_0$. Thus, we conclude that there is no loss in SNR in maximum-likelihood sequence estimation of the information symbols when the channel dispersion has length 2.

Example 8. The controlled intersymbol interference in a partial-response signal may be viewed as having been generated by a time-dispersive channel. Thus, the intersymbol interference from a duobinary pulse may be represented by the (normalized) channel characteristic

$$F(z) = \sqrt{\frac{1}{2}} + \sqrt{\frac{1}{2}} z^{-1} \tag{4-154}$$

Similarly, the representation for a modified duobinary pulse is

$$F(z) = \sqrt{\frac{1}{2}} - \sqrt{\frac{1}{2}} z^{-2} \tag{4-155}$$

The minimum distance $\delta_{\min}^2 = 1$ for any error event of the form

$$\varepsilon(z) = \pm(1 - z^{-1} - z^{-2} \cdots - z^{-(n-1)}),\ n \geq 1 \tag{4-156}$$

for the channel given by Equation 4-154, since

$$\alpha(z) = \pm\sqrt{\frac{1}{2}} \mp \sqrt{\frac{1}{2}} z^{-n}$$

Similarly, when

$$\varepsilon(z) = \pm(1 + z^{-2} + z^{-4} + \cdots + z^{-2(n-1)}),\ n \geq 1 \tag{4-157}$$

$\delta^2_{\min} = 1$ for the channel given by Equation 4-155 since

$$\alpha(z) = \pm\sqrt{\frac{1}{2}} \mp \sqrt{\frac{1}{2}} z^{-2n}$$

Hence the MLSE of these two partial-response signals result in no loss in SNR. In contrast, the suboptimum symbol-by-symbol detection described previously resulted in a 2.1-dB loss.

The constant $K_{\delta_{\min}}$ is easily evaluated for these two signals. With precoding, the number of output symbol errors (Hamming weight) associated with the error events in Equations 4-156 and 4-157 is two. Hence,

$$K_{\delta_{\min}} = 2\sum_{n=1}^{\infty}\left(\frac{M-1}{M}\right)^n = 2(M-1) \tag{4-158}$$

On the other hand, without precoding, these error events result in n symbol errors, and, hence,

$$K_{\delta_{\min}} = 2\sum_{n=1}^{\infty} n\left(\frac{M-1}{M}\right)^n = 2M(M-1) \tag{4-159}$$

As a final exercise, we consider the evaluation of $\delta^2_{\min}$ from the quadratic form in Equation 4-138. The matrix $\boldsymbol{A}$ of the quadratic form is positive-definite; hence, all its eigenvalues are positive. If $\{\mu_k(\boldsymbol{\varepsilon})\}$ are the eigenvalues and $\boldsymbol{v}_k(\boldsymbol{\varepsilon})$ are the corresponding orthonormal eigenvectors of $\boldsymbol{A}$ for an error event $\boldsymbol{\varepsilon}$, then the quadratic form in Equation 4-138 can be expressed as

$$\delta^2(\boldsymbol{\varepsilon}) = \sum_{k=1}^{L+1} \mu_k(\boldsymbol{\varepsilon})[\boldsymbol{f}^T \boldsymbol{v}_k(\boldsymbol{\varepsilon})]^2 \tag{4-160}$$

In other words, $\delta^2(\boldsymbol{\varepsilon})$ is expressed as a linear combination of the squared projections of the channel vector $\boldsymbol{f}$ onto the eigenvectors of $\boldsymbol{A}$. Each squared projection of the sum is weighted by the corresponding eigenvalue $\mu_k(\boldsymbol{\varepsilon})$, $k = 1, 2, \cdots, L+1$. Then

$$\delta^2_{\min} = \min_{\boldsymbol{\varepsilon}} \delta^2(\boldsymbol{\varepsilon}) \tag{4-161}$$

It is interesting to note that the worst channel characteristic of a given length $L+1$ can be obtained by finding the eigenvector corresponding to the minimum eigenvalue. Thus, if $\mu_{\min}(\boldsymbol{\varepsilon})$ is the minimum eigenvalue for a given error event $\boldsymbol{\varepsilon}$ and $\boldsymbol{v}_{\min}(\boldsymbol{\varepsilon})$ is the corresponding eigenvector, then

$$\begin{aligned} \mu_{\min} &= \min_{\boldsymbol{\varepsilon}} \mu_{\min}(\boldsymbol{\varepsilon}) \\ \boldsymbol{f} &= \min_{\boldsymbol{\varepsilon}} \boldsymbol{v}_{\min}(\boldsymbol{\varepsilon}) \end{aligned}$$

and

$$\delta^2_{\min} = \mu_{\min}$$

Example 9. Let us determine the worst time-dispersive channel of length 3 ($L = 2$) by finding the minimum eigenvalue of $\boldsymbol{A}$ for different error events. Thus,

$$F(z) = f_0 + f_1 z^{-1} + f_2 z^{-2}$$

where f_0, f_1, and f_2 are the components of the eigenvector of $\boldsymbol{A}$ corresponding to the minimum eigenvalue. An error event of the form

$$\varepsilon(z) = 1 - z^{-1}$$

results in a matrix

$$\boldsymbol{A} = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix}$$

which has the eigenvalues $\mu_1 = 2$, $\mu_2 = 2 + \sqrt{2}$, $\mu_3 = 2 - \sqrt{2}$. The eigenvector corresponding to μ_3 is

$$\boldsymbol{v}_3^T = \begin{bmatrix} \frac{1}{2} & \sqrt{\frac{1}{2}} & \frac{1}{2} \end{bmatrix} \tag{4-162}$$

We may also consider the dual error event

$$\boldsymbol{\varepsilon}(z) = 1 + z^{-1}$$

which results the matrix

$$\boldsymbol{A} = \begin{bmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{bmatrix}$$

This matrix has eigenvalues identical to those of the one for $\varepsilon(z) = 1 - z^{-1}$. The corresponding eigenvector for $\mu_3 = 2 - \sqrt{2}$ is

$$\boldsymbol{v}_3^T = \begin{bmatrix} -\frac{1}{2} & \sqrt{\frac{1}{2}} & -\frac{1}{2} \end{bmatrix} \tag{4-163}$$

Any other error events lead to larger values for $\mu_{\min}$. Hence, $\mu_{\min} = 2 - \sqrt{2}$ and the worst-case channel is eigher

$$\begin{bmatrix} \frac{1}{2} & \sqrt{\frac{1}{2}} & \frac{1}{2} \end{bmatrix} \text{ or } \begin{bmatrix} -\frac{1}{2} & \sqrt{\frac{1}{2}} & -\frac{1}{2} \end{bmatrix}$$

The loss in SNR from the channel is

$$-10 \log \delta^2_{\min} = -10 \log \mu_{\min} = 2.3 \text{ dB}$$

Repetitions of the above computation for channels with $L = 3, 4$, and 5 yield the results given in Table 4.3.

Table 4.3: Maximum Performance Loss and Corresponding Channel Characteristics

Channel length $L+1$	Performance loss $-10 \log \delta_{\min}^2$ dB	Minimum-distance channel
3	2.3	0.50, 0.71, 0.50
4	4.2	0.38, 0.60, 0.60, 0.38
5	5.7	0.29, 0.50, 0.58, 0.50, 0.29
6	7.0	0.23, 0.42, 0.52, 0.52, 0.42, 0.23

4.4 Linear Equalization

The MLSE for a channel with ISI has a computational complexity that grows exponentially with the length of the channel time dispersion. If the size of the symbol alphabet is M and the number of interfering symbols contributing to ISI is L, the Viterbi algorithm computes M^{L+1} metrics for each new received symbol. In most channels of practical interest, such a large computational complexity is prohibitively expensive to implement.

In this and the following sections, we describe suboptimum channel equalization approaches to compensate for the ISI. One approach employs a linear transversal filter, which is described in this section. This filter structure has a computational complexity that is a linear function of the channel dispersion length L.

The linear filter most often used for equalization is the transversal filter shown in Figure 4-20. Its input is the sequence $\{v_k\}$ given in Equation 4-110 and its output in the estimate of the information sequence $\{I_k\}$. The estimate of the kth symbol may be expressed as

$$\hat{I}_k = \sum_{j=-K}^{K} c_j v_{k-j} \tag{4-164}$$

where $\{c_j\}$ are the $2K+1$ complex-valued tap weight coefficients of the filter. The estimate $\hat{I}_k$ is quantized to the nearest (in distance) information symbol to form the decision $\tilde{I}_k$. If $\tilde{I}_k$ is not identical to the transmitted information symbol I_k, an error has been made.

