INFORMATION THEORY OF MODULATION CODES AND WAVEFORMS

INTRODUCTION

The fundamental problem of communication is the conveying of information (which may take several different forms) from a generating source through a communication medium to a desired destination. This conveyance of information, invariably, is achieved by transmitting signals that contain the desired information in some form and that efficiently carry the information through the communication medium. We refer to the process of superimposing an information signal onto another for efficient transmission as *modulation*.

Several factors dictate modulating the desired information signal into another signal more suitable for transmission. The following factors affect the choice of modulation signals:

1. The need to use signals that efficiently propagate through the communication medium at hand. For example, if the communication medium is the atmosphere (or free space), one might use a radio frequency (RF) signal at some appropriate frequency, whereas for underwater communications, one might use an acoustical signal.
2. Communication media invariably distort stochastically signals transmitted through them, which makes information extraction at the receiver nonperfect and most often nonperfect. Thus, a need exists to design modulation signals that are robust to the stochastic (and other) effects of the channel, to minimize its deleterious effects on information extraction.
3. It is highly desirable that communication systems convey large amounts of information per unit time. The price we pay in increasing the in formation rate is often an increase in the required transmitted signal bandwidth. We are interested in modulation signals that can accommodate large information rates at as small a required bandwidth as possible.
4. The power requirements (i.e., average power and peak power) of the transmitted signals to achieve a certain level of performance in the presence of noise introduced during transmission are of paramount importance, especially in power-limited scenarios, such as portable radio and deepspace communications. Our preference is for signals that require as little power as possible for a desired performance level.

The problem of designing modulation signals that possibly optimize some aspect of performance, or satisfy some constraints imposed by the communication medium or the hardware, is known generally as signal design. Signal design problems are important and widely prevalent in communications.

Currently, a proliferation of products make use of modulation to transmit information efficiently. Perhaps the most prevalent and oldest examples are commercial broadcast stations that use frequency modulation (FM) or amplitude modulation (AM) to transmit audio signals through the atmosphere. Another example are data modems that are used to transmit and receive data through telephone lines. These two examples have obvious similarities but also some very important differences. In the broadcast station example, the information to be communicated (an audio signal) is analog and is used to directly modulate a radio-frequency (RF) carrier, which is an example of *analog modulation.* On the other hand, the data communicated through a modem come from the serial port of a computer and are discrete (in fact they are binary; that is, they take two possible values, "0" or "1"), which results in a *digitally modulated* signal. Clearly, the difference between analog and digital modulation is not in the nature of the transmitted signals, because the modulation signals are analog in both cases. Rather, the difference is in the nature of the set of possible modulation signals, which is discrete (and in fact finite) for digitally modulated signals and infinitely uncountable for analog modulation.

The simplest possible digital modulation system consists of two modulation signals. One signal corresponds to the transmission of a "0" and the other of a "1," which is called *binary modulation.* Binary digits (bits) are communicated using binary modulation by assigning a signal in a one-to-one correspondence to each of the two possible logical values of a bit. This mapping between bits and signals is done at a rate equal to the bit rate (i.e., the number of bits/second arriving at the input of the modulator). In response to each transmitted modulation signal, the channel produces a received signal at its output, which is a randomly distorted replica of the transmitted signal. To extract the information superimposed on the modulation signals, a processor, called a receiver or a detector, processes the noisy signal received. The function of the detector is to decide which of the two (in this case) possible signals was transmitted, and in doing so correctly, it recovers the correct value for the transmitted bit. Because of the presence of stochastic noise in the received signal, the receiver may make an incorrect decision for some transmitted bits. The probability of making a decision error in extracting the transmitted bits is known as the bit-error probability or the bit-error rate (BER). The performance of communication systems using digital modulation is invariably measured by their achieved BER, as a function of the transmitted energy per information bit. Receivers that achieve the smallest possible BER for a given channel and modulation signal set are called optimal.

Binary modulation systems are the simplest to implement and detect, but they are not necessarily the most efficient in communicating information. Modulators with larger signal sets use a smaller bandwidth to transmit a given information bit rate. For example, one can envision having a modulation signal set containing four (instead of two) signals: $s_1(t)$, $s_2(t)$, $s_3(t)$, $s_4(t)$. With four signals, we can assign to each a two-bit sequence in a one-to-one cor-

respondence, for example, as follows:

$$s_1(t) \Leftrightarrow 00$$
$$s_2(t) \Leftrightarrow 01$$
$$s_3(t) \Leftrightarrow 10$$
$$s_4(t) \Leftrightarrow 11$$

In this case, each time a transmitted signal is detected correctly, the receiver extracts two (correct) bits. The bit rate has also doubled compared with a binary modulator for the same signaling rate (transmitted signals per second). Because bandwidth is proportional to the signaling rate, we have effectively doubled our transmission efficiency using a modulator with four signals instead of two. Of course, the job of the receiver is now harder because it has to make a four-way decision, instead of just a binary decision, and everything else being the same, the probability of making an erroneous decision increases. We refer to the above modulator as a 4-ary modulator (or a quaternary modulator).

Clearly, the idea can be extended to modulation signal sets that contain $M = 2^k$ signals, for some integer $k = 1, 2, 3 \cdots$. In this case, each transmitted signal carries k bits. We refer to modulators that use M signals as M-ary modulators. As in the 4-ary modulator example above, the advantage of a larger number of modulation signals is that the number of signals per second that needs to be transmitted to accommodate a certain number of bits per second decreases as M increases. Because the number of signals per second determines to a large extent the bandwidth required, more signals means a smaller required bandwidth for a given number of transmitted bits per second, which is a desirable result. The price paid for large signal sets is in complexity and, as previously pointed out, in possibly reduced performance for the same expended average energy per bit.

Although many analog modulation (communication) systems are still in use, the trend is for systems to become digital. Currently, two prominent examples of analog systems becoming digital are cellular phones and digital TV broadcasts. Digital modulation techniques are by far the more attractive.

ANALOG MODULATION

The most prevalent medium for everyday communication is through RF (sinusoidal) carriers. Three quantities exist whose knowledge determines exactly the shape of an RF signal: (1)its amplitude; (2)its phase; and (3)its frequency, as indicated in equation 1:

$$s(t) = A(t)\cos[2\pi\ f_c t + \phi(t)] \tag{1}$$

where f_c is the frequency of the sinusoidal signal in Hertz. Information can be conveyed by modulating the amplitude, the instantaneous frequency, or the phase of the carrier (or combinations of the three quantities).

Amplitude Modulation

Let the information signal *m(t)* be baseband and bandlimited to some bandwidth W Hz. A baseband signal bandlimited to W Hz has a frequency spectrum centered at the origin and contains substantially no energy above W HZ. We assume, which is a good assumption in practice, that $W << f_c$. In amplitude modulation (AM), the information signal modulates the amplitude of the carrier according to:

$$u(t) = Am(t)\cos(2\pi\ f_c t + \phi) \tag{2}$$

where ϕ is some fixed carrier phase. Insight into the process of modulation is obtained by looking at the Fourier transform of the modulated signal, given by (see, for example, Reference (1))

$$U(\ f) = \frac{A}{2}[M(\ f - \ f_c)e^{j\phi} + M(\ f + \ f_c)e^{-j\phi}] \tag{3}$$

Figure 1plots the magnitude of the Fourier transform of the modulated signal for a simple choice (for presentation purposes) of the Fourier transform of the information signal. It is easy to see that, whereas the information signal has bandwidth W, the modulated signal has a bandwidth of $2W$. Also, as can be observed from equation 3, no unmodulated carrier component exists, which would be manifested as delta functions at the carrier frequency f_c. We refer to this scheme as double-sideband, suppressed-carrier (DSB-SC), amplitude modulation.

Demodulation of DSB-SC amplitude modulated signals can be achieved by multiplying the received signal by a locally generated replica of the carrier, which is generated by a local oscillator (LO). For best performance, the locally generated carrier must match as closely as possible the frequency and phase of the received carrier. It is usually reasonable to assume the receiver generated carrier frequency matches the carrier frequency in the received signal well.[1] Neglecting noise, for simplicity, and assuming perfect frequency synchronization, the demodulator is described mathematically by

$$\begin{aligned} z(t) &= u(t)\cos(2\pi\ f_c t + \hat{\phi}) \\ &= \frac{A}{2}\, m(t)\cos(\phi - \hat{\phi}) + \frac{A}{2}\, m(t)\cos(4\pi\ f_c t + \phi + \hat{\phi}) \end{aligned} \tag{4}$$

where $\hat{\phi}$ is the phase of the locally generated carrier. Now the component in equation 4at twice the carrier frequency is easily filtered out by low-pass filtering to yield

$$\hat{m}(t) = \frac{A}{2} m(t)\cos(\phi - \hat{\phi}) \tag{5}$$

which is a scaled version of the modulation signal. In the presence of noise, to maximize the signal-to-noise ratio (SNR), it is important that the phase error $(\phi - \hat{\phi})$ be small. The problem of phase synchronization is an important one and is often practically achieved using a phase-locked loop (PLL) (see, for example, References (2)–(5).) When the locally generated carrier is perfectly phase and frequency locked to the phase and frequency of the received signal, detection of the information is referred to as coherent. This is in contrast to noncoherent detection, when the phase of the locally generated carrier does not match that of the received signal. Clearly, coherent detection achieves the ultimate limit in performance. It can be approached in practice by using sophisticated algorithms, at the cost of increased complexity.

