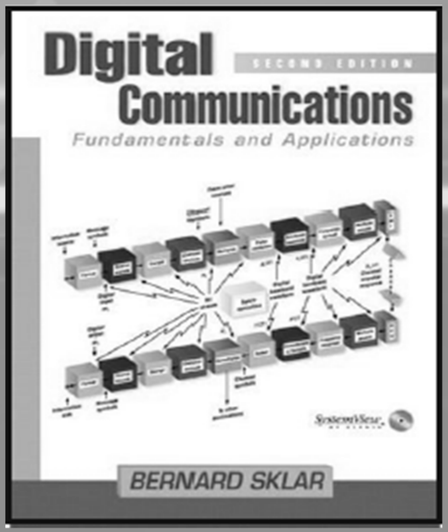


ENE 467
Digital Communications

TEACHING BY

ASST. PROF. SUWAT PATTARAMALAI, PH.D.

9. Modulation and Coding Trade-Off

- Outcome
  - Can explain goals of communication system designer
  - Can explain error probability plane, Nyquist minimum bandwidth, Shannon-Hartley Capacity Theorem, bandwidth efficiency plane
  - Can explain Modulation and Coding trade off
  - Defining, designing, and evaluating digital communication systems
  - Bandwidth efficient modulation
  - Modulation and coding for bandlimited channels

Goals of the communications system designer

GOALS OF THE COMMUNICATIONS SYSTEM DESIGNER

System trade-offs are fundamental to all digital communication designs. The goals of the designer may include any of the following (1) to maximize transmission bit rate R; (2) to minimize probability of bit error P_B; (3) to minimize required power, or equivalently, to minimize required bit energy to noise power spectral density E_b/N_0; (4) to minimize required system bandwidth W; (5) to maximize system utilization, that is, to provide reliable service for a maximum number of users with minimum delay and with maximum resistance to interference; and (6) to minimize system complexity, computational load, and system cost. A system designer may seek to achieve all these goals simultaneously. However, goals 1 and 2 are clearly in conflict with goals 3 and 4; they call for simultaneously maximizing R, while minimizing P_B, E_b/N_0, and W. There are several constraints and theoretical limitations that necessitate the trading off of any one system requirement with each of the others:

The Nyquist theoretical minimum bandwidth requirement
The Shannon-Hartley capacity theorem (and the Shannon limit)
Government regulations (e.g., frequency allocations)
Technological limitations (e.g., state-of-the-art components)
Other system requirements (e.g., satellite orbits)

Some of the realizable modulation and coding trade-offs can best be viewed as a change in operating point on one of two performance planes. These planes will be referred to as the error probability plane and the bandwidth efficiency plane, and they are described in the following sections.

ERROR PROBABILITY PLANE

Figure 9.1 illustrates the family of P_B versus E_b/N_0 curves for the coherent detection of orthogonal signaling (Figure 9.1a) and multiple phase signaling (Figure 9.1b). The modulator uses one of its $M = 2^k$ waveforms to represent each k-bit sequence, where M is the size of the symbol set. Figure 9.1a illustrates the potential bit error improvement with orthogonal signaling as k (or M) is increased. For orthogonal signal sets, such as orthogonal frequency shift keying (FSK) modulation, increasing the size of the symbol set can provide an improvement in P_B, or a reduction in the E_b/N_0 required, at the cost of increased bandwidth. Figure 9.1b illustrates potential bit error degradation with nonorthogonal signaling as k (or M) increases. For nonorthogonal signal sets, such as multiple phase shift keying (MPSK) modulation, increasing the size of the symbol set can reduce the bandwidth requirement, but at the cost of a degraded P_B, or an increased E_b/N_0 requirement. We shall refer to these families of curves (Figure 9.1a or b) as *error probability performance curves,* and to the plane on which they are plotted as an *error probability plane.* Such a plane describes the locus of operating points available for a particular type of modulation and coding. For a given system information rate, each curve in the plane can be associated with a different fixed minimum required bandwidth; therefore, the set of curves can be termed *equibandwidth curves.* As the curves move in the direction of the ordinate, the required transmission bandwidth increases; as the curves move in the opposite direction, the required bandwidth decreases. Once a modulation and coding scheme and an available E_b/N_0 are determined, system operation is characterized by a particular point in the error probability plane. Possible trade-offs can be viewed as changes in the operating point on one of the curves or as changes in the operating point from one curve to another curve of the family. These trade-offs are seen in Figure 9.1a and b as changes in the system operating point in the direction shown by the arrows. Movement of the operating point along line 1, between points a and b, can be viewed as trading off between P_B and E_b/N_0 performance (with W fixed). Similarly, movement along line 2, between points c and d, is seen as trading P_B versus W (with E_b/N_0 fixed). Finally, movement along line 3, between points e and f, illustrates trading W versus E_b/N_0 (with P_B fixed). Movement along line 1 is effected by increasing or decreasing the available E_b/N_0. This can be achieved, for example, by increasing transmitter power, which means that the trade-off might be accomplished simply by "turning a knob," even after the system is configured. However, the other trade-offs (movement along line 2 or line 3) involve some changes in the system modulation or coding scheme, and therefore need to be accomplished during the system design phase. The advent of *software radios* [1] will even allow changes to a system's modulation and coding by programmable means.