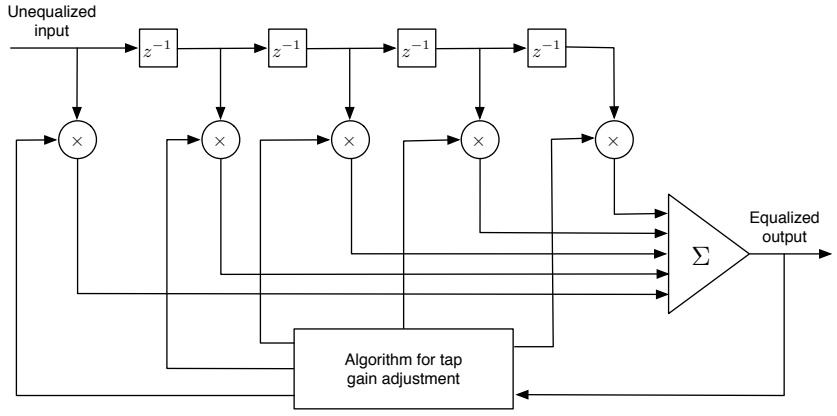

Figure 4-20: Linear transversal filter.

Considerable research has been performed on the criterion for optimizing the filter coefficients $\{c_k\}$. Since the most meaningful measure of performance for a digital communication system is the average probability of error, it is desirable to choose the coefficients to minimize this performance index. However, the probability of error is a highly non-linear function of $\{c_j\}$. Consequently, the probability of error as a performance index for optimizing the tap weight coefficients of the equalizer is computationally complex.

Two criteria have found widespread use in optimizing the equalizer coefficients $\{c_j\}$. One is the peak distortion criterion and the other is the mean-square-error criterion.

4.4.1 Peak Distortion Criterion

The peak distortion is simply defined as the worst-case intersymbol interference at the output of the equalizer. The minimization of this performance index is called the *peak distortion criterion.* First we consider the minimization of the peak distortion assuming that the equalizer has an infinite number of taps. Then we shall discuss the case in which the transversal equalizer spans a finite time duration.

We observe that the cascade of the discrete-time linear filter model having an impulse response $\{f_n\}$ and an equalizer having an impulse response $\{c_n\}$ can be represented by a single equivalent filter having the impulse response

$$q_n = \sum_{j=-\infty}^{\infty} c_j f_{n-j} \tag{4-165}$$

That is, $\{q_n\}$ is simply the convolution of $\{c_n\}$ and $\{f_n\}$. The equalizer is assumed to have an infinite number of taps. Its output at the kth sampling instant can be expressed in the form

$$\hat{I}_k = q_0 I_k + \sum_{n \neq k} I_n q_{k-n} + \sum_{j=-\infty}^{\infty} c_j \eta_{k-j} \tag{4-166}$$

The first term in Equation 4-166 represents a scaled version of the desired symbol. For convenience, we normalize q_0 to unity. The second term is the intersymbol interference. The peak value of this interference, which is called the *peak distortion*, is

$$\begin{aligned} \mathcal{D}(\boldsymbol{c}) &= \sum_{n=-\infty | n \neq 0}^{\infty} |q_n| \\ &= \sum_{n=-\infty | n \neq 0}^{\infty} \left| \sum_{j=-\infty}^{\infty} c_j f_{n-j} \right| \end{aligned} \tag{4-167}$$

Thus, $\mathcal{D}(\boldsymbol{c})$ is a function of the equalizer tap weights.

With an equalizer having an infinite number of taps, it is possible to select the tap weights so that $\mathcal{D}(\boldsymbol{c}) = 0$, *i.e.*, $q_n = 0$ for all n except $n = 0$. That is, the intersymbol interference can be completely eliminated. The values of the tap weights for accomplishing this goal are determined from the condition

$$q_n = \sum_{j=-\infty}^{\infty} c_j f_{n-j} = \begin{cases} 1 & ; \quad n = 0 \\ 0 & ; \quad n \neq 0 \end{cases} \tag{4-168}$$

By taking the z transform of Equation 4-168, we obtain

$$Q(z) = C(z)F(z) = 1 \tag{4-169}$$

or, simply,

$$C(z) = \frac{1}{F(z)} \tag{4-170}$$

where $C(z)$ denotes the z transform of the $\{c_j\}$. Note that the equalizer, with transfer function $C(z)$, is simply the inverse filter to the linear filter model $F(z)$. In other words, complete

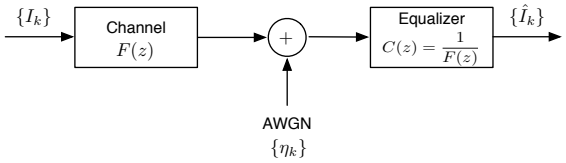

Figure 4-21: Block diagram of channel with zero-forcing equalizer.

elimination of the intersymbol interference requires the use of an inverse filter to $F(z)$. We call such a filter a *zero-forcing filter*. Figure 4-21 illustrates in block diagram the equivalent discrete-time channel and equalizer.

The cascade of the noise-whitening filter having the transfer function $1/F^*(l/z^*)$ and the zero-forcing equalizer having the transfer function $1/F(z)$ results in an equivalent zero-forcing equalizer having the transfer function

$$C'(z) = \frac{1}{F(z)F^*(1/z^*)} = \frac{1}{X(z)} \tag{4-171}$$

as shown in Figure 4-22. This combined filter has as its input the sequence $\{y_k\}$ of samples from the matched filter, given by Equation 4-104. Its output consists of the desired symbols corrupted only by additive zero-mean Gaussian noise. The impulse response of the combined filter is

$$\begin{aligned} c'_k &= \frac{1}{2\pi j} \oint_{-\infty}^{\infty} C'(z) z^{k-1} dz \\ &= \frac{1}{2\pi j} \oint_{-\infty}^{\infty} \frac{z^{k-1}}{X(z)} dz \end{aligned} \tag{4-172}$$

where the integration is performed on a closed contour that lies within the region of convergence of $C'(z)$. Since $X(z)$ is a polynomial with $2L$ roots $(\rho_1, \rho_2, \cdots, \rho_L, 1/\rho_1^*, 1/\rho_2^*, \cdots, 1/\rho_L^*)$, it follows that $C'(z)$ must converge in an annular region in the z plane that includes the unit circle ($z = e^{j\theta}$). Consequently, the closed contour in the integral can be the unit circle.

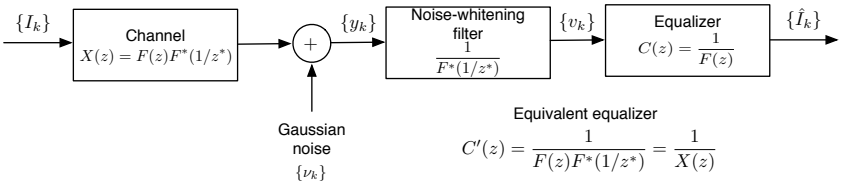

Figure 4-22: Block diagram of channel with equivalent zero-forcing equalizer.

The performance of the infinite-tap equalizer that completely eliminates the inter-symbol interference can be expressed in terms of the SNR at its output. For mathematical convenience, we normalize the received signal energy to unity[6]. This implies that $q_0 = 1$ and that the expected value of $|I_k|^2$ is also unity. Then the SNR is simply the reciprocal of the noise variance σ_n^2 at the output of the equalizer[7].

[6]This normalization is used throughout this chapter for mathematical convenience.

[7]If desired, one can multiply this normalized SNR at the output of the equalizer by the signal energy.

The value of σ_n^2 can be simply determined by observing that the noise sequence $\{v_k\}$ at the input to the equivalent zero-forcing equalizer $C'(z)$ has zero-mean and a power spectral density

$$\mathcal{S}_{v,v}(\omega) = N_0 X(e^{j\omega T}), \quad |\omega| \leq \frac{\pi}{T} \tag{4-173}$$

where $X(e^{j\omega t})$ is obtained from $X(z)$ by the substitution $z = e^{j\omega T}$. Since $C'(z) = 1/X(z)$, it follows that the noise sequence at the output of the equalizer has a power spectral density

$$\mathcal{S}_{n,n}(\omega) = \frac{N_0}{X(e^{j\omega T})}, \quad |\omega| \leq \frac{\pi}{T} \tag{4-174}$$

Consequently, the variance of the noise variable at the output of the equalizer is

$$\begin{aligned} \sigma_n^2 &= \frac{T}{2\pi} \int_{-\pi/T}^{\pi/T} \mathcal{S}_{nn}(\omega) d\omega \\ &= \frac{TN_0}{2\pi} \int_{-\pi/T}^{\pi/T} \frac{d\omega}{X(e^{j\omega T})} \end{aligned} \tag{4-175}$$

and the SNR for the zero-forcing equalizer is

$$\gamma_\infty = \frac{1}{\sigma_n^2} = \left[\frac{TN_0}{2\pi} \int_{-\pi/T}^{\pi/T} \frac{d\omega}{X(e^{j\omega T})} \right]^{-1} \tag{4-176}$$

where the subscript on γ indicates that the equalizer has an infinite number of taps.