A simpler, noncoherent, detector can be used if the transmitted carrier contains an unmodulated component (or a "pilot tone") resulting in what is referred to as DSB mod-

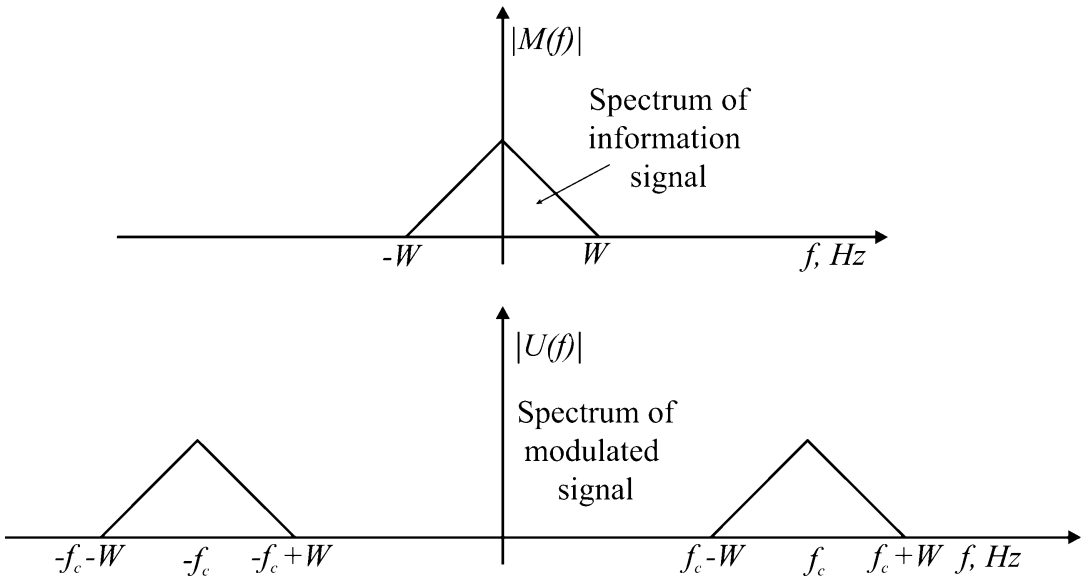

Figure 1. The magnitude of the Fourier transform of a DSB-SC amplitude-modulated signal. (Figure is not to scale).

ulation. In conventional AM (such as in broadcasting), the modulated signal takes the form

$$u(t) = A[1 + am(t)]\cos(2\pi\ f_c t + \phi)$$

with the constraint that $|m(t)| \leq 1$; $a,\ 0 \leq a \leq 1$, is the modulation index. Figure 2shows an example of a conventionally modulated AM signal. Clearly, the modulated signal for conventional AM has a strong unmodulated component at the carrier frequency that carries no information but uses power, and thus, a significant power penalty exists in using it. The benefit resulting from the reduced power efficiency is that simple receivers can now be used to detect the signal. The loss in power efficiency can be justified in broadcasting, where conventional AM is used, because in this case only one high-power transmitter draws power from the power grid, with millions of (now simpler and therefore less costly) receivers.

Demodulation of AM Signals

The most popular detection method for conventional AM is envelope detection. This method consists of passing the received modulated signal [usually after RF amplification and down conversion to some intermediate frequency (IF)] through a rectifier followed by a simple low-pass filter (in the form of a simple, passive, RC circuit). This simple detector is shown in Fig. 3.

Double-sideband amplitude modulation is wasteful in bandwidth, requiring a bandwidth that is twice the baseband signal bandwidth. It can be shown that the two sidebands are redundant, and that the information signal can be obtained if only one sideband was transmitted, which reduces the required bandwidth by a factor of two compared with DSB-AM. At the same time, an improvement in power efficiency, occurs because transmitting the redundant sideband requires not only extra bandwidth but also extra power. When only one sideband is transmitted, the resulting signal is referred to as single sideband (SSB). The general form of a single-sideband signal is

$$u(t) = A[m(t)\cos(2\pi\ f_c t) \pm \hat{m}(t)\sin(2\pi\ f_c t)] \tag{6}$$

where $\hat{m}(t)$ is the Hilbert transform of $m(t)$ given by

$$\hat{m}t = m(t) * \frac{1}{\pi t} \Leftrightarrow \hat{M}(\ f) = M(\ f)H(\ f)$$

where *H(f)* is the Fourier transform of $h(t) = 1/\pi t$ and is given by

$$H(\ f) = \begin{cases} -j, & f > 0 \\ j, & f < 0 \\ 0, & f = 0, \end{cases}$$

In equation 6 , the plus or minus sign determines whether the upper or the lower sideband is chosen. Figure 4shows the spectrum of an upper sideband SSB signal. For a more complete exposition to SSB, including modulation and demodulation methods, consult References (1) and (6–9).

Another amplitude modulation scheme, widely used in TV broadcasting, is vestigial sideband (VSB). The reader is referred to References (1) and (6–9) for more information.

Angle Modulation

Angle modulation of a sinusoidal carrier includes phase modulation (PM) and frequency modulation (FM). In phase modulation, the information signal modulates the instantaneous phase of a high-frequency sinusoidal carrier, whereas in frequency modulation, the information signal directly modulates the instantaneous frequency of the carrier. As the instantaneous frequency and phase of a signal are simply related (the instantaneous frequency is the scaled derivative of the instantaneous phase), clearly PM and FM are also closely related and have similar properties. For angle modulation, the modulated signal is given by

$$u(t) = A\cos[2\pi\ f_c t + \phi(t)]$$

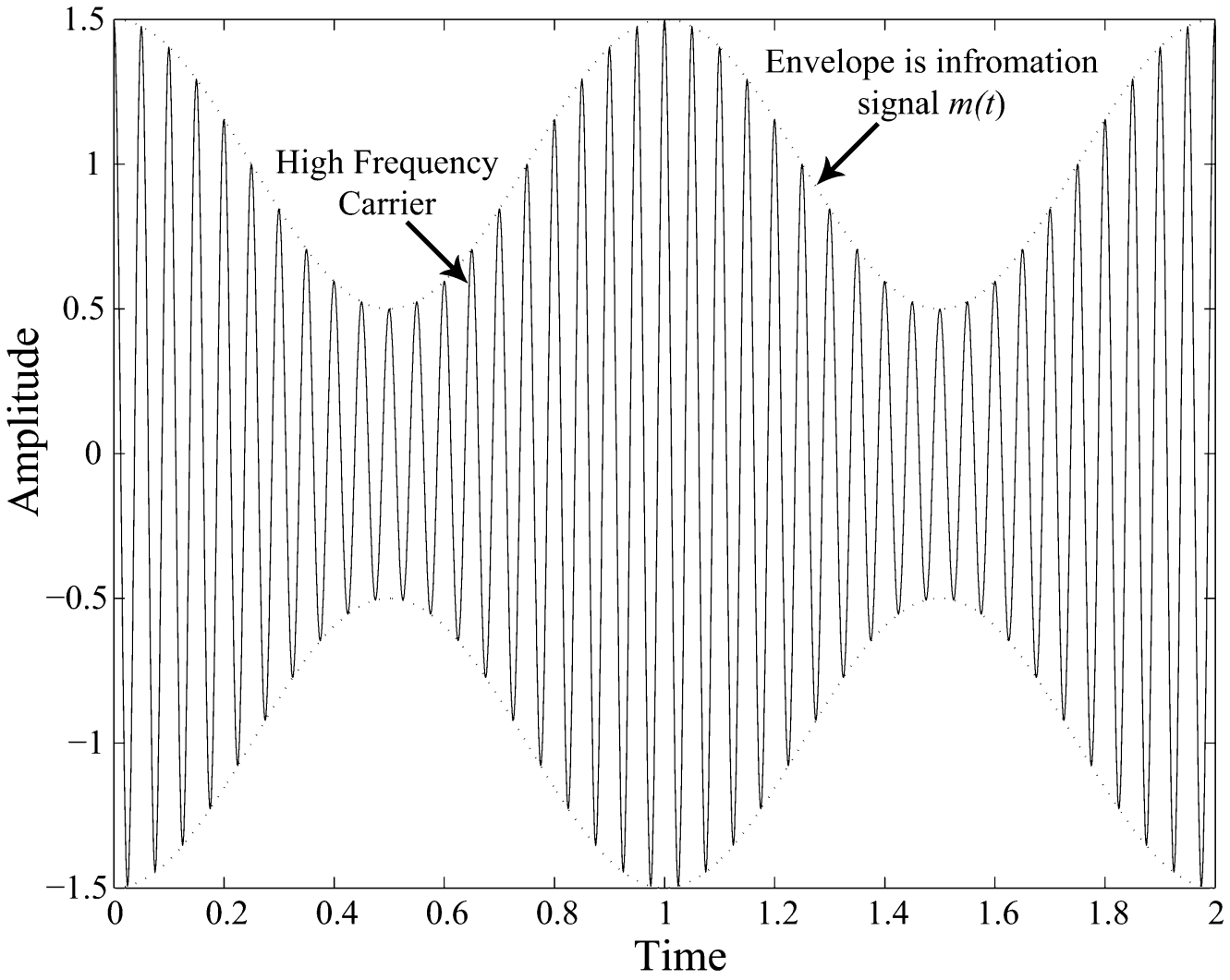

Figure 2. Illustration of a conventionally amplitude-modulated signal.