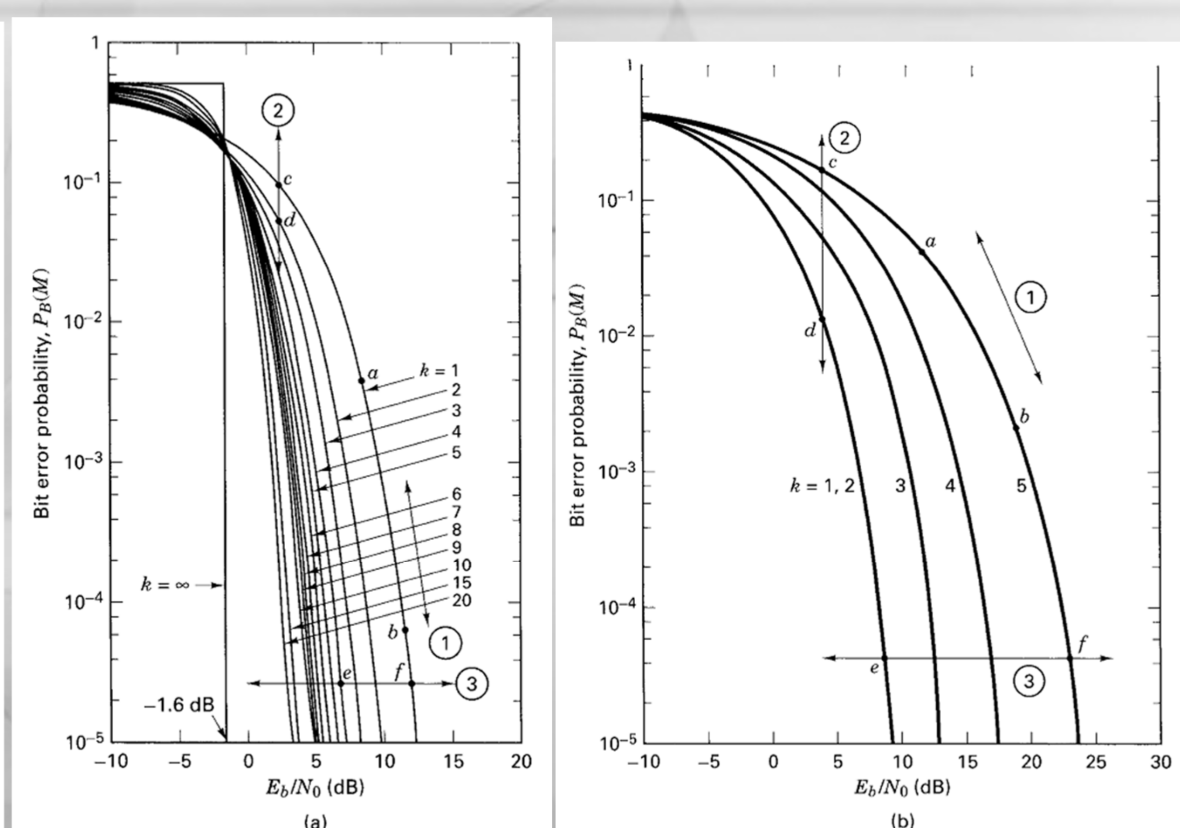


Figure 9.1 Bit error probability versus E_b/N_0 for coherently detected M-ary signaling. (a) Orthogonal signaling. (b) Multiple phase signaling.