The spectral characteristics $X(e^{j\omega T})$ corresponding to the Fourier transform of the sampled sequence $\{x_n\}$ has an interesting relationship to the analog filter $H(\omega)$ used at the receiver. Since

$$x_k = \int_{-\infty}^{\infty} h^*(t) h(t + kT) dt$$

use of Parseval's theorem yields

$$x_k = \frac{1}{2\pi} \int_{-\infty}^{\infty} |H(\omega)|^2 e^{j\omega T} dt \tag{4-177}$$

where $H(\omega)$ is the Fourier transform of $h(t)$. But the integral in Equation 4-177 can be expressed in the form

$$x_k = \frac{1}{2\pi} \int_{-\pi/T}^{\pi/T} \left[\sum_{n=-\infty}^{\infty} \left| H\left(\omega + \frac{2\pi n}{T}\right) \right|^2 \right] e^{j\omega kT} d\omega \tag{4-178}$$

Now, the Fourier transform of $\{x_k\}$ is

$$X(e^{j\omega T}) = \sum_{k=-\infty}^{\infty} x_k e^{-j\omega kT} \tag{4-179}$$

and the inverse transform yields

$$x_k = \frac{T}{2\pi} \int_{-\pi/T}^{\pi/T} X(e^{j\omega T}) e^{j\omega kT} d\omega \tag{4-180}$$

From a comparison of Equations 4-178 and 4-180, we obtain the desired relationship beween $X(e^{j\omega T})$ and $H(\omega)$. That is,

$$X(e^{j\omega T}) = \frac{1}{T}\sum_{n=-\infty}^{\infty}\left|H\left(\omega+\frac{2\pi n}{T}\right)\right|^2, \quad |\omega| \le \frac{\pi}{T} \tag{4-181}$$

where the right-hand side of Equation 4-181 is called the *folded spectrum* of $|H(\omega)|^2$. We also observe that $|H(\omega)|^2 = X(\omega)$, where $X(\omega)$ is the Fourier transform of the waveform $x(t)$ and $x(t)$ is the response of the matched filter to the input pulse $h(t)$. Therefore the right-hand side of Equation 4-181 can also be expressed in terms of $X(\omega)$.

Substitution for $X(e^{j\omega T})$ in Equation 4-176 using the result in Equation 4-181 yields the desired expression for the SNR in the form

$$\gamma_\infty = \left[\frac{T^2 N_0}{2\pi}\int_{-\pi/T}^{\pi/T}\frac{d\omega}{\sum_{n=-\infty}^{\infty}\left|H\left(\omega+\frac{2\pi n}{T}\right)\right|^2}\right]^{-1} \tag{4-182}$$

We observe that if the folded spectral characteristic of $H(\omega)$ possesses any zeros, the integrand becomes infinite and the SNR goes to zero. In other words, the performance of the equalizer is poor whenever the folded spectral characteristic possesses nulls or takes on small values. This behavior occurs primarily because the equalizer, in eliminating the intersymbol interference, enhances the additive noise. For example, if the channel contains a spectral null in its frequency response, the linear zero-forcing equalizer attempts to compensate for this by introducing an infinite gain at that frequency. But this compensates for the channel distortion at the expense of enhancing the additive noise. On the other hand, an ideal channel coupled with an appropriate signal design that results in no intersymbol interference will have a folded spectrum that satisfies the condition

$$\sum_{n=-\infty}^{\infty}\left|H\left(\omega+\frac{2\pi n}{T}\right)\right|^2 = T, \quad |\omega| \le \frac{\pi}{T} \tag{4-183}$$

In this case, the SNR achieves its maximum value, namely,

$$\gamma_\infty = \frac{1}{N_0} \tag{4-184}$$

Finite-length equalizer

Let us now turn our attention to an equalizer having $2K+1$ taps. Since $c_j = 0$ for $|j| > K$, the convolution of $\{f_n\}$ with $\{c_n\}$ is zero outside the range $-K \le n \le K+L-1$. That is, $q_n = 0$ for $n < -K$ and $n > K+L-1$. With q_0 normalized to unity, the peak distortion is

$$\mathcal{D}(\boldsymbol{c}) = \sum_{n=-K|n\neq 0}^{K+L-1}|q_n| = \sum_{n=-K|n\neq 0}^{K+L-1}\left|\sum_j c_j f_{n-j}\right| \tag{4-185}$$

Although the equalizer has $2K+1$ adjustable parameters, there are $2K+L$ nonzero values in the response $\{q_n\}$. Therefore, it is generally impossible to completely eliminate the intersymbol interference at the output of the equalizer. There is always some residual interference when the optimum coefficients are used. The problem is to minimize $\mathcal{D}(\boldsymbol{c})$ with respect to the coefficients $\{c_j\}$.

The peak distortion given by Equation 4-185 has been shown by Lucky (1965) to be a convex function of the coefficients $\{c_j\}$. That is, it possesses a global minimum and no local minima. Its minimization can be carried out numerically using, for example, the method of steepest descent. Little more can be said for the general solution to this minimization problem. However, for one special but important case, the solution for the minimization of $\mathcal{D}(\boldsymbol{c})$ is known. This is the case in which the distortion at the input to the equalizer, defined as

$$D_0 = \frac{1}{|f_0|} \sum_{n=1}^{L} |f_n| \tag{4-186}$$

is less than unity. This condition is equivalent to having the eye open prior to equalization. That is, the intersymbol interference is not severe enough to close the eye. Under this condition, the peak distortion $\mathcal{D}(\boldsymbol{c})$ is minimized by selecting the equalizer coefficients to force $q_n = 0$ for $1 \leq |n| \leq K$ and $q_0 = 1$. In other words, the general solution to the minimization of $\mathcal{D}(\boldsymbol{c})$, when $D_0 < 1$, is the zero-forcing solution for $\{q_n\}$ in the range $1 \leq |n| \leq K$. However, the values of $\{q_n\}$ for $K + 1 \leq n \leq K + L - 1$ are nonzero, in general. These nonzero values constitute the residual intersymbol interference at the output of the equalizer.

4.4.2 Mean-Square-Error (MSE) Criterion

In the MSE criterion, the tap weight coefficients $\{c_j\}$ of the equalizer are adjusted to minimize the mean square value of the error

$$\varepsilon_k = I_k - \hat{I}_k \tag{4-187}$$

where I_k is the information symbol transmitted in the kth signaling interval and $\hat{I}_k$ is the estimate of that symbol at the output of the equalizer, defined in Equation 4-164. When the information symbols $\{I_k\}$ are complex-valued, the performance index for the MSE criterion, denoted by J, is defined as

$$J = E\left\{|\varepsilon_k|^2\right\} = E\left\{|I_k - \hat{I}_k|^2\right\} \tag{4-188}$$

On the other hand, when the information symbols are real-valued, the performance index is simply the square of the real part of ε_k. In either case, J is a quadratic function of the equalizer coefficients $\{c_j\}$. In the following discussion, we consider the minimization of the complex-valued form given in Equation 4-188.

Infinite-length equalizer

First, we shall derive the tap weight coefficients that minimize J when the equalizer has an infinite number of taps. In this case, the estimate $\hat{I}_k$ is expressed as

$$\hat{I}_k = \sum_{j=-\infty}^{\infty} c_j v_{k-j} \tag{4-189}$$

Substitution of Equation 4-189 into the expression for J given in Equation 4-188 and expansion of the result yields a quadratic function of the coefficients $\{c_j\}$. This function can be easily minimized with respect to the $\{c_j\}$ to yield a set (infinite in number) of linear equations for the $\{c_j\}$. Alternatively, the set of linear equations can be obtained by invoking the orthogonality

principle in mean square estimation. That is, we select the coefficients $\{c_j\}$ to render the error ε_k orthogonal to the signal sequence $\{v_{k-l}^*\}$ for $-\infty < l < \infty$. Thus,

$$E\left\{\varepsilon_k v_{k-l}^*\right\} = 0, \quad -\infty < l < \infty \tag{4-190}$$

Substitution for ε_k in Equation 4-190 yields

$$E\left\{\left(I_k - \sum_{j=-\infty}^{\infty} c_j v_{k-j}\right) v_{k-l}^*\right\} = 0$$

or, equivalently,

$$\sum_{j=-\infty}^{\infty} c_j E\left\{v_{k-j} v_{k-l}^*\right\} = E\left\{I_k v_{k-l}^*\right\}, \quad -\infty < l < \infty \tag{4-191}$$

To evaluate the moments in Equation 4-191, we use the expression for v_k given in Equation 4-110. Thus, we obtain

$$\begin{aligned} E\left\{v_{k-j} v_{k-l}^*\right\} &= \sum_{n=0}^{L} f_n^* f_{n+l-j} + N_0 \delta(l-j) \\ &= \begin{cases} x_{l-j} + N_0 \delta(l-j) & |l-j| \le L \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{4-192}$$

and

$$E\left\{I_k v_{k-l}^*\right\} = \begin{cases} f_{-l}^* & -L \le l \le 0 \\ 0 & \text{otherwise} \end{cases} \tag{4-193}$$

Now, if we substitute Equations 4-192 and 4-193 into Equation 4-191 and take the z transform of both sides of the resulting equation, we obtain

$$C(z)\left[F(z)F^*(1/z^*) + N_0\right] = F^*(1/z^*) \tag{4-194}$$

Therefore, the transfer function of the equalizer based on the MSE criterion is

$$C(z) = \frac{F^*(1/z^*)}{F(z)F^*(1/z^*) + N_0} \tag{4-195}$$

When the noise-whitening filter is incorporated into $C(z)$, we obtain an equivalent equalizer having the transfer function

$$\begin{aligned} C'(z) &= \frac{1}{F(z)F^*(1/z^*) + N_0} \\ &= \frac{1}{X(z) + N_0} \end{aligned} \tag{4-196}$$

We observe that the only difference between this expression for $C'(z)$ and the one based on the peak distortion criterion is the noise spectral density factor N_0 that appears in Equation 4-196. When N_0 is very small in comparison with the signal, the coefficients that minimize the peak distortion $\mathcal{D}(\boldsymbol{c})$ are approximately equal to the coefficients that minimize the MSE performance index J. That is, in the limit as $N_0 \to 0$, the two criteria yield the same solution for

the tap weights. Consequently, when $N_0 = 0$, the minimization of the MSE results in complete elimination of the intersymbol interference. On the other hand, that is not the case when $N_0 \neq 0$. In general, when $N_0 \neq 0$, there is both residual intersymbol interference and additive noise at the output of the equalizer.