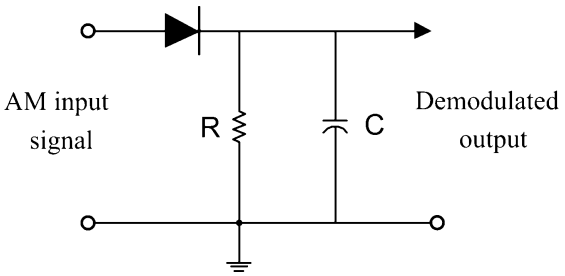

Figure 3. A simple demodulator for conventional AM signals.

where

$$\phi(t) = \begin{cases} d_p m(t) & \text{PM} \\ 2\pi d_f \displaystyle\int_{-\infty}^{t} m(\tau)d\tau & \text{FM} \end{cases}$$

The constants d_p and d_f are the *phase* and *frequency deviation* constants, respectively. These constants, along with the peak amplitude of the information signal, define the peak phase deviation and peak frequency deviation constants, given by

$$\Delta\phi = d_p \cdot |m(t)|$$

and

$$\Delta f = d_f \cdot |m(t)|$$

In turn, the peak deviation constants define the phase and frequency modulation indices according to

$$\beta_p = \Delta\phi$$

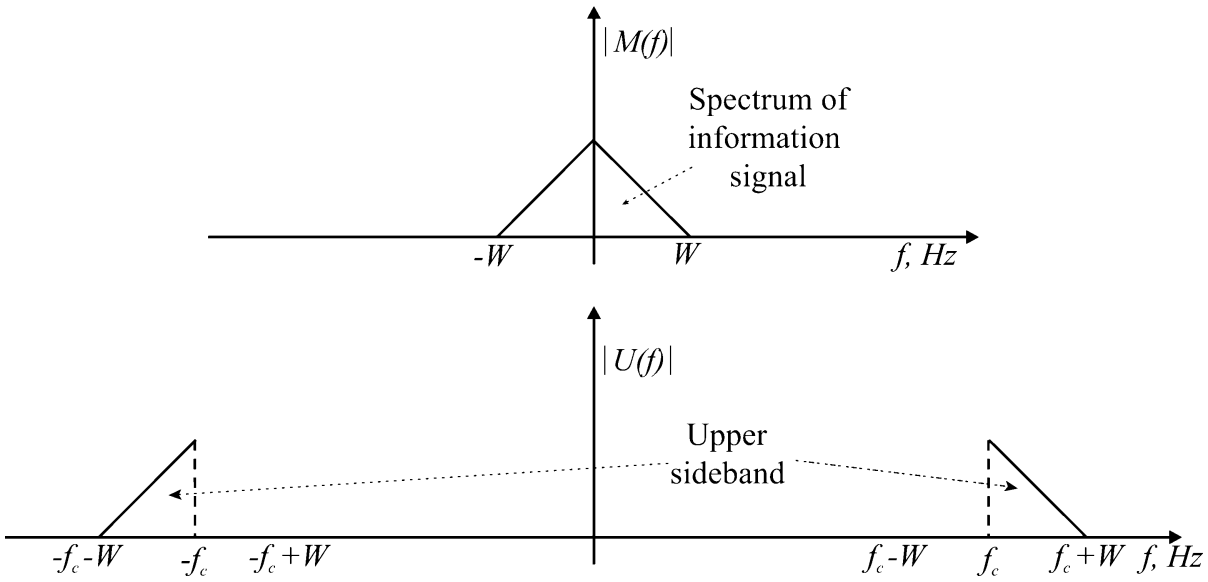

Figure 4. The spectrum of an upper sideband SSB signal.

and

$$\beta_f = \frac{\Delta_f}{W}$$

where W is the bandwidth of the information signal signal $m(t)$. As an example, the peak frequency deviation for FM broadcasts is 75 KHz, and the signal bandwidth is limited to 15 KHz, which yields a modulation index of 5. For illustration, Fig. 5shows typical waveforms for frequency and phase modulation.

The spectrum of an angle-modulated signal is much more difficult to obtain mathematically than in the AM case because angle modulation is nonlinear. Moreover, strictly speaking, angle-modulated signals have an infinite bandwidth. However, an approximation for the effective bandwidth (i.e., the frequency band containing most of the signal energy) of angle-modulated signals is given by Carson's rule:

$$B = 2(\beta + 1)W$$

where β is the phase- or frequency-modulation index and W is the bandwidth of the information signal. The bandwidth of the modulated signal increases linearly as the modulation index increases. FM systems with a small modulation index are called narrowband FM, whereas systems with a large modulation index are called wideband FM. One popular and practical way to generate wideband FM is to first generate a narrowband FM signal (which is easily generated) and then, through frequency multiplication, to convert it into a wideband FM signal at an appropriate carrier frequency. Wideband FM is used in broadcasting, and narrowband FM is used in point-to-point FM radios.

Detection of FM or PM signals takes several different forms, including (PLLs) and discriminators, which convert FM into AM that is then detected as such. For more information on ways to modulate and demodulate angle modulated signals, consult References (1,3,5), and (9).

DIGITAL MODULATION

A wide variety of digital modulation methods exists, depending on the communication medium and the mode of communication, both of which impose constraints on the nature of transmitted signals. For example, for optical systems that use an optical carrier [generated by a light-emitting diode (LED) or a laser], various modulation schemes are particularly suitable, which may not be suitable for RF communications systems. Similarly, modulation schemes used in magnetic recording systems may not be suitable for other systems. Generally, as indicated in the Introduction, the modulation must be matched to the channel under consideration.

Signal Space

In designing and describing digital modulation schemes, it is often desirable to consider modulation signals as points in some appropriate signal space, spanned by a set of orthonormal-basis signals. The dimensionality of the signal space equals the number of orthonormal-basis signals that span it.

A set of signals $\{\phi_1(t), \phi_2(t), \cdots, \phi_N(t)\}$, for $0 \le t \le T$ is orthonormal if the following condition holds:

$$\int_0^T \phi_i(t)\phi_j(t)dt = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$$

If $s(t)$ is any signal in the N-dimensional space spanned by these signals, then it can be expressed as

$$s(t) = \sum_{i=1}^{N} s_i \phi_i(t)$$

for some set or real numbers $s_1, s_2, \cdot, s_N$. The N coefficients uniquely describing $s(t)$ are obtained using

$$s_k = \int_0^T s(t)\phi_k(t)dt, \;\; k = 1, 2, \cdots, N$$

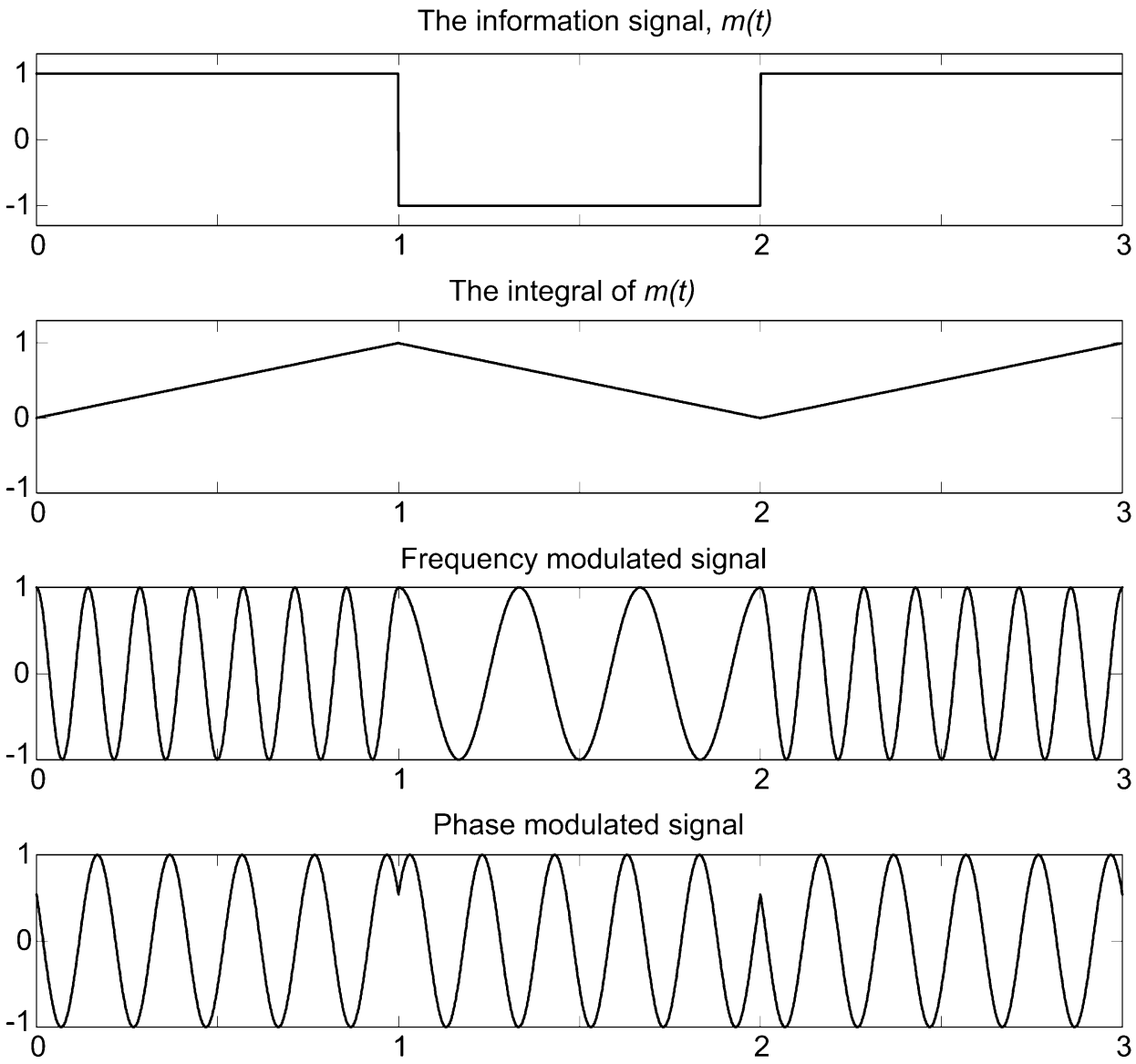

Figure 5. Illustration of frequency- and phase-modulated signals.

Figure 6illustrates the concept of signal space for the special case of two dimensions. In the figure, four distinct signals are represented as points in the signal space.

Perhaps the most widely known and used modulation schemes are those pertaining to RF communication, some of which are examined next.