Nyquist Minimum Bandwidth

NYQUIST MINIMUM BANDWIDTH

Every realizable system having some nonideal filtering will suffer from intersymbol interference (ISI)—the tail of one pulse spilling over into adjacent symbol intervals so as to interfere with correct detection. Nyquist [2] showed that the theoretical minimum bandwidth (Nyquist bandwidth) needed for the baseband transmission of R_s symbols per second without ISI is $R_s/2$ hertz. This is a basic theoretical constraint, limiting the designer's goal to expend as little bandwidth as possible. (See Section 3.3.) In practice, the Nyquist minimum bandwidth is expanded by about 10% to 40%, because of the constraints of real filters. Thus, *typical* baseband digital communication throughput is reduced from the ideal 2 symbols/s/Hz to the range of about 1.8 to 1.4 symbols/s/Hz. From its set of M symbols, the modulation or coding system assigns to each symbol a k-bit meaning, where $M = 2^k$. Thus, the number of bits per symbol can be expressed as $k = \log_2 M$, and the data rate or bit rate R must be k times faster than the symbol rate R_s, as expressed by the basic relationship

$$R = kR_s \quad \text{or} \quad R_s = \frac{R}{k} = \frac{R}{\log_2 M} \tag{9.1}$$

For signaling at a fixed symbol rate, Equation (9.1) shows that, as k is increased, the data rate R is increased. In the case of MPSK, increasing k, thereby results in an increased bandwidth efficiency R/W measured in bits/s/Hz. For example, movement along line 3, from point e to point f in Figure 9.1b, represents trading E_b/N_0 for a reduced bandwidth requirement. In other words, with the same system bandwidth, one can transmit MPSK signals at an increased date rate and hence at an increased R/W.

Example

Example 9.1 Digital Modulation Schemes Fall into One of Two Classes

In some sense, all digital modulation schemes fall into one of two classes with opposite behavior characteristics. The first class constitutes orthogonal signaling, and its error performance follows the curves shown in Figure 9.1a. The second class constitutes nonorthogonal signaling (the constellation of signal phasors can be depicted on a plane). Figure 9.1b illustrates an MPSK example of such nonorthogonal signaling. However, any phase/amplitude modulation (e.g., QAM) falls into this second class. In the context of Figure 9.1, answer the following questions:

(a) Does error-performance improve or degrade with increasing M, for M-ary signaling?
(b) The choices available in digital communications almost always involves a trade-off. If error-performance improves, what price must we pay?
(c) If error-performance degrades, what benefit is exhibited?

Solution

(a) When examining Figure 9.1, we see that error-performance improvement or degradation depends upon the class of signaling in question. Consider the orthogonal signaling in Figure 9.1a, where error-performance improves with increased k or M. Recall that there are only two fair ways to compare error-performance with such curves. A vertical line can be drawn through some fixed value of E_b/N_0, and as k or M is increased, it is seen that P_B is reduced. Or, a horizontal line can be drawn through some fixed P_B requirement, and as k or M is increased, it is seen that the E_b/N_0 requirement is reduced. Similarly, it can be seen that the curves in Figure 9.1b, for nonorthogonal signaling such as MPSK, behave in the opposite fashion. Error-performance degrades as k or M is increased.

(b) In the case of orthogonal signaling, where error performance improves with increasing k of M, what is the cost? In terms of the orthogonal signaling we are most familiar with, MFSK, when $k = 1$ and $M = 2$ there are two tones in the signaling set. When $k = 2$ and $M = 4$, there are four tones in the set. When $k = 3$ and $M = 8$, there are eight tones, and so forth. With MFSK, only one tone is sent during each symbol time, but the available transmission bandwidth consists of the entire set of tones. Hence, as k or M is increased, it should be clear that the cost of improved error-performance is an expansion of required bandwidth.