A measure of the residual intersymbol interference and additive noise is obtained by evaluating the minimum value of J, denoted by $J_{\min}$ when the transfer function of the equalizer given by Equation 4-195. Since $J = E\left\{|\varepsilon_k|^2\right\} = E\left\{\varepsilon_k I_k^*\right\} - E\left\{\varepsilon_k \hat{I}_k^*\right\}$, and since $E\left\{\varepsilon_k \hat{I}_k^*\right\} = 0$ by virtue of the orthogonality conditions given in Equation 4-190, it follows that

$$\begin{aligned} J_{\min} &= E\left\{\varepsilon_k I_k^*\right\} \\ &= E\left\{|I_k|^2\right\} - \sum_{j=-\infty}^{\infty} c_j E\left\{v_{k-j} I_k^*\right\} \\ &= 1 - \sum_{j=-\infty}^{\infty} c_j f_{-j} \end{aligned} \tag{4-197}$$

This particular form for $J_{\min}$ is not very informative. More insight on the performance of the equalizer as a function of the channel characteristics is obtained when the summation in Equation 4-197 is transformed into the frequency domain. This can be accomplished by first noting that the summation in Equation 4-197 is the convolution of $\{c_j\}$ with $\{f_j\}$, evaluated at a shift of zero. Thus. if $\{b_k\}$ denotes the convolution of these two sequences, the summation in Equation 4-197 is simply equal to b_0. Since the z transform of the sequence $\{b_k\}$ is

$$\begin{aligned} B(z) &= C(z)F(z) \\ &= \frac{F(z)F^*(1/z^*)}{F(z)F^*(1/z^*) + N_0} \\ &= \frac{X(z)}{X(z) + N_0} \end{aligned} \tag{4-198}$$

the term b_0 is

$$\begin{aligned} b_0 &= \frac{1}{2\pi j} \oint \frac{B(z)}{z} dz \\ &= \frac{1}{2\pi j} \oint \frac{X(z)}{z[X(z) + N_0]} dz \end{aligned} \tag{4-199}$$

The contour integral in Equation 4-199 can be transformed into an equivalent line integral by the change of variable $z = e^{j\omega T}$. The result of this change of variable is

$$b_0 = \frac{T}{2\pi} \int_{-\pi/T}^{\pi/T} \frac{X(e^{j\omega T})}{X(e^{j\omega T}) + N_0} d\omega \tag{4-200}$$

Finally, substitution of the result in Equation 4-200 for the summation in Equation 4-197 yields the desired expression for the minimum MSE in the form

$$\begin{aligned} J_{\min} &= 1 - \frac{T}{2\pi} \int_{-\pi/T}^{\pi/T} \frac{X(e^{j\omega T})}{X(e^{j\omega T}) + N_0} d\omega \\ &= \frac{T}{2\pi} \int_{-\pi/T}^{\pi/T} \frac{N_0}{X(e^{j\omega T}) + N_0} d\omega \end{aligned}$$

$$= \frac{T}{2\pi}\int_{-\pi/T}^{\pi/T} \frac{N_0}{T^{-1}\sum_{n=-\infty}^{\infty}|H(\omega+2\pi n/T)|^2 + N_0} d\omega \tag{4-201}$$

In the absence of intersymbol interference, $X(e^{j\omega T}) = 1$ and, hence,

$$J_{\min} = \frac{N_0}{1+N_0} \tag{4-202}$$

We observe that $0 \leq J_{\min} \leq 1$. Furthermore, the relationship between the output (normalized by the signal energy) SNR γ_∞ and $J_{\min}$ must be

$$\gamma_\infty = \frac{1-J_{\min}}{J_{\min}} \tag{4-203}$$

More importantly, this relation between γ_∞ and $J_{\min}$ also holds when there is residual intersymbol interference in addition to the noise.

Finite-length equalizer

Let us now turn our attention to the case in which the transversal equalizer spans a finite time duration. The output of the equalizer in the kth signaling interval is

$$\hat{I}_k = \sum_{j=-K}^{K} c_j v_{k-j} \tag{4-204}$$

The MSE for the equalizer having $2K+1$ taps, denoted by $J(K)$, is

$$J(K) = E\left\{|I_k - \hat{I}_k|^2\right\} = E\left\{\left|I_k - \sum_{j=-K}^{K} c_j v_{k-j}\right|^2\right\} \tag{4-205}$$

Minimization of $J(K)$ with respect to the tap weights $\{c_j\}$ or, equivalently, forcing the error $\varepsilon_k = I_k - \hat{I}_k$ to be orthogonal to the signal samples v_{j-l}^*, $|l| \leq K$ yields the following set of simultaneous equations:

$$\sum_{j=-K}^{K} c_j \Gamma_{lj} = \xi_l, \quad l = -K, \cdots, -1, 0, 1, \cdots, K \tag{4-206}$$

where

$$\Gamma_{lj} = \begin{cases} x_{l-j} + N_0\delta(l-j) & |l-j| \leq L \\ 0 & \text{otherwise} \end{cases} \tag{4-207}$$

and

$$\xi_l = \begin{cases} f_{-l}^* & -L \leq l \leq 0 \\ 0 & \text{otherwise} \end{cases} \tag{4-208}$$

It is convenient to express the set of linear equations in matrix form. Thus,

$$\boldsymbol{\Gamma C} = \boldsymbol{\xi} \tag{4-209}$$

where $\boldsymbol{C}$ denotes the column vector of $2K+1$ tap weight coefficients, $\boldsymbol{\Gamma}$ denotes the $(2K+1) \times (2K+1)$ Hermitian covariance matrix with elements Γ_{ij} and $\boldsymbol{\xi}$ is a $(2K+1)$-dimensional column vector with elements ξ_i. The solution of Equation **??** is

$$\boldsymbol{C}_{\text{opt}} = \boldsymbol{\Gamma}^{-1}\boldsymbol{\xi} \tag{4-210}$$

Thus, the solution for $\boldsymbol{C}_{\text{opt}}$ involves inverting the matrix $\boldsymbol{\Gamma}$. The optimum tap weight coefficients given by Equation 4-210 minimize the performance index $J(K)$, with the result that the minimum value of $J(K)$ is

$$\begin{aligned} J_{\min}(K) &= 1 - \sum_{j=-K}^{0} c_j f_{-j} \\ &= 1 - \boldsymbol{\xi}^H \boldsymbol{\Gamma}^{-1} \boldsymbol{\xi} \end{aligned} \tag{4-211}$$

where H represents the conjugate transpose. $J_{\min}(K)$ may be used in Equation 4-203 to compute the output SNR for the linear equalizer with $2K+1$ tap coefficients.

4.5 Decision-Feedback Equalization

In Section 4.3.2 we developed an equivalent discrete-time model of the channel with ISI and additive noise, as shown in Figure 4-16. We observed that the additive Gaussian noise in this model is colored. Then we simplified this model by inserting a noise-whitening filter prior to the equalizer, so that the resulting discrete-time model of the channel has AWGN as shown in Figure 4-17. To recover the information sequence that is corrupted by ISI. we considered two types of equalization methods, one based on the MLSE criterion that is efficiently implemented by the Viterbi algorithm and the other employed a linear transversal filter. We recall that the MLSE is the optimum detector in the sense that it minimizes the probability of a sequence error while the linear equalizer is suboptimum.

In this section, we consider a nonlinear type of channel equalizer for mitigating the ISI, which is also suboptimum, but whose performance is generally better than that of the linear equalizer. The nonlinear equalizer consists of two filters, a feedforward filter and a feedback filter, arranged as shown in Figure 4-29, and it is called a *decision-feedback equalizer* (DFE). The input to the feedforward filter is the received signal sequence. The feedback filter has as its input the sequence of decisions on previously detected symbols. Functionally, the feedback filter is used to remove that part of the ISI from the present estimated symbol caused by previously detected symbols. Since the detector feeds hard decisions to the feedback filter, the DFE is nonlinear.

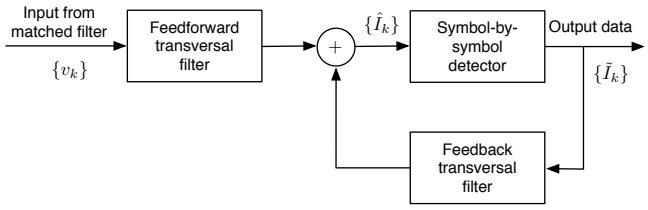

Figure 4-23: Structure of decision-feedback equalizer.