Phase-Shift Keying

Under phase-shift keying (PSK), the information bits determine the phase of a carrier, which takes values from a discrete set in accordance with the information bits. The general form of M-ary PSK signals (i.e., a PSK signal set containing signals) is given by

$$s_i(t) = \sqrt{\frac{2E}{T}} \cos(2\pi\ f_c t + \theta_i), \quad i = 1, 2, \cdots, M, \quad 0 \le t \le T \tag{7}$$

where

$$\theta_i = \frac{2\pi(i-1)}{M}$$

and

$$E = \int_0^T s_i^2(t)dt$$

is the *signal energy*. Equation (7)is rewritten in a slightly different form as

$$\begin{aligned} s_i(t) &= \sqrt{E}[\cos(\theta_i)\sqrt{\frac{2}{T}}\cos(2\pi\ f_c t) - \sin(\theta_i)\sqrt{\frac{2}{T}}\sin(2\pi\ f_c t)] \\ &= \sqrt{E}[\cos(\theta_i)\phi_1(t) - \sin(\theta_i)\phi_2(t)] \end{aligned}$$

where $\phi_1(t)$ and $\phi_2(t)$ are easily observed to be orthonormal. Thus, PSK signals are points in a two-dimensional space spanned by $\phi_1(t)$ and $\phi_2(t)$. Figure 7illustrates various PSK signal constellations, including binary PSK (BPSK) and 4-

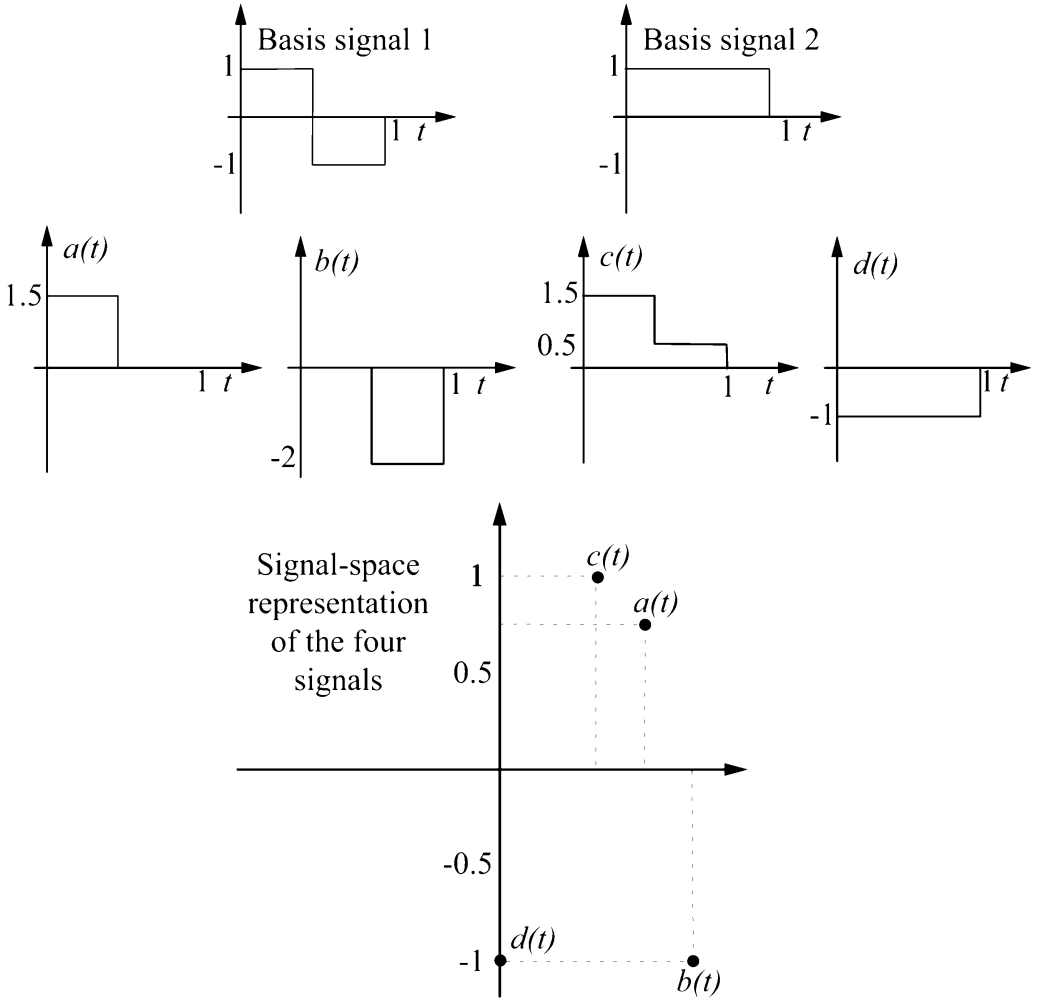

Figure 6. Illustration of the concept of signal space. The two signals on top are the basis signals. Signals *a(t)*, *b(t)*, *c(t)*, and *d(t)* are represented in signal space as points in the two-dimensional space spanned by the two basis signals.

ary PSK, also known as quadrature PSK (QPSK). The figure also illustrates the mapping of information bits to each signal in the constellation. The illustrated mapping, known as Gray coding, has the property that adjacent signals are assigned binary sequences that differ in only one bit. This property is desirable in practice, because, when a detection error is made, it is more likely to be to a signal adjacent to the transmitted signal. Then Gray coding results in a single bit error for the most likely signal errors.

Performance in Additive Gaussian Noise. The simplest channel for data transmission is the additive, white, Gaussian noise (AWGN) channel. For this channel, the transmitted signal is corrupted by and additive Gaussian process, resulting in a received signal given by

$$r(t) = s_i(t) + n(t), \quad 0 \le t \le T \tag{8}$$

where $n(t)$ is zero-mean, white Gaussian noise of spectral density $N_0/2$.

For PSK signals, the optimum receiver (detector), also known as a maximum-likelihood (ML) receiver, decides which of the M possible PSK signals was transmitted by finding the modulation signal that maximizes

$$l_1 = \int_0^T r(t)s_i(t)dt$$

This signal is the well-known correlation receiver, where the most likely signal transmitted is chosen as the one most correlated with the received signal. The correlation receiver involves a multiplication operation, followed by integration. Because processing is linear, it is possible to obtain the same result by passing the received signal through a linear filter with an appropriate impulse response and sampling it at an appropriate instant. The impulse response $h_i(t)$ of the linear filter is easily derived as

$$h_i(t) = s_i(T - t)$$

This linear filter implementation of the optimum receiver is called a matched-filter receiver.

For binary PSK, the probability that the optimal receiver makes a decision error is given by

$$P_{\text{BPSK}}(e) = \frac{1}{2}\text{erfc}(\sqrt{\frac{E}{N_0}}) \tag{9}$$

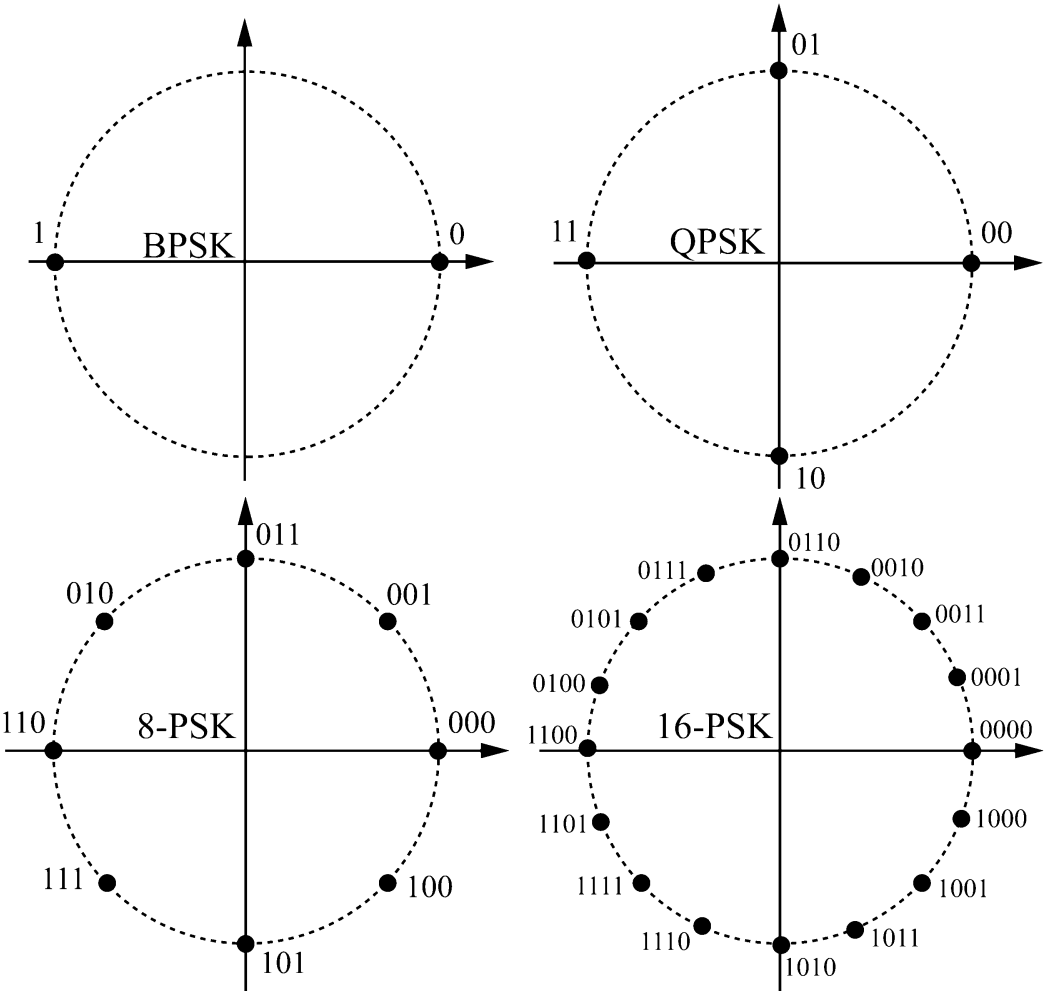

Figure 7. Signal space representation of various PSK constellations. The bit assignments correspond to Gray coding.

where

$$\mathrm{erfc}(x) = 1 - \frac{2}{\sqrt{\pi}} \int_0^x e^{-y^2} dy$$

is the complimentary error-function. In equation 9 , the ratio E/N_0 is the SNR, which determines performance. The performance of QPSK is also derived easily and is given by

$$P_{\mathrm{QPSK}}(e) = P_{\mathrm{BPSK}}(e)[2 - P_{\mathrm{BPSK}}(e)]$$

where $\mathrm{P_{BPSK}(e)}$ is as given in equation 9 . An exact expression for the error probability of larger PSK constellations also exists and is found, for example, in Chapter 9 of Reference (1). Figure 8shows the error probability of various PSK constellations as a function of the SNR per information bit.