(c) In the case of nonorthogonal signaling, such as MPSK or QAM, where error-performance degrades as k or M is increased, one might rightfully guess that the tradeoff will entail a reduction in the required bandwidth. Consider the following example. Suppose we require a data rate of $R = 9600$ bit/s. And, suppose that the modulation chosen is 8-ary PSK. Then, using Equation (9.1), we find that the symbol rate is

$$R_s = \frac{R}{\log_2 M} = \frac{9600 \text{ bit/s}}{3 \text{ bit/symbol}} = 3200 \text{ symbol/s}$$

If we decide to use 16-ary PSK for this example, the symbol rate would then be

$$R_s = = \frac{9600 \text{ bit/s}}{4 \text{ bit/symbol}} = 2400 \text{ symbol/s}$$

If we continue in this direction and use 32-ary PSK, the symbol rate becomes

$$R_s = = \frac{9600 \text{ bit/s}}{5 \text{ bit/symbol}} = 1920 \text{ symbol/s}$$

Do you see what happens as the operating point in Figure 9.1b is moved along a horizontal line from the $k = 3$ curve to the $k = 4$ curve, and finally to the $k = 5$ curve? For a given data rate and bit-error probability, each such movement allows us to signal at a slower rate. Whenever you hear the words, "slower signaling rate," that is tantamount to saying that the transmission bandwidth can be reduced. Similarly, any case of increasing the signaling rate, corresponds to a need for increasing the transmission bandwidth.

Shannon-Hartley Capacity Theorem

SHANNON–HARTLEY CAPACITY THEOREM

Shannon [3] showed that the system capacity C of a channel perturbed by additive white Gaussian noise (AWGN) is a function of the average received signal power S, the average noise power N, and the bandwidth W. The capacity relationship (Shannon–Hartley theorem) can be stated as

$$C = W \log_2\left(1 + \frac{S}{N}\right) \tag{9.2}$$

When W is in hertz and the logarithm is taken to the base 2, as shown, the capacity is given in bits/s. It is theoretically possible to transmit information over such a channel at any rate R, where $R \le C$, with an *arbitrarily small* error probability by using a sufficiently complicated coding scheme. For an information rate $R > C$, it is not possible to find a code that can achieve an arbitrarily small error probability. Shannon's work showed that the values of S, N, and W *set a limit on transmission rate, not on error probability*. Shannon [4] used Equation (9.2) to graphically exhibit a bound for the achievable performance of practical systems. This plot, shown in Figure 9.2, gives the normalized channel capacity C/W in bits/s/Hz as a function of the channel signal-to-noise ratio (SNR). A related plot, shown in Figure 9.3, indicates the normalized channel bandwidth W/C in Hz/bits/s as a function of SNR in the channel. Figure 9.3 is sometimes used to illustrate the power-bandwidth trade-off inherent in the ideal channel. However, it is not a pure trade-off [5] because the detected noise power is proportional to bandwidth:

$$N = N_0 W \tag{9.3}$$

Substituting Equation 9.3 into Equation 9.2 and rearranging terms yields

$$\frac{C}{W} = \log_2\left(1 + \frac{S}{N_0 W}\right) \tag{9.4}$$

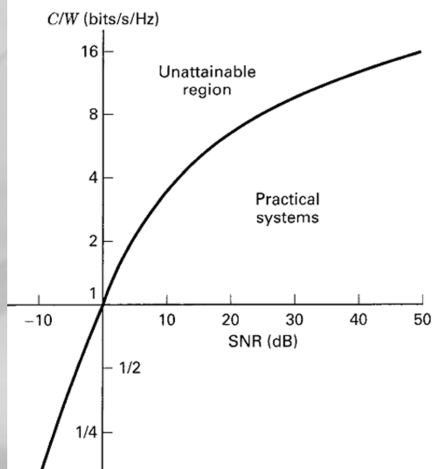


Figure 9.2 Normalized channel capacity versus channel SNR.