In the case where the feedforward and feedback filters have infinite-duration impulse responses, Price (1972) showed that the optimum feedforward filter in a zero-forcing DFE is the noise-whitening filter with system function $1/F^*(1/z^*)$. Hence, in the zero-forcing DFE, the feedforward filter whitens the additive noise and results in an equivalent discrete-time channel having the system function $F(z)$.

In our treatment, we focus on finite-duration impulse response filters and apply the MSE criterion to optimize their coefficients.

4.5.1 Coefficient Optimization

From the description given above. it follows that the equalizer output can be expressed as

$$\hat{I}_k = \sum_{j=-K_1}^{0} c_j v_{k-j} + \sum_{j=1}^{K_2} c_j \tilde{I}_{k-j} \tag{4-212}$$

where $\hat{I}_k$ is an estimate of the kth information symbol, $\{c_j\}$ are the tap coefficients of the filter, and $\{\tilde{I}_{k-1}, \cdots, \tilde{I}_{k-K_2}\}$ are previously detected symbols. The equalizer is assumed to have (K_1+1) taps in its feedforward section and K_2 in its feedback section.

Both the peak distortion criterion and the MSE criterion result in a mathematically tractable optimization of the equalizer coefficients, as can be concluded from the papers by George *et al.* (1971), Price (1972), Salz (1973), and Proakis (1975). Since the MSE criterion is more prevalent in practice, we focus our attention on it. Based on the assumption that previously detected symbols in the feedback filter are correct, the minimization of MSE

$$J(K_1, K_2) = E\left\{\left|I_k - \hat{I}_k\right|^2\right\} \tag{4-213}$$

leads to the following set of linear equations for the coefficients of the feedforward filter:

$$\sum_{j=-K_1}^{0} \psi_{lj} c_j = f_{-l}^*, \quad l = -K_1, \cdots, -1, 0 \tag{4-214}$$

where

$$\psi_{lj} = \sum_{m=0}^{-l} f_m^* f_{m+l-j} + N_0 \delta(l-j), \quad l, j = -K_1, \cdots, -1, 0 \tag{4-215}$$

The coefficients of the feedback filter of the equalizer are given in terms of the coefficients of the feedforward section by the following expression:

$$c_k = -\sum_{j=-K_1}^{0} c_j f_{k-j}, \quad k = 1, 2, \cdots, K_2 \tag{4-216}$$

The values of the feedback coefficients result in complete elimination of intersymbol interference from previously detected symbols, provided that previous decisions are correct and that $K_2 \geq L$ (see Problem 9.51).

4.5.2 Performance Characteristics of DFE

We now turn our attention to the performance achieved with decision-feedback equalization. The exact evaluation of the performance is complicated to some extent by occasional incorrect decisions made by the detector, which then propagate down the feedback section. In the absence of decision errors, the minimum MSE is given as

$$J_{\min}(K_1) = 1 - \sum_{j=-K_1}^{0} c_j f_{-j} \tag{4-217}$$

By going to the limit ($K_1 \to \infty$) of an infinite number of taps in the feedforward filter, we obtain the smallest achievable MSE, denoted as $J_{\min}$. With some effort $J_{\min}$ can be expressed in terms of the spectral characteristics of the channel and additive noise, as shown by Salz (1973). This more desirable form for $J_{\min}$ is

$$J_{\min} = \exp\left\{\frac{T}{2\pi}\int_{-\pi/T}^{\pi/T} \ln\left[\frac{N_0}{X(e^{j\omega T}) + N_0}\right] d\omega\right\} \tag{4-218}$$

The corresponding output SNR is

$$\gamma_\infty = \frac{1 - J_{\min}}{J_{\min}}$$

$$= -1 + \exp\left\{\frac{T}{2\pi}\int_{-\pi/T}^{\pi/T} \ln\left[\frac{X(e^{j\omega T}) + N_0}{N_0}\right] d\omega\right\} \tag{4-219}$$

We observe again, that in the absence of intersymbol interference, $X(e^{j\omega T}) = 1$, and hence, $J_{\min} = N_0/(1+N_0)$. The corresponding output SNR is $\gamma_\infty = 1/N_0$.

Example 10. It is interesting to compare the value of $J_{\min}$ for the decision-feedback equalizer with the value of $J_{\min}$ obtained with the linear MSE equalizer. For example, let us consider the discrete-time equivalent channel consisting of two taps f_0 and f_1. The minimum MSE for this channel is

$$\begin{aligned} J_{\min} &= \exp\left\{\frac{T}{2\pi}\int_{-\pi/T}^{\pi/T} \ln\left[\frac{N_0}{1+N_0+2|f_0||f_1|\cos(\omega T+\theta)}\right] d\omega\right\} \\ &= N_0 \exp\left[-\frac{T}{2\pi}\int_{-\pi}^{\pi} \ln\left(1+N_0+2|f_0||f_1|\cos\omega\right) d\omega\right] \\ &= \frac{2N_0}{1+N_0+\sqrt{(1+N_0)^2-4|f_0f_1|^2}} \end{aligned} \tag{4-220}$$

Note that $J_{\min}$ is maximized when $|f_0| = |f_1| = \sqrt{\frac{1}{2}}$. Then

$$\begin{aligned} J_{\min} &= \frac{2N_0}{1+N_0+\sqrt{(1+N_0)^2-1}} \\ &\approx 2N_0, \quad N_0 \ll 1 \end{aligned} \tag{4-221}$$

The corresponding output SNR is

$$\gamma_\infty \approx \frac{1}{2N_0}, \quad N_0 \ll 1 \tag{4-222}$$

Therefore, there is a 3-dB degradation in output SNR due to the presence of intersymbol interference. In comparison, the performance loss for the linear equalizer is very severe. Its output SNR as given by Equalizer **??** is $\gamma_\infty \approx (2/N_0)^{1/2}$ for $N_0 \ll 1$.

Example 11. Consider the exponentially decaying channel characteristic of the form

$$f_k = (1-a^2)^{1/2}a^k, \quad k = 0, 1, 2, \cdots \tag{4-223}$$

where $a < 1$. The output SNR of the decision-feedback equalizer is

$$\begin{aligned} \gamma_\infty &= -1 + \exp\left\{\frac{1}{2\pi}\int_{-\pi}^{\pi} \ln\left[\frac{1+a^2+(1-a^2)/N_0-2a\cos\omega}{1+a^2-2a\cos\omega}\right] d\omega\right\} \\ &= -1 + \frac{1}{2N_0}\left\{1-a^2+N_0(1+a^2)+\sqrt{\left[1-a^2+N_0(1+a^2)\right]^2-4a^2N_0^2}\right\} \end{aligned}$$

$$\approx \frac{(1-a^2)\left[1+N_0(1+a^2)/(1-a^2)\right]-N_0}{N_0}$$
$$\approx \frac{1-a^2}{N_0}, \quad N_0 \ll 1 \tag{4-224}$$

Thus, the loss in SNR is $10\log_{10}(1-a^2)$ dB. In comparison, the linear equalizer has loss of $10\log_{10}[(1-a^2)/(1+a^2)]$ dB.

These results illustrate the superiority of the decision-feedback equalizer over the linear equalizer when the effect of decision errors on performance is neglected. It is apparent that a considerable gain in performance can be achieved relative to the linear equalizer by the inclusion of the decision-feedback section, which eliminates the intersymbol interference from previously detected symbols.

One method of assessing the effect of decision errors on the error rate performance of the decision-feedback equalizer is Monte Carlo simulation on a digital computer. For purposes of illustration, we offer the following results for binary PAM signaling through the equivalent discrete-time channel models shown in Figure **??**b and c.

The results of the simulation are displayed in Figure 4-24. First of all, a comparison of these results with those presented in Figure 9**??** leads us to conclude that the decision-feedback equalizer yields a significant improvement in performance relative to the linear equalizer having the same number of taps. Second, these results indicate that there is still a significant degradation in performance of the decision-feedback equalizer due to the residual intersymbol interference, especially on channels with severe distortion such as the one shown in Figure **??**c. Finally, the performance loss due to incorrect decisions being fed back is 2 dB, approximately, for the channel responses under consideration. Additional results on the probability of error for a decision-feedback equalizer with error propagation may be found in the papers by Duttweiler et al. (1974) and Beaulieu (1994).

The structure of the DFE that is analyzed above employs a T-spaced filter for the feedforward section. The optimality of such a structure is based on the assumption that the analog filter preceding the DFE is matched to the channel-corrupted pulse response and its output is sampled at the optimum time instant. In practice, the channel response is not known a priori, so it is not possible to design an ideal matched filter. In view of this difficulty, it is customary in practical applications to use a fractionally spaced feedforward filter. Of course, the feedback filter tap spacing remains at T . The use of the FSE for the feedforward filter eliminates the system sensitivity to a timing error.