Baseband Pulse-Amplitude Modulation

Pulse-amplitude modulation (PAM) is the digital equivalent of AM. The difference is that now only discrete amplitudes are allowed for transmission. M-ary PAM is a one-dimensional signaling scheme described mathematically by

$$s_i(t) = (2i - 1 - M)\sqrt{E}\, p(t), \quad i = 1, 2, \cdots, M, \quad 0 \le t \le T$$

where *p(t)* is a unit-energy baseband pulse. Figure 9shows the signal-space representation of PAM signals assuming $E = 1$. In contrast to PSK signals, clearly not every signal has the same energy; in which case, the constellation is described by its average energy:

$$E_{\mathrm{av}} = \frac{E}{M} \sum_{i=1}^{M} (2i - 1 - M)^2 = (\frac{M^2 - 1}{3})E$$

Performance in Additive Gaussian Noise. Based on the data *r(t)* received (as given in equation 8), the maximum-likelihood receiver for PAM signaling chooses as the most likely signal transmitted the signal that maximizes

$$l_i = (2i - 1 - M) \cdot r - \frac{\sqrt{E}}{2}(2i - 1 - M)^2$$

where

$$r = \int_0^T r(t)\, p(t) dt$$

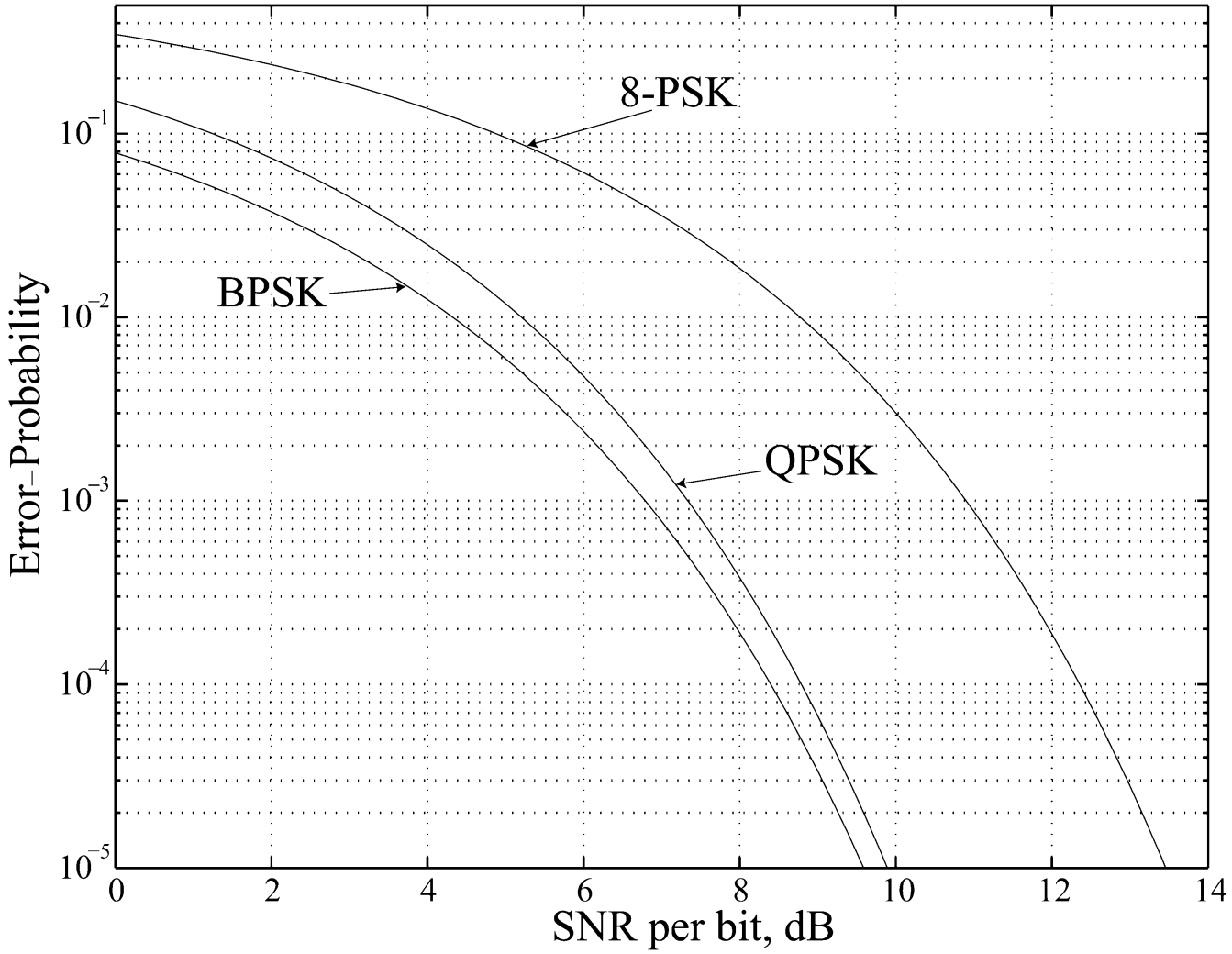

Figure 8. Symbol error probability for BPSK, QPSK, and 8-PSK as a function of the SNR per bit.

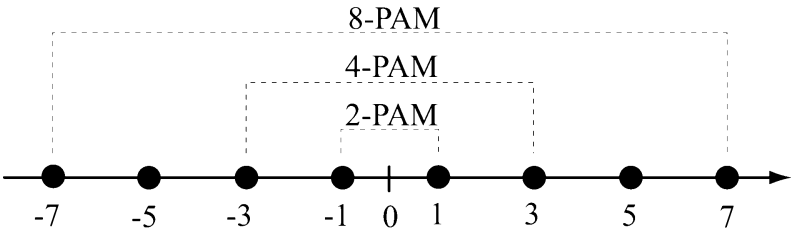

Figure 9. The signal space representation of various PAM constellations.

In signal space, the decision boundaries for this receiver are midway between constellation points, and a decision is made accordingly, based on where r falls on the real line. The error probability for M-ary PAM signals is given by

$$\mathrm{P_{PAM}(e)} = \frac{(M-1)}{M}\mathrm{erfc}\left(\sqrt{\frac{3}{M^2-1}\frac{E_{\mathrm{av}}}{N_0}}\right)$$

The error probability for various PAM constellations is shown in Fig. 10as a function of SNR per bit.

Quadrature Amplitude-Modulation

Quadrature amplitude modulation (QAM) is a popular scheme for high-rate, high-bandwidth efficiency systems. QAM is a combination of both amplitude and phase modulation. Mathematically, M-ary QAM is described by

$$s_i(t) = \sqrt{E}\, p(t)[A_i \cos(2\pi\ f_c t) + B_i \sin(2\pi\ f_c t)], \quad 0 \le t \le T,$$
$$i = 1, 2, \cdots, M$$

where A_i and B_i take values from the set $\{\pm 1, \pm 3, \pm 5, \cdots\}$ and E and $p(t)$ are as defined earlier. The signal space rep-