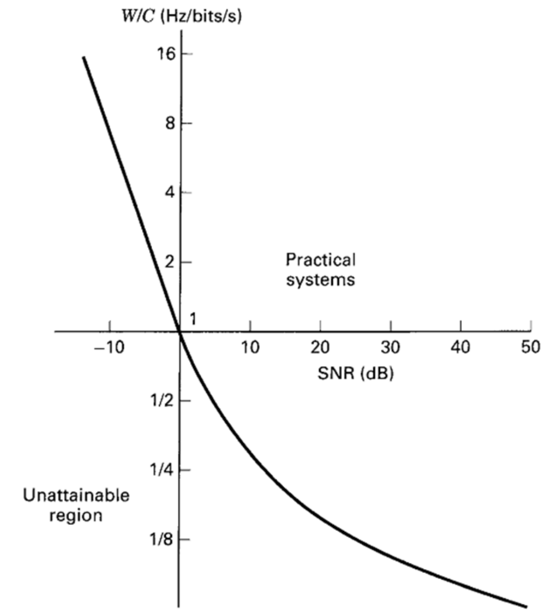


Figure 9.3 Normalized channel bandwidth versus channel SNR.

For the case where transmission bit rate is equal to channel capacity, $R = C$, we can use the identity presented in Equation 3.30 to write

$$\frac{S}{N_0 C} = \frac{E_b}{N_0} \tag{9.5}$$

Hence, we can modify Equation 9.4 as follows:

$$\frac{C}{W} = \log_2\left[1 + \frac{E_b}{N_0}\left(\frac{C}{W}\right)\right] \tag{9.6a}$$

$$2^{C/W} = 1 + \frac{E_b}{N_0}\left(\frac{C}{W}\right) \tag{9.6b}$$

$$\frac{E_b}{N_0} = \frac{W}{C}\left(2^{C/W} - 1\right) \tag{9.6c}$$

Shannon-Hartley Capacity Theorem

Figure 9.4 is a plot of W/C versus E_b/N_0 in accordance with Equation (9.6c). The asymptotic behavior of this curve as $C/W \to 0$ (or $W/C \to \infty$) is discussed in the next section.

9.4.1 Shannon Limit

There exists a limiting value of E_b/N_0 below which there can be no error-free communication at any information rate. Using the identity

$$\lim_{x \to 0} (1 + x)^{1/x} = e$$

we can calculate the limiting value of E_b/N_0 as follows: Let

$$x = \frac{E_b}{N_0}\left(\frac{C}{W}\right)$$

Then, from Equation (9.6a),

$$\frac{C}{W} = x \log_2 (1 + x)^{1/x}$$

and

$$1 = \frac{E_b}{N_0} \log_2 (1 + x)^{1/x}$$

In the limit, as $C/W \to 0$, we get

$$\frac{E_b}{N_0} = \frac{1}{\log_2 e} = 0.693 \qquad (9.7)$$

or, in decibels,

$$\frac{E_b}{N_0} = -1.6 \text{ dB}$$

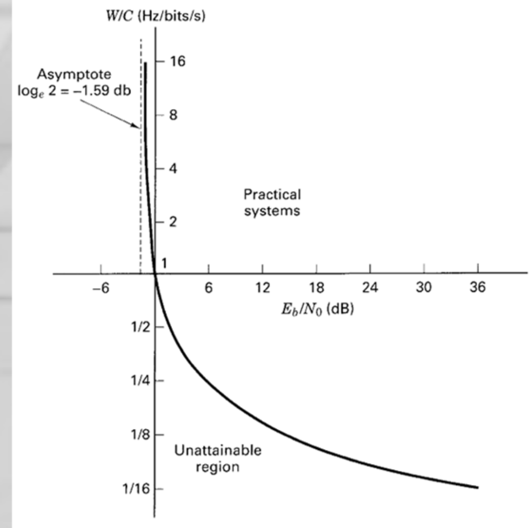


Figure 9.4 Normalized channel bandwidth versus channel E_b/N_0.