Performance comparison with the MLSE

We conclude this subsection on the performance of the DFE by comparing its performance against that of the MLSE. For the two-path channel with $f_0 = f_1 = \sqrt{\frac{1}{2}}$, we have shown that the MLSE suffers no SNR loss while the decision-feedback equalizer suffers a 3-dB loss. On channels with more distortion, the SNR advantage of the MLSE over decision-feedback equalization is even greater. Figure **??** illustrates a comparison of the error rate performance of these two equalization techniques, obtained via Monte Carlo simulation, for binary PAM and the channel characteristics shown in Figure **??**b and c. The error rate curves for the two methods have different slopes; hence

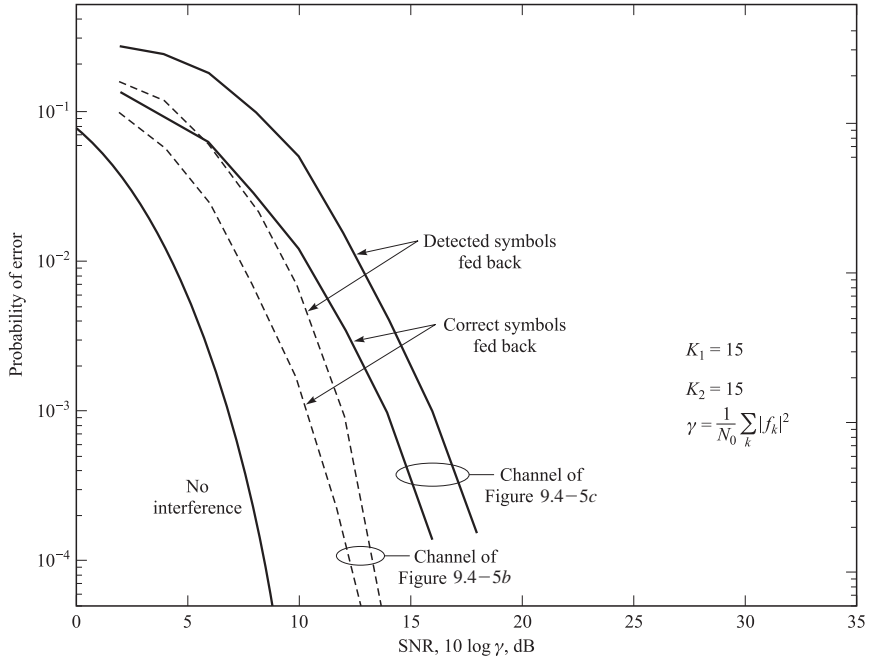

Figure 4-24: Performance of decision-feedback equalizer with and without error propagation.

the difference in SNR increases as the error probability decreases. As a benchmark, the error rate for the AWGN channel with no intersymbol interference is also shown in Figure **??**.

4.5.3 Predictive Decision-Feedback Equalizer

Belfiore and Park (1979) proposed another DFE structure that is equivalent to the one shown in Figure 4-29 under the condition that the feedforward filter has an infinite number of taps. This structure consists of an FSE as a feedforward filter and a linear predictor as a feedback filter, as shown in the configuration given in Figure 4-25. Let us briefly consider the performance characteristics of this equalizer, based on the MSE criterion.

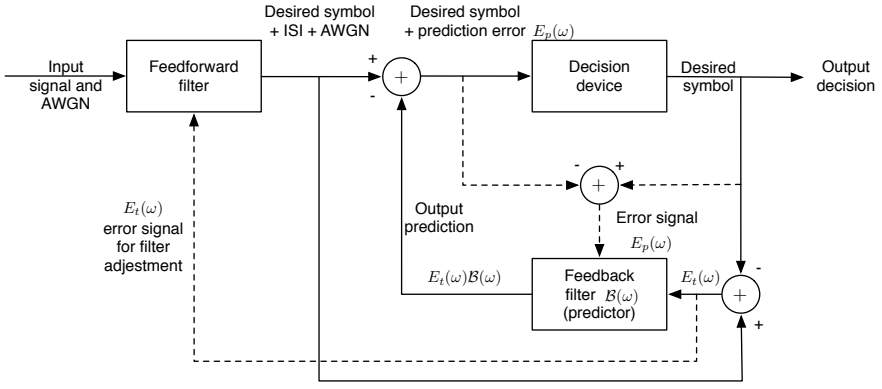

Figure 4-25: Block diagram of predictive DFE.

First of all, the noise at the output of the infinite length feedforward filter has the power spectral density

$$\frac{N_0 X(e^{j\omega T})}{|N_0 + X(e^{j\omega T})|^2}, \quad |\omega| \leq \frac{\pi}{T} \tag{4-225}$$

The residual intersymbol interference has the power spectral density

$$\left|1 - \frac{X(e^{j\omega T})}{N_0 + X(e^{j\omega T})}\right|^2 = \frac{N_0^2}{|N_0 + X(e^{j\omega T})|^2}, \quad |\omega| \leq \frac{\pi}{T} \tag{4-226}$$

The sum of these two spectra represents the power spectral density of the total noise and intersymbol interference at the output of the feedforward filter. Thus, on adding Equations 4-225 and 4-226, we obtain

$$|E_t(\omega)|^2 = \frac{N_0}{N_0 + X(e^{j\omega T})}, \quad |\omega| \leq \frac{\pi}{T} \tag{4-227}$$

As we have observed previously. if $X(e^{j\omega T}) = 1$, the channel is ideal and, hence, it is not possible to reduce the MSE any further. On the other hand, if there is channel distortion, the power in the error sequence at the output of the feedforward filter can be reduced by means of linear prediction based on past values of the error sequence.

If $\mathcal{B}(\omega)$ represents the frequency response of the infinite length feedback predictor, *i.e.*,

$$\mathcal{B}(\omega) = \sum_{n=1}^{\infty} b_n e^{-j\omega nT} \tag{4-228}$$

then the error at the output of the predictor is

$$E_p(\omega) = E_t(\omega) - E_r(\omega)\mathcal{B}(\omega) = E_t(\omega)[1 - \mathcal{B}(\omega)] \tag{4-229}$$

The minimization of the mean square value of this error, *i.e.*,

$$J = \frac{1}{2\pi} \int_{-\pi/T}^{\pi/T} |1 - \mathcal{B}(\omega)|^2 \, |E_t(\omega)|^2 \, d\omega \tag{4-230}$$

over the predictor coefficients $\{b_n\}$ yields the optimum predictor in the form

$$\mathcal{B}(\omega) = 1 - \frac{G(\omega)}{g_0} \tag{4-231}$$

where $G(\omega)$ is the solution to the spectral factorization

$$G(\omega)G^*(-\omega) = \frac{1}{|E_r(\omega)|^2} \tag{4-232}$$

and

$$G(\omega) = \sum_{n=0}^{\infty} g_n e^{-j\omega nT} \tag{4-233}$$

The output of the infinite length linear predictor is a white noise sequence with power spectral density $1/g_0^2$ and the corresponding minimum MSE is given by Equation 4-218. Therefore, the MSE performance of the infinite length predictive DFE is identical to the conventional DFE.

Although these two DFE structures result in equivalent performance if their lengths are infinite, the predictive DFE is suboptimum if the lengths of the two filters are finite. The reason for the optimality of the conventional DFE is relatively simple. The optimization of its tap coefficients in the feedforward and feedback filters is done jointly. Hence, it yields the minimum MSE. On the other hand, the optimizations of the feedforward filter and the feedback predictor in the predictive DFE are done separately. Hence. its MSE is at least as large as that of the conventional DFE. In spite of this suboptimality of the predictive DFE, it is suitable as an equalizer for trellis-coded signals, where the conventional DFE is not as suitable, as described in the next chapter.

4.5.4 Equalization at the Transmitter: Tomlinson-Harashima Precoding

If the channel response is known to the transmitter, the equalizer can be placed at the transmitter end of the communication system. Thus, the noise enhancement that is generally inherent when the equalizer (linear or DFE) is placed at the receiver is avoided. In practice, however, channel characteristics generally vary with time, so it is cumbersome to place the entire equalizer at the transmitter.

In wireline channels, the channel characteristics do not vary significantly with time. Therefore, it is possible to place the feedback filter of the DFE at the transmitter and the feedforward filter at the receiver. This approach has the advantage that the problem of error propagation due to incorrect decisions in the feedback filter is completely eliminated. Thus, the tail (postcursors) in the channel response is cancelled without any penalty in the SNR. The linear fractionally spaced feedforward part of the DFE, which ideally is the WMF, can be designed to compensate for ISI that results from any small time variation in the channel response. The synthesis of the feedback filter of the DFE at the transmitter side is usually performed after the response of the channel is measured at the receiver by the transmission of a channel probe signal and the receiver sends to the transmitter the coefficients of the feedback filter.

The one problem with this approach to implementing the DFE is that the signal points at the transmitter, after subtracting the postcursors of the ISI, generally have a larger dynamic range than the original signal constellation and, consequently, require a larger transmitter power. This problem can be avoided by precoding the information symbols prior to transmission as described by Tomlinson (1971) and Harashima and Miyakawa (1972).