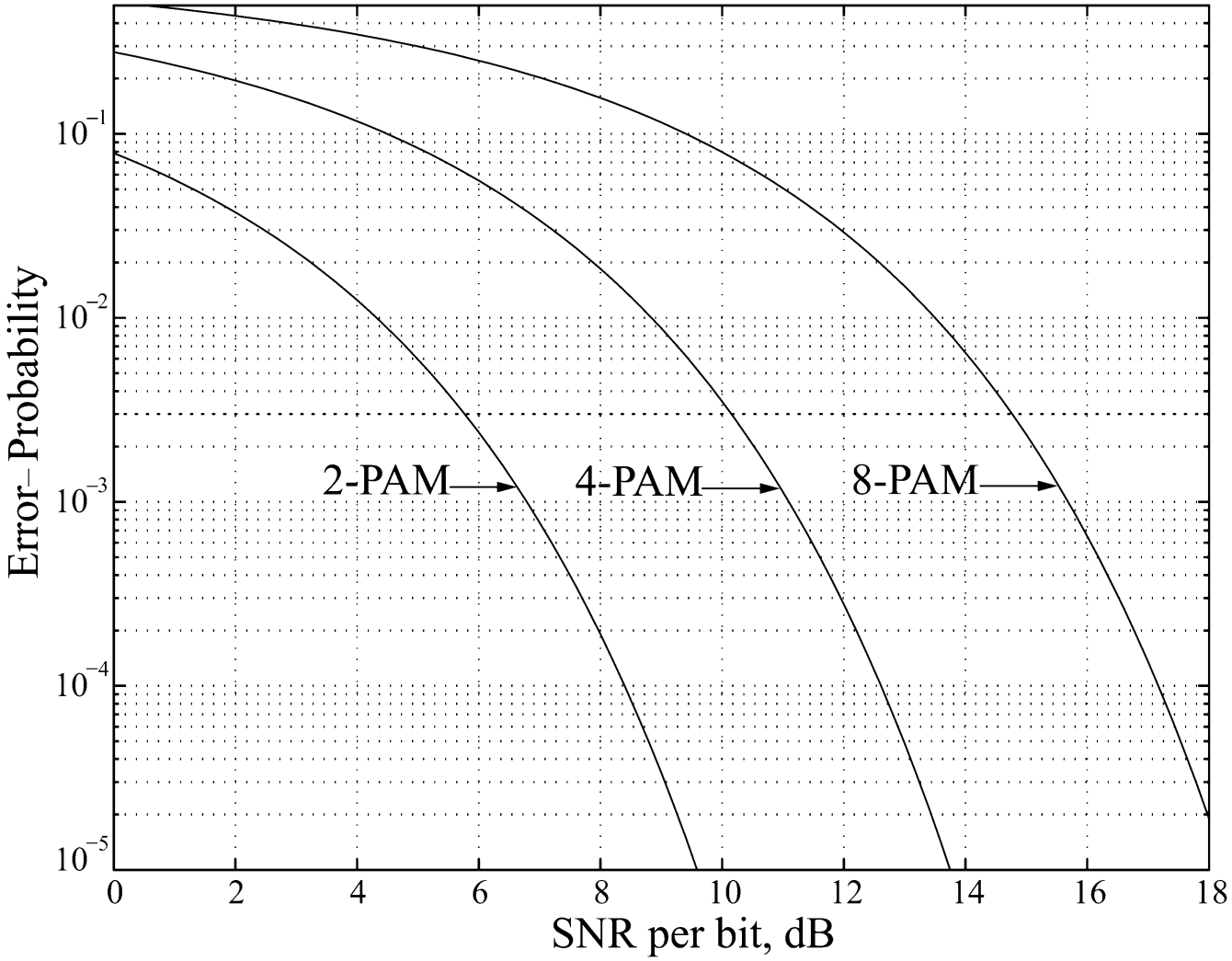

Figure 10. Symbol error probability for 2-, 4-, and 8-PAM as a function of SNR per bit.

resentation of QAM signals is shown in Fig. 11for various values of M, which are powers of 2; that is, $M = 2^k,\ k = 2, 3, \cdots$. For even values of k, the constellations are *square*, whereas for odd values, the constellations have a cross shape and are thus called *cross* constellations. For square constellations, QAM corresponds to the independent amplitude modulation of an in-phase carrier (i.e., the cosine carrier) and a quadrature carrier (i.e., the sine carrier).

Performance in Additive Gaussian Noise. The optimum receiver for QAM signals chooses the signal that maximizes

$$l_i = A_i r_c + B_i r_s - \frac{\sqrt{E}}{4}(A_i^2 + B_i^2)$$

where

$$r_c = \int_0^T r(t)\, p(t)\cos(2\pi\ f_c t)dt$$

and

$$r_s = \int_0^T r(t)\, p(t)\sin(2\pi\ f_c t)dt$$

For square constellations that correspond to independent PAM of each carrier, an exact error probability is derived easily and is given by

$$P_{\mathrm{QAM}}(\mathrm{e}) = 1 - [1 - (1 - \frac{1}{\sqrt{M}})\mathrm{erfc}(\sqrt{\frac{3}{2(M-1)} \cdot \frac{E_{\mathrm{av}}}{N_0}})]^2$$

For cross constellations, tight upper bounds and good approximations are available. Figure 12plots the symbol error probability of various square QAM constellations as a function of SNR per bit.

Frequency-Shift Keying

As the name implies, frequency-shift keying (FSK) modulates the frequency of a carrier to convey information. FSK is one of the oldest digital modulation techniques and was the modulation of choice for the first, low-rate modems. Its main attribute, which makes it of interest in some appli-

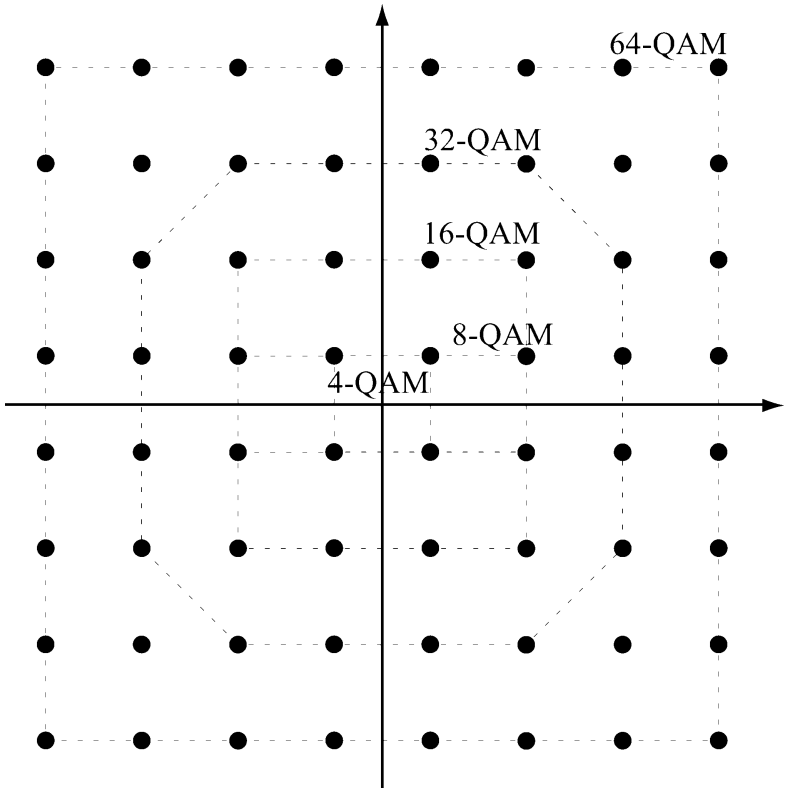

Figure 11. Signal space representation of various QAM constellations.

cations, is that it can be detected noncoherently (as well as coherently), which reduces the cost of the receiver. Mathematically, the modulated M-ary FSK signal is described by

$$s_i(t) = \sqrt{\frac{2E}{T}}\cos[2\pi(f_c + f_i)t], \quad 0 \le t \le T, \quad i = 1, 2, \cdots, M$$

where

$$f_i = (\frac{2i - 1 - M}{2})\Delta \mathrm{r}\ f$$

Δf is the minimum frequency separation between modulation tones. For orthogonal signaling (i.e., when the correlation between all pairs of distinct signals is zero), the minimum tone spacing is $1/2T$. This a condition is often imposed in practice. Orthogonal signaling performs well as a function of energy per bit, but it is also bandwidth-inefficient, which makes it impractical for high-speed, band limited applications.

Performance in Additive Gaussian Noise. FSK is detected coherently or incoherently. Coherent detection requires a carrier phase synchronization subsystem at the receiver that generates locally a carrier phase-locked to the received carrier. The optimum receiver for coherent detection makes decisions by maximizing the following (implementation assumes phase-coherence):

$$l_i = \int_0^T r(t)s_i(t)dt$$

For binary (orthogonal) signaling, the error probability is given simply by

$$\mathrm{P_{FSK}(e)} = \frac{1}{2}\mathrm{erfc}\,(\sqrt{\frac{E}{2N_0}}), \quad \text{(coherent FSK)}$$

which is 3 dB worse than BPSK. For M-ary signaling, an exact expression exists in integral form and is found, for example, in Reference (10). Noncoherent detection does not assume phase coherence and does not attempt to phase-lock the locally generated carrier to the received signal. In this case, it is easy to argue that the phase difference between the LO carrier and the received carrier is completely randomized. An optimum receiver is also derived in this case, and it is one that maximizes over the set of frequency tones

$$l_i = r_{ci}^2 + r_{si}^2$$

where

$$r_{ci}^2 = \int_0^T r(t)\cos[2\pi(f_c + f_i)t]dt$$

and

$$r_{si}^2 = \int_0^T r(t)\sin[2\pi(f_c + f_i)t]dt$$

The exact error-probability performance of this noncoherent receiver is available in analytical form, but it is complicated to compute for the general M-ary case (see, for example, Reference (10)). For the binary case, the error prob-