This value of E_b/N_0 is called the *Shannon limit.* On Figure 9.1a the Shannon limit is the P_B versus E_b/N_0 curve corresponding to $k \to \infty$. The curve is discontinuous, going from a value of $P_B = \frac{1}{2}$ to $P_B = 0$ at $E_b/N_0 = -1.6$ dB. It is not possible in practice to reach the Shannon limit, because as k increases without bound, the bandwidth requirement and the implementation complexity increases without bound. Shannon's work provided a theoretical proof for the existence of codes that could improve the P_B performance, or reduce the E_b/N_0 required, from the levels of the uncoded binary modulation schemes to levels approaching the limiting curve. For a bit error probability of 10^{-5}, binary phase-shift-keying (BPSK) modulation requires an E_b/N_0 of 9.6 dB (the optimum uncoded binary modulation). Therefore, for this case, Shannon's work promised the existence of a theoretical performance improvement of 11.2 dB over the performance of optimum uncoded binary modulation, through the use of coding techniques. Today, most of that promised improvement (as much as 10 dB) is realizable with turbo codes (see Section 8.4). Optimum system design can best be described as a search for rational compromises or trade-offs among the various constraints and conflicting goals. The modulation and coding trade-off, that is, the selection of modulation and coding techniques to make the best use of transmitter power and channel bandwidth, is important, since there are strong incentives to reduce the cost of generating power and to conserve the radio spectrum.

Shannon-Hartley Capacity

9.4.2 Entropy

To design a communications system with a specified message handling capability, we need a metric for measuring the information content to be transmitted. Shannon [3] developed such a metric, H, called the entropy of the message source (having n possible outputs). *Entropy* is defined as the average amount of information per source output and is expressed by

$$H = -\sum_{i=1}^{n} p_i \log_2 p_i \quad \text{bits/source output} \tag{9.8}$$

where p_i is the probability of the ith output and $\Sigma p_i = 1$. In the case of a binary message or a source having only two possible outputs, with probabilities p and $q = (1 - p)$, the entropy is written

$$H = -(p \log_2 p + q \log_2 q) \tag{9.9}$$

and is plotted versus p in Figure 9.5.

The quantity H has a number of interesting properties, including the following:

1. When the logarithm in Equation (9.8) is taken to the base 2, as shown, the unit for H is average bits per event. The unit *bit*, here, is a measure of *information content* and is not to be confused with the term "bit," meaning "binary digit."
2. Ther term "entropy" has the same uncertainty connotation as it does in certain formulations of statistical mechanics. For the information source with two equally likely possibilities (e.g., the flipping of a fair coin), it can be seen from Figure 9.5 that the uncertainty in the event, and hence the average information content, is maximum. As the probabilities depart from the equally likely case, the average information content decreases. In the limit, when one of the probabilities goes to zero, H also goes to zero. We know the result before the event happens, so the result conveys no additional information.

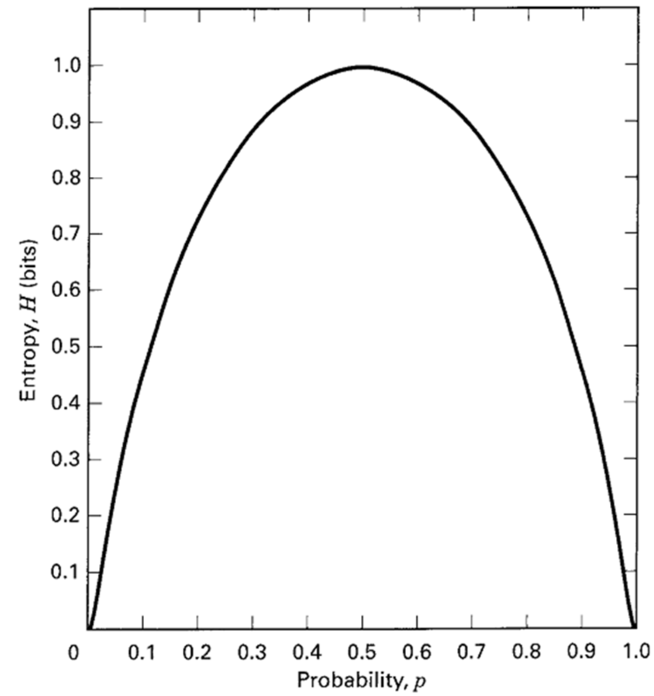


Figure 9.5 Entropy versus probability (two events).

3. To illustrate that information content is related to a priori probability (if the a priori message probability at the receiver is zero or one, we need not send the message), consider the following example: At the end of her nine-month pregnancy, a woman enters the delivery room of a local hospital to give birth. Her husband waits anxiously in the waiting room. After some time, a physician approaches the husband and says: "Congratulations, you are the father of a child." How much information has the physician given the father *beyond the medical outcome*? Almost none; the father has known with virtual certainty that a child was forthcoming. Had the physician said, "you are the father of a boy" or "you are the father of a girl," he would have transmitted 1 bit of information, since there was a 50% chance that the child could have been a boy or a girl.