We describe the precoding technique for a P AM signal constellation. Since a square QAM signal constellation may be viewed as two P AM signal sets on quadrature carriers, the precoding is easily extended to QAM. For simplicity, we assume that the feedforward filter in the DFE is the WMF and that the channel response, characterized by the parameters $\{f_i, 0 \le i \le L\}$, is perfectly known to the transmitter and the receiver. The information symbols $\{I_k\}$ are assumed to take the values $\{\pm 1, \pm 3, \cdots, \pm(M-1)\}$.

In the precoding, the ISI due to the postcursors $\{f_i, 0 \le i \le L\}$ is subtracted from the symbol to be transmitted and, if the difference falls outside of the range $(-M,\ M]$, it is reduced to the range by subtracting an integer multiple of $2M$ from this difference. Hence, the precoder output may be expressed as

$$a_k = I_k - \sum_{L}^{j=1} f_j a_{k-j} + 2Mb_k \tag{4-234}$$

where $\{b_k\}$ represents the appropriate integer that brings $\{a_k\}$ to the desired range. In other words, $\{a_k\}$ is reduced to the desired range by performing a modulo-$2M$ operation.

The modulo operation is defined mathematically by the function

$$m_y(x) = x - yz$$

where $y > 0$ and $z = \left\lfloor \frac{x+y/2}{y} \right\rfloor$ is a unique integer such that $m_y(x) \in [-y/2, y/2]$. In our case $y = 2M$. By using the z transform to describe the operation of the precoder, we have

$$A(z) = I(z) - [F(z) - 1]A(z) + 2MB(z) \tag{4-235}$$

where the channel coefficient f_0 is normalized to unity for convenience. Hence, the transmitted sequence is

$$A(z) = \frac{I(z) + 2MB(z)}{F(z)} \tag{4-236}$$

Since the channel response is $F(z)$, the received signal sequence may be expressed as

$$\begin{aligned} V(z) &= A(z) + W(z) \\ &= [I(z) + 2MB(z)] + W(z) \end{aligned} \tag{4-237}$$

where $W(z)$ represents the AWGN term. Therefore, the received data sequence term $I(z) + 2MB(z)$ at the input to the detector is free of ISI and $I(z)$ can be recovered from $V(z)$ by use of a symbol-by-symbol detector that decodes the symbols modulo-$2M$. Figure 4-26 illustrates the block diagram of the system that implements the precoder and the feedback filter of the DFE at the transmitter.

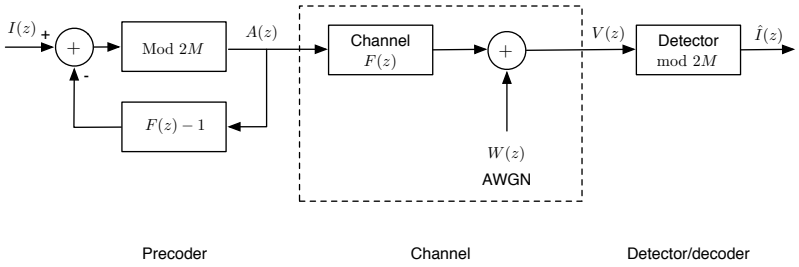

Figure 4-26: Tomlinson-Harashima precoding.

The placement of the feedback filter at the transmitter makes it possible to use the DFE in conjunction with trellis-coded modulation (TCM). Since the equalizer at the receiver is a linear filter, decisions from the output of the Viterbi (TCM) detector can be used to adjust the coefficients of the equalizer. In this case, the Viterbi detector performs the modulo-$2M$ operations in its metric computations.

4.6 Reduced Complexity ML Detectors

The performance results of the three basic equalization methods described above, namely, MLSE, linear equalization (LE), and decision-feedback equalization (DFE), clearly show the superiority of MLSE in channels with severe ISI. Such channels are encountered in wireless communications and in high-density magnetic recording systems. The performance advantage of MLSE has motivated a significant amount of research on methods that retain the performance characteristics of MLSE, but do so at a reduced complexity.

The early work on the design of reduced complexity MLSE focused on methods that reduce the length of the ISI span by preprocessing the received signal prior to the maximum-likelihood detector. Falconer and Magee (1973) and Beare (1978) used a linear equalizer to reduce the span of the ISI to some small specified length prior to the Viterbi detector. Lee and Hill (1977) employed a DFE in place of the LE. Thus, the large ISI span in the channel is reduced to a sufficiently small length, called the *desired impulse response*, so that the complexity of the Viterbi detector following the LE or DFE is manageable. We may view this role of the LE or the DFE, prior to the Viterbi detector, as equalizing the channel response to a specified partial-response characteristic of short duration (the desired impulse response) which the Viterbi detector can handle with sufficiently small complexity. The choice of the desired impulse response is tailored to the ISI characteristics of the channel. This approach to reducing the complexity of the Viterbi detector has proved to be very effective in high-density magnetic recording systems, as illustrated in the papers by Siegel and Wolf (1991), Tyner and Proakis (1993), Moon and Carley (1988), and Proakis (1998).

Another general approach is to reduce the complexity of the Viterbi detector directly, by reducing the number of surviving sequences. The papers by Vermuelen and Hellman (1974), Fredricsson (1974), and Foschini (1977) describe algorithms that reduce the number ofsurviving sequences in the Viterbi detector. Other works on this class of methods include the papers by Clark *et al.*. (1984, 1985) and Wesolowski (1987a).

The most effective approach in terms of performance for reducing the complexity of the Viterbi detector directly is the method described in the papers by Bergmans *et al.*(1987), Eyuboglu and Qureshi (1988), and Duel-Hallen and Heegard (1989). The filter preceding the Viterbi detector is the whitened matched filter (WMF) described previously. The WMF reduces the channel to one that has a minimum phase characteristic. The basic algorithm described in these papers for reducing the computational complexity of the Viterbi detector employs decision feedback within the Viterbi detector to reduce the effective length of the ISI from L terms to L_0 terms, where $L_0 < L$. This may be accomplished in one of two ways, as described by Bergmans *et al.*(1987), either by using "global feedback" or "local feedback" from preliminary decisions that are present in the Viterbi detector. The use of global feedback is illustrated in Figure 4-27, where preliminary decisions obtained by using the most probable surviving sequence from the Viterbi detector are used to synthesize the tail in the ISI due to the channel coefficients $(f_{L_0+1}, f_{L_0+2}, \cdots, f_{L-1}, f_L)$. Thus, for M-ary modulations, the computational complexity of the Viterbi detector is reduced from M^L to M^{L_0}, which amounts to a reduction by the factor M^{L-L_0}. The primary drawback of using global feedback is that if one or more of the symbols $\hat{I}_{k-L_0-1}, \cdots, \hat{I}_{k-L}$ in the most probable surviving sequence are incorrect, the subtraction of the tail in the ISI is also incorrect and, thus, the metric computations are corrupted by the residual ISI resulting from this imperfect cancellation.

To remedy this problem, one may use the preliminary decisions corresponding to each surviving sequence to cancel the ISI in the tail of the corresponding surviving sequence. Thus, the ISI will be perfectly cancelled when the correct sequence is among the surviving sequences, even

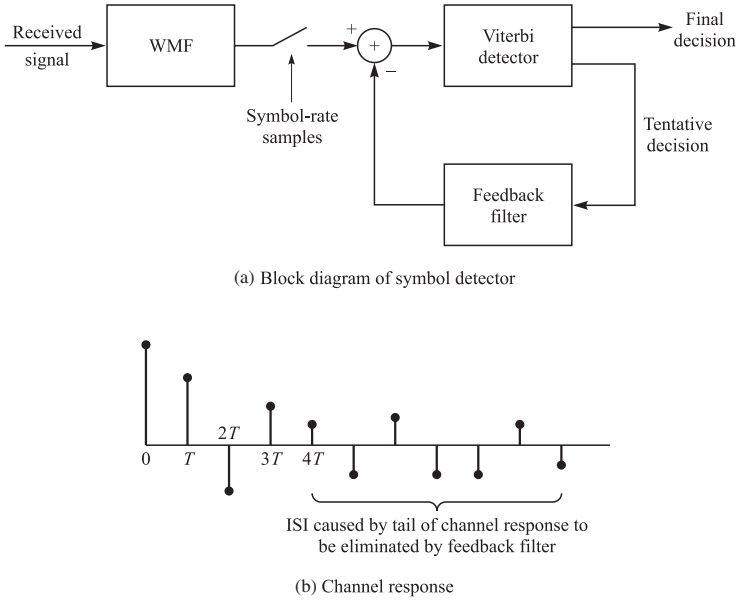

Figure 4-27: Reduced complexity ML sequence detector using feedback from the Viterbi detector.

if it is not the most probable sequence. Bergmans *et al.*(1987) described this approach as using "local feedback" to perform the tail cancellation. It is interesting to note that if L_0 is selected as unity ($L_0 = 1$), the Viterbi detector reduces to the simple feedback filter of a conventional DFE. At the other extreme, when $L_0 = L$, we have a full complexity Viterbi detector. The analytical and simulation results given in the paper by Bergmans *et al.*(1987) clearly illustrate that local feedback gives superior performance to global feedback.