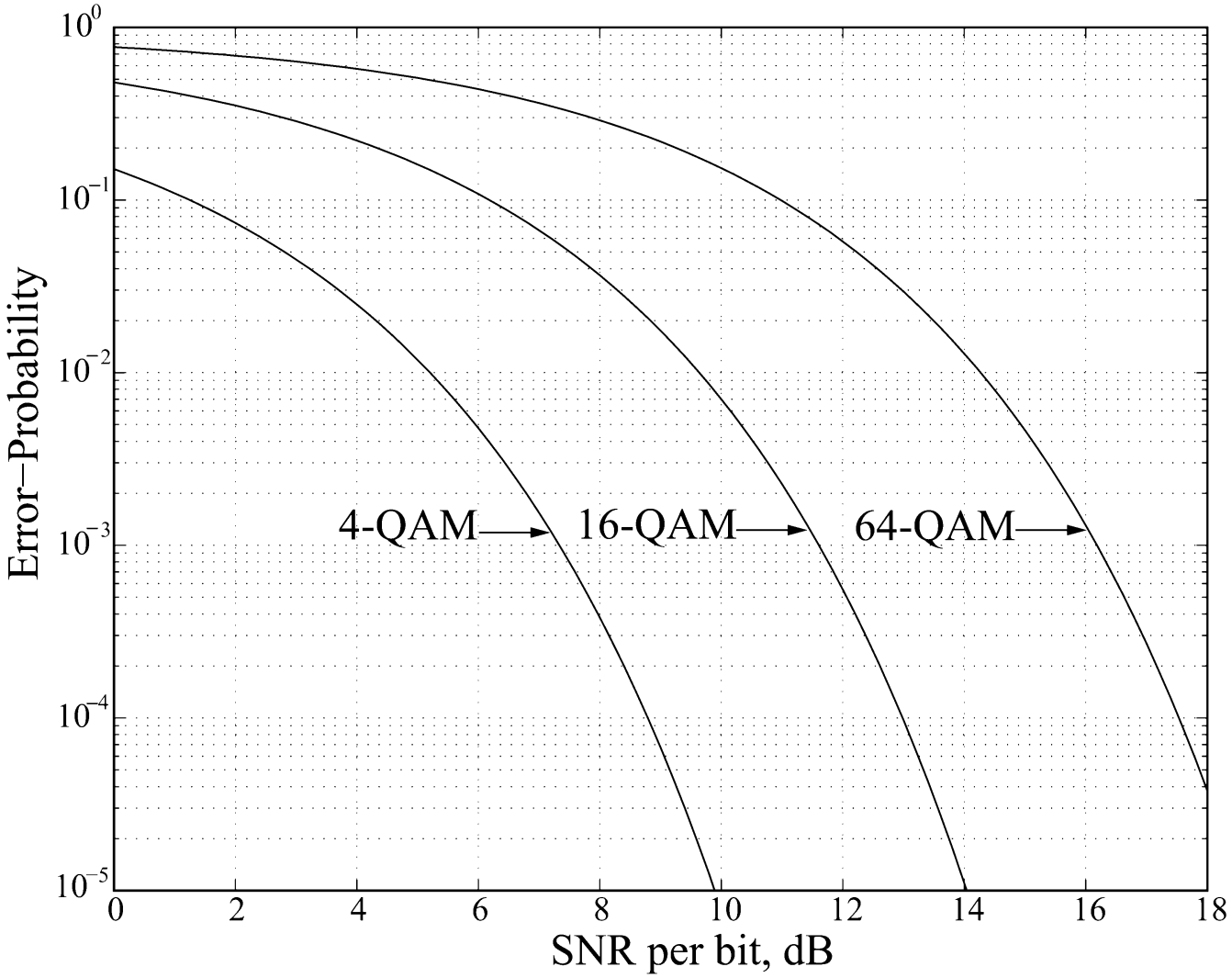

Figure 12. Symbol error probability as a function of SNR per bit for 4-, 16-, and 64-QAM.

ability has a simple form given by

$$\mathrm{P_{FSK}(e)} = \frac{1}{2} e^{-\frac{E}{2N_0}} \quad \text{(noncoherent FSK)}$$

Figure 13compares the performance of coherent and incoherent binary FSK. At an error probability of about 10^{-6}, noncoherent detection is inferior only slightly more than half a decibel compared with coherent detection. However, this small loss is well compensated for by the fact that no carrier phase synchronization is needed for the former.

Continuous-Phase Modulation

All modulation schemes described so far are memoryless, in the sense that the signal transmitted in a certain symbol interval does not depend on any past (or future) symbols. In many cases, for example, when a need exists to shape the transmitted signal spectrum to match that of the channel, it is necessary to constrain the transmitted signals in some form. Invariably, the imposed constraints introduce memory into the transmitted signals. One important class of modulation signals with memory are continuous-phase modulation (CPM) signals. These signals constrain the phase of the transmitted carrier to be continuous, thereby reducing the spectral sidelobes of the transmitted signals. Mathematically, the modulation signals for CPM are described by the expression

$$u(t) = A\cos[2\pi\ f_c t + \phi(t; \mathbf{d})]$$

where

$$\phi(t; \mathbf{d}) = 2\pi \sum_{k=-\infty}^{n} d_k h_k q(t - kT), \quad nT \le t \le (n+1)T$$

The d_k are the modulation symbols and h_k are the modulation indices, which may vary from symbol to symbol. For binary modulation, the modulation symbols are either 1 or -1. Finally, $q(t)$ is the integral of some baseband pulse $p(t)$ containing no impulses (thus guaranteeing that $q(t)$ is continuous)

$$q(t) = \int_{-\infty}^{t} p(\tau) d\tau$$

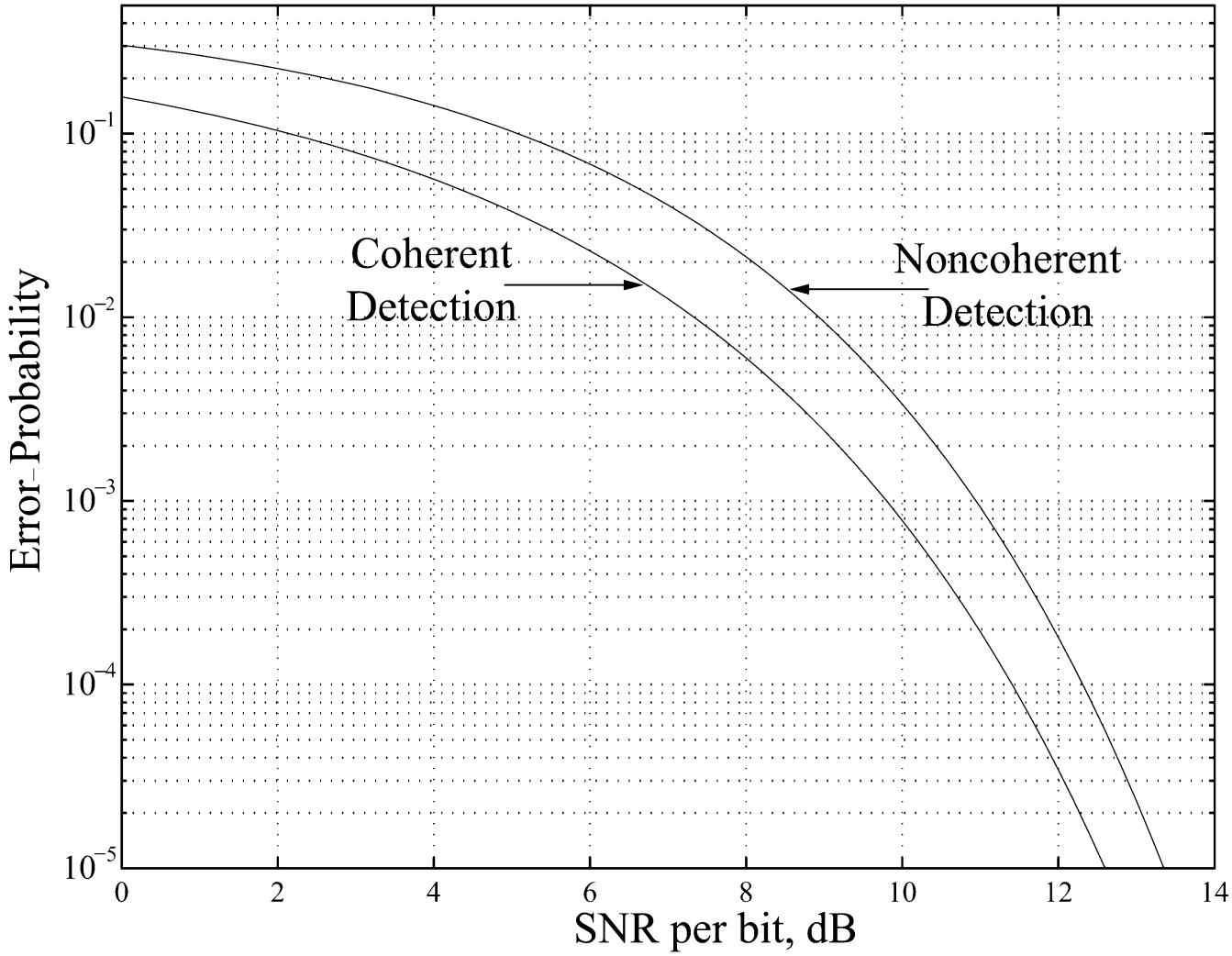

Figure 13. Error probability comparison between coherent and noncoherent FSK.

When $p(t)$ is zero for $t \geq T$, we have what is called *full-response* CPM, otherwise, we have *partial-response* CPM. In general, partial-response CPM achieves better spectral sidelobe reduction than does full-response CPM. A special case of CPM in which the modulation indices are all equal and $p(t)$ is a rectangular pulse of duration T seconds is called continuous-phase FSK (CPFSK). If, $h = 1/2$, we have what is called minimum-shift keying (MSK). A variation of MSK, in which the rectangular baseband pulse is first passed through a filter with a Gaussian-shape impulse response for further reduction in the spectral sidelobes, is called Gaussian MSK (GMSK). Various simple ways for detecting GMSK are available, which combined with its spectral efficiency, has made it a popular modulation scheme. In particular, it is the modulation scheme originally used for the European digital cellular radio standard, known as GSM. For more information on CPM signaling, including spectral characteristics and performance in noise, refer to Reference (10).

Modulation Codes

Another technique for shaping the spectrum of transmitted modulation signals is putting constraints on the sequence of bits sent to the modulator. This coding of bits to shape the spectrum of the transmitted modulation signals is called modulation coding or line coding. Important examples of the use of such codes are in magnetic and optical recording channels. Simple examples of modulation codes are found in the baseband transmission of binary data where a pulse is sent for a binary "1" and its negative for a "0" (called antipodal signaling). If the pulse amplitude does not return to zero in response to consecutive similar bits, then we have nonreturn-to zero (NRZ) signaling. If the pulse returns to zero, then we have return-to-zero (RZ) signaling. The encoding of bits using NRZ and RZ signaling is illustrated in Fig. 14.