Example

Example 9.2 Average Information Content in the English Language

(a) Calculate the average information in bits/character for the English language, assuming that each of the 26 characters in the alphabet occurs with equal likelihood. Neglect spaces and punctuation.

(b) Since the alphabetic characters do not appear with equal frequency in the English language (or any other language), the answer to part (a) will represent an upper bound on average information content per character. Repeat part (a) under the assumption that the alphabetic characters occur with the following probabilities:

$p = 0.10$: for the letters a, e, o, t

$p = 0.07$: for the letters h, i, n, r, s

$p = 0.02$: for the letters c, d, f, l, m, p, u, y

$p = 0.01$: for the letters b, g, j, k, q, v, w, x, z

Solution

(a)
$$H = -\sum_{i=1}^{26} \frac{1}{26} \log_2 \left(\frac{1}{26}\right) = 4.7 \text{ bits/character}$$

(b)
$$H = -(4 \times 0.1 \log_2 0.1 + 5 \times 0.07 \log_2 0.07 + 8 \times 0.02 \log_2 0.02 + 9 \times 0.01 \log_2 0.01) = 4.17 \text{ bits/character}$$

If we want to express the 26 letters of the alphabet with some binary-digit coding scheme, we generally need five binary digits for each character. Example 9.2 demonstrates that there may be a way to encode the English language with a fewer number of binary digits per character, *on the average*, by exploiting the fact that the average amount of information contained within each character is less than 5 bits. The subject of source coding, which deals with this exploitation, is treated in Chapter 13.

Shannon-Hartley Capacity Theorem

9.4.3 Equivocation and Effective Transmission Rate

Suppose that we are transmitting information at a rate of 1000 binary symbols/s over a binary symmetric channel (defined in Section 6.3.1), and that the a priori probability of transmitting either a one or a zero is equally likely. Suppose also that the noise in the channel is so great that the probability of receiving a one is $\frac{1}{2}$, whatever was transmitted, and similarly for receiving a zero. In such a case, half the received symbols would be correct *due to chance alone,* and the system might appear to be providing 500 bits/s while actually no information is being received at all. Equally "good" reception could be obtained by dispensing with the channel entirely and "flipping a coin" within the receiver. The proper correction to apply to the amount of information transmitted is the amount of information that is lost in the channel. Shannon [3] uses a correction factor called *equivocation* to account for the uncertainty in the received signal. Equivocation is defined as the *conditional entropy* of the message X, given Y, or

$$\begin{aligned} H(X|Y) &= -\sum_{X,Y} P(X,Y)\log_2 P(X|Y) \\ &= -\sum_{Y} P(Y) \sum_{X} P(X|Y)\log_2 P(X|Y) \end{aligned} \tag{9.10}$$

where X is the transmitted source message, Y is the received signal, $P(X, Y)$ is the joint probability of X and Y, and $P(X|Y)$ is the conditional probability of X given Y. Equivocation can be thought of as the uncertainty that message X was sent, having received Y. For an *error-free channel,* $H(X|Y) = 0$, because having received Y, there is complete certainty about the message X. However, for a channel with a nonzero probability of symbol error, $H(X|Y) > 0$, because the channel introduces uncertainty. Consider a binary sequence, X, where the a priori source probabilities are $P(X=1) = P(X=0) = \frac{1}{2}$, and where, on the average, the channel produces one error in a received sequence of 100 bits ($P_B = 0.01$). Using Equation (9.10), the equivocation $H(X|Y)$ is expressed as

$$\begin{aligned} H(X|Y) &= -[(1-P_B)\log_2(1-P_B) + P_B\log_2 P_B] \\ &= -(0.99\log_2 0.99 + 0.01\log_2 0.01) \\ &= 0.081 \text{ bit/received symbol} \end{aligned}$$

Thus, the channel introduces 0.081 bit of uncertainty to each received symbol.