4.7 Iterative Equalization and Decoding-Turbo Equalization

Iterative decoding and the turbo-coding principle that was described in Section **??** can be applied to channel equalization. Suppose the transmitter of a digital communication system employs a binary systematic convolutional encoder followed by a block interleaver and a modulator. The channel is a linear time-dispersive channel that introduces ISI. In such a case, we may view the channel as an inner encoder in a serially concatenated code. Hence, we can apply iterative decoding based on the MAP criterion.

Figure 4-28: Iterative equalization and decoding.

The basic configuration of the iterative equalizer-decoder is shown in Figure 4-28. The input to the MAP equalizer is the sequence $\{v_k\}$ from the WMF. The equalizer computes the logarithm of the likelihood ratio of the coded bits, denoted as $L^E(\hat{x})$, which represents the a posteriori values of the coded bits. The outer decoder receives as an input the extrinsic part of $L^E(\hat{x})$, which is defined as

$$L_e^E(\hat{x}) = L^E(\hat{x}) - L_e^D(\hat{x}) \tag{4-238}$$

where $L_e^D(\hat{x})$ is the extrinsic part of the outer decoder output after interleaving. $L_e^E(\hat{x})$ is deinterleaved prior to being fed to the outer decoder.

The outer decoder computes the logarithm of the likelihood ratio of the coded bits, denoted by $L^D(\hat{x}')$ and the information bits, denoted as $L^D(\hat{I})$. The extrinsic part of $L^D(\hat{x}')$, denoted as $L_e^D(\hat{x}')$, is the incremental information about the current bit obtained by the decoder after all the information for all the received bits. The extrinsic information is computed as

$$L_e^D(\hat{x}') = L^D(\hat{x}') - L_e^E(\hat{x}') \tag{4-239}$$

$L_e^D(\hat{x}')$ is interleaved to produce $L_e^D(\hat{x})$ and fed to the MAP equalizer. We emphasize the importance of feeding back only the extrinsic part $L^D(\hat{x})$, thus, minimizing the between the a priori information used by the equalizer and previous equaloutputs. Similarly, we reduce the a posteriori information $L^E(\hat{x})$ by the a priori information values $L_e^D(\hat{x})$ to obtain the extrinsic information value $L_e^E(\hat{x})$, which is fed to the outer decoder after deinterleaving.

The computation of the log-likelihood ratios is described in the paper by Bauch *et al.*(1997). The power of this iterative equalization-decoding scheme can be assessed from the performance results given in this paper. Figure **??** illustrates the bit error probability obtained through simulation of the five-tap time-invariant channel given in **??**c. The outer decoder used is a rate 1/2 recursive systematic convolutional with constraint length $K = 5$. The interleaver used was a pseudorandom block interleaver of length $N = 4096$ bits. Binary PSK was used for modulation. The graph illustrates the performance gain as the number of iterations is increased. We observe that after six iterations, the performance of the iterative equalizer-decoder is within 0.8 dB of the performance of the encoded data without ISI, at a bit error probability of 10^{-4}. Hence, the iterative equalizer eliminates nearly the entire loss due to ISI. In contrast, the optimum (noniterative) Viterbi detector for this channel suffers a loss of approximately 7 dB, due to ISI, as can be observed from Figure **??**b. Therefore, the iterative equalizer has achieved a performance gain of about 6 dB, aside from the coding gain due to the convolutional code. The performance of this method of iterative equalization has been evaluated for cellular radio channels by Bauch

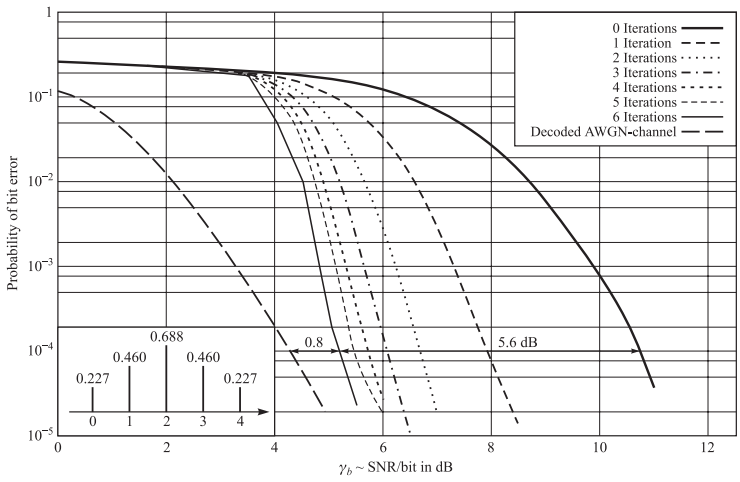

Figure 4-29: Channel taps and bit error rate for a time-invariant channel. [from Bauch *et al.*(1997)]

et al.(1998). An implementation ofiterative equalization-decoding using non-linear circuits is described in a paper by Hagenauer *et al.*(1999).

An alternative approach to iterative equalization-decoding is to employ a parallel concatenated code (turbo code) followed by a block interleaver and a modulator at the transmitter side. The receiver employs a MAP equalizer followed by a turbo decoder. The extrinsic information generated by the turbo decoder is fed back to the MAP equalizer. Thus, we have an iterative equalizer-turbo decoder structure, which is called a turbo equalizer. Turbo equalization is treated by Raphaeli and Zarai (1998) and Douillard *et al.*(1995).

4.8 Bibliographical Notes and References

The pioneering work on signal design for bandwidth-constrained channels was done by Nyquist (1928). The use of binary partial-response signals was originally proposed by Lender (1963) and was later generalized by Kretzmer (1966). Other early work on problems dealing with intersymbol interference (ISI) and transmitter and receiver optimization with constraints on ISI was done by Gerst and Diamond (1961), Tufts (1965), Smith (1965), and Berger and Tufts (1967). "Faster than Nyquist" transmission has been studied by Mazo (1975) and Foschini (1984).

Channel equalization for digital communications was developed by Lucky (I965, 1966), who focused on linear equalizers that were optimized using the peak distortion criterion. The mean-square-error criterion for optimization of the equalizer coefficients was proposed by Widrow (1966).

Decision-feedback equalization was proposed and analyzed by Austin (1967). Analyses of the performance of the DFE can be found in the papers by Monsen (1971), George *et al.*(1971), Price (1972), Salz (1973). Duttweiler *et al.*(1974), and Altekar and Beaulieu (1993).

The use of the Viterbi algorithm as the optimal maximum-likelihood sequence estimator for symbols corrupted by ISI was proposed and analyzed by Forney (1972) and Omura (1971). Its use for carrier-modulated signals was considered by Ungerboeck (1974) and MacKenchnie (1973).

The use of iterative MAP algorithms in suppressing ISI in coded systems, called turbo equalization, represents a major new advance in suppression of intersymbol interference in signal transmission through band-limited channels. It is anticipated that iterative MAP equalization algorithms will be incorporated in future communication systems. The implementation of turbo equalization. described in the paper by Hagenauer *et al.*(1999), is the first attempt at implementing an iterative MAP equalization algorithm in a coded system.

Problems

1. A channel is said to be *distortionless* if the response $y(t)$ to an input $x(t)$ is $Kx(t-t_0)$, where K and t_0 are contants. Show that if the frequency response of the channel is $A(f)e^{j\theta(f)}$, where $A(f)$ and $\theta(f)$ are real, the necessary and sufficient conditions for distortionless transmission are $A(f) = K$ and $\theta(f) = 2\pi f t_0 \pm n\pi$, $n = 0, 1, 2, \cdots$.

2. Consider a four-level PAM system with possible transmitted levels, 3, 1, -1, and -3. The channel through which the data is transmitted introduces intersymbol interference over two successive symbols. The equivalent discrete-time channel model is shown in Figure 4-30. $\{\eta_k\}$ is a sequence of real-valued independent zero-mean Gaussian noise variables

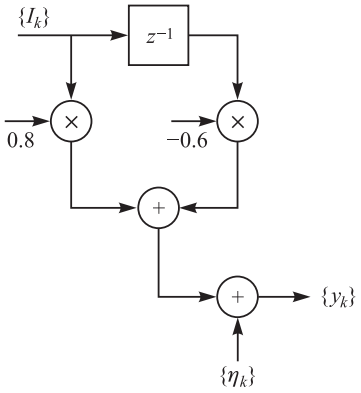

Figure 4-30: Problem 2

with variance $\sigma^2 = N_0$. The received sequence is

$$\begin{aligned} y_1 &= 0.8I_1 + n_1 \\ y_2 &= 0.8I_2 - 0.6I_1 + n_2 \\ y_3 &= 0.8I_3 - 0.6I_2 + n_3 \\ &\vdots \\ y_k &= 0.8I_k - 0.6I_{k-1} + n_k \end{aligned}$$

   (a) Sketch the tree structure, showing the possible signal sequences for the received signals y_1, y_2, and y_3.

   (b) Suppose the Viterbi algorithm is used to detect the information sequence. How many probabilities must be computed at each stage of the algorithm?

   (c) How many surviving sequences are there in the Viterbi algorithm for this channel?

   (d) Suppose that the received signals are

$$\begin{aligned} y_1 &= 0.5 \\ y_2 &= 2.0 \\ y_3 &= -1.0 \end{aligned}$$

   Determine the surviving sequences through stage y_3 and the corresponding metrics.