It is often desirable to have a transmitted pulse sequence, in response to random input bits, with no spectral component at zero frequency (i.e., in dc). This condition is desirable, for example, when the modulation signals are sent through a channel with a null at dc. If the bits arriving

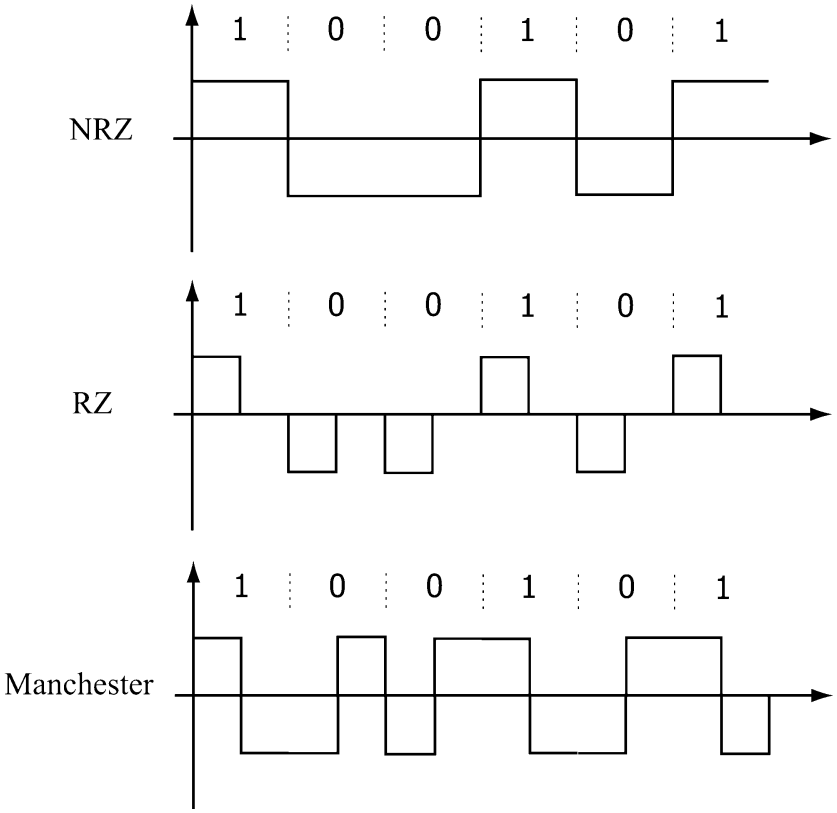

Figure 14. Illustration of NRZ, RZ, and Manchester coding.

at the input of the modulator are truly random (each with probability 1/2 of being zero or one) and independent, then the expected value of the dc component of the transmitted signal is zero. However, at any given time (even though the average is zero), a significant dc component may be caused by the transmission of a long sequence of zeros or ones. Besides the creation of a dc component, these long sequences of zeros or ones also negatively affect the performance of the timing recovery system at the receiver, whose function is to establish time synchronization (essential before data detection).

Biphase or Manchester pulses have the property of zero dc over each bit interval. These pulses and their encoding are illustrated in Fig. 14, along with NRZ and RZ signaling. An important property of a line code that describes the dc variations of a baseband signal is the running digital sum (RDS) (11). The RDS is the running sum of the baseband amplitude levels. It has been shown that, if the RDS for a modulation code is bounded, then the code has a null at dc (12). This process facilitates transmission of the modulated data through channels with a null at dc and avoids a form of intersymbol-interference (ISI) known as baseline wander. A converse result also shows that modulation codes, generated by finite-state machines, which have a spectral null at dc, have a bounded RDS (13).

Run-Length Limited Codes. Run-length limited (RLL) codes are an important class of modulation codes, which are often used in magnetic recording systems. RLL codes impose constraints on the minimum and maximum number of consecutive zeros between ones and are also called *(d, k)* codes, where d is the minimum number of zeros and k is the maximum number of zeros between ones. The minimum number of zeros between ones ensures that ISI is kept small, and the maximum number of zeros between ones ensures that the transmitted signal has enough transitions in it to aid in timing recovery. RLL codes (and in fact a much larger class of codes) are conveniently described by finite-state machines (FSMss). An FSM consists of a set of interconnected states that describe the allowable bit transitions (paths). The interconnections between all possible pairs of states are often described by a two-dimensional state transition matrix, which is known also as the *adjacency* matrix. A one at the i,j position in the matrix means that there is a path from state i to state j. A zero means that no path exists between the two states. Figure 15shows the FSM for the (1,3) *(d, k)* code. It consists of four states, and its adjacency matrix is given by

$$\mathrm{A} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

Clearly, the constraints imposed on the binary sequences (in the form of d and k) limit the number of possible sequences of a given length n, which satisfy the constraint to a subset of the total number of 2^n possible sequences. If the number of sequences of length n satisfying the *(d, k)* constraints is $M(n)$, then the *capacity* of the code is defined

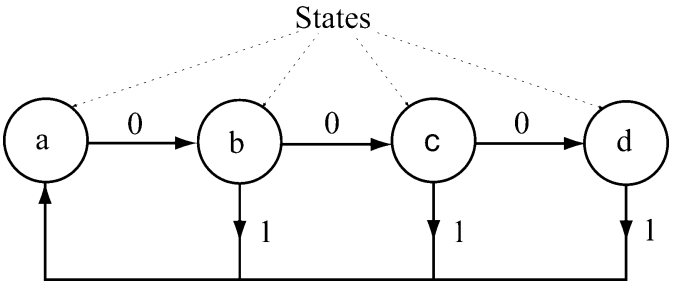

Figure 15. The finite-state machine for the ((1, 3)) RLL code.

by

$$C(d,k) = n \to \infty \lim \frac{1}{n} \log_2[M(n)] \tag{10}$$

For a fixed n, the ratio on the right-hand side of equation 10is called the rate of the code (which is the fraction of information bits per transmitted bit). It can be shown that the rate of the code is monotonically nondecreasing in n. Thus, the capacity of the code is the largest achievable rate. Shannon (14,15) has shown that the capacity of a FSM (the *(d, k)* code is just an example) is given by

$$C(d,k) = \log_2(\lambda_{\max})$$

where $\lambda_{\max}$ is the largest real eigenvalue of the adjacency matrix of the FSM. As an example, the eigenvalues of the adjacency matrix for the (1,3) code are 1.4656, -1.0000, $-0.2328 + 0.7926$i, and $-0.2328 - 0.7926$i The largest real eigenvalue is 1.4656, and thus, the capacity of the code is $\log_2(1.4656) = 0.5515$. For an excellent overview of information theory, including Shannon's result above, consult Reference (16).

The fact that an FSM is found that produces sequences satisfying the necessary constraints does not automatically imply that a code has been constructed. The problem of assigning information bits to encoded bits still exists. The problem of constructing such codes from their FSM representation has been studied by Adler et al. (17). An excellent tutorial paper on the topic can be found in Reference (18). Practical examples of applying the results of Reference (17) are, for example, in References (19) and (20). Another important class of codes that shapes the spectrum of the transmitted data and achieves a coding gain in the process is the class of matched spectral null (MSN) codes. The interested reader is referred to the paper by Karabed and Siegel (21)for more details.

Yet, another, very important class of modulation signals includes those signals that combine coding and modulation for improved performance. These combined modulation and coding techniques and, in particular, trellis-coded modulation (TCM) became better known from the breakthrough paper of Unger-boeck (22). In contrast to previous classic coding techniques that separate the coding and modulation problems, TCM achieves a coding gain (i.e., improved performance) without expanding bandwidth. It is thus very appealing in band limited applications, such as telephone modems, where it has been widely employed.

[1] This assumption is not as easy to justify when the receiver moves relative to the transmitter, because of the frequency offset caused by the Doppler effect.

BIBLIOGRAPHY

1. Proakis J.; Salehi M. *Communication Systems Engineering*; Prentice-Hall: Englewood Cliffs, NJ, 1994.
2. Gardner F. M. *Phaselock Techniques*; Wiley: New York, 1966.
3. Viterbi A. J. *Principles of Coherent Communications*; McGraw-Hill: New York, 1966.
4. Lindsey W. C. *Synchronization Systems in Communications*; Prentice-Hall: Englewood Cliffs, NJ, 1972.
5. Blanchard A. *Phase-Locked Loops: Application to Coherent Receiver Design*; Wiley: New York, 1976.
6. Stremler F. G. *Introduction to Communication Systems*; 3rd ed.; Addison-Wesley: Reading, MA, 1990.
7. Haykin S. *Communication Systems*, 3rd ed.; Wiley: New York, 1994.
8. Roden M. S. *Analog and Digital Communication Systems*; Prentice-Hall: Englewood Cliffs, NJ, 1991.
9. Couch L. W. *Modern Communication Systems*; Prentice-Hall: Englewood Cliffs, NJ, 1995.
10. Proakis J. *Digital Communications*, 3rd ed.; McGraw-Hill: New York, 1995.
11. Franaszek P. A. Sequence-State Coding for Digital Transmission. *Bell Syst. Tech. J.*, 1968, **47**, 143.
12. Calderbank A. R.; Mazo J. Spectral Nulls and Coding with Large Alphabets. *IEEE Commun. Mag.*December 1991.
13. Yoshida S.; Yajima Y. On the Relationship Between Encoding Automaton and the Power Spectrum of its Output Sequence. *Trans. IECE* 1976, **E59**,p. 97.
14. Shannon C. E. A Mathematical Theory of Communication. *Bell Syst. Tech. J.* 1948, **27**,pp 379–423.
15. Shannon C. E. A Mathematical Theory of Communication. *Bell Syst. Tech. J.* 1948, **27**,pp. 623–656.
16. Cover T. M.; Thomas J. A. Elements of Information Theory; Wiley Interscience: New York, 1991.
17. Adler R. L.; Coppersmith D.; Hassner M. Algorithms for Sliding Block Codes. *IEEE Trans. Inform. Theory* 1983, **IT-29**,pp. 5–22.
18. Marcus B. H.; Siegel P. H.; Wolf J. K. Finite-State Modulation Codes for Data Storage. *IEEE J. Select. Areas Commun.* 1992, **10**,pp. 5–37.
19. Calderbank A. R.; Georghiades C. N. Synchronizable Codes for the Optical OPPM Channel. *IEEE Trans. Inform. Theory* 1994, **40**,pp. 1097–1107.
20. Soljanin E.; Georghiades C. N. Coding for Two-Head Recording Systems. *IEEE Trans. Inform. Theory* 1995, **41**,pp. 747–755.

21. Karabed R.; Siegel P. Matched-Spectral Null Codes for Partial Response Channels. *IEEE Trans. Inform. Theory* 1991, **IT-37**,pp. 818–855.
22. Ungerboeck G. Channel Coding with Multilevel/Phase Signals. *IEEE Trans. Inform. Theory* 1982, **IT-28**,pp. 55–67.

COSTAS N. GEORGHIADES
Texas A&M University