Shannon showed that the average effective information content H_{eff} at the receiver is obtained by subtracting the equivocation from the entropy of the source. Therefore,

$$H_{\text{eff}} = H(X) - H(X|Y) \tag{9.11}$$

For a system transmitting equally likely binary symbols, the entropy $H(X)$ is 1 bit/symbol. When the symbols are received with $P_B = 0.01$, the equivocation is 0.081 bit/received symbol as was calculated above. Then, using Equation (9.11), the effective entropy of the received signal H_{eff} is

$$H_{\text{eff}} = 1 - 0.081 = 0.919 \text{ bit/received symbol}$$

Thus, if $R = 1000$ binary symbols transmitted per second, for example, the effective information bit rate R_{eff} can then be expressed as

$$\begin{aligned} R_{\text{eff}} &= RH_{\text{eff}} \\ &= 1000 \text{ symbols/s} \times 0.919 \text{ bit/symbol} = 919 \text{ bits/s} \end{aligned} \tag{9.12}$$

Notice that in the extreme case, where $P_B = 0.5$,

$$\begin{aligned} H(X|Y) &= -(0.5\log_2 0.5 + 0.5\log_2 0.5) \\ &= 1 \text{ bit/symbol} \end{aligned}$$

and, applying Equations (9.12) and (9.11) to the $R = 1000$ symbols/s example, yields

$$R_{\text{eff}} = 1000 \text{ symbols/s}\,(1-1) = 0 \text{ bit/s}$$

as should be expected.

Example

Example 9.3 Apparent Contradiction in the Shannon Limit

Plots of P_B versus E_b/N_0 typically display a smooth increase of P_B as E_b/N_0 is decreased. For example, the bit error probability for the curves in Figure 9.1 shows P_B *tending* to 0.5 in the limit as E_b/N_0 approaches zero. Thus there is apparently always a nonvanishing information rate, regardless of how small E_b/N_0 becomes. This *appears to contradict* the Shannon limit of $E_b/N_0 = -1.6$ dB, below which no error-free information rate can be supported per unit bandwidth, or below which even an infinite bandwidth cannot support a finite information rate (see Figure 9.4).

(a) Suggest a way of resolving the apparent contradiction.
(b) Show how Shannon's equivocation correction can resolve it for a binary PSK system where the source has an entropy of 1 bit/symbol. Consider that the operating point on Figure 9.1b corresponds to $E_b/N_0 = 0.1$ (−10 dB).

Solution

(a) The value of E_b, traditionally used in link calculations for practical systems, is invariably the received signal energy per *transmitted symbol.* However, the meaning of E_b in Equation (9.6) is the signal energy per bit of *received information.* The information loss caused by the noisy channel must be taken into account to resolve the apparent contradiction.

(b) Following Equation (4.79) for BPSK, we write

$$P_B = Q(\sqrt{2E_b/N_0}) = Q(0.447)$$

where Q is defined in Equation (3.43) and tabulated in Table B.1. From the tabulation, P_B is found to be 0.33. Next, we solve for the equivocation and effective entropy:

$$H(X|Y) = -[(1 - P_B)\log_2(1 - P_B) + P_B \log_2 P_B]$$

$$= -(0.67 \log_2 0.67 + 0.33 \log_2 0.33)$$
$$= 0.915 \text{ bit/symbol}$$

$$H_{\text{eff}} = H(X) - H(X|Y)$$
$$= 1 - 0.915$$
$$= 0.085 \text{ bit/symbol}$$

Hence,

$$\left(\frac{E_b}{N_0}\right)_{\text{eff}} = \frac{(E_b/N_0) \text{ joules per symbol/watts per hertz}}{H_{\text{eff}} \text{ bits/symbol}}$$

$$= \frac{0.1}{0.085} = 1.176 \; \frac{\text{joules per bit}}{\text{watts/Hz}}$$

$$= 0.7 \text{ dB}$$

Thus, the effective value of E_b/N_0 is equal to 0.7 dB per received information bit, which is well above Shannon's limit of – 1.6 dB.

9.5 BANDWIDTH-EFFICIENCY PLANE

Using Equation (9.6), we can plot normalized channel bandwidth W/C in Hz/bits/s versus E_b/N_0, as shown in Figure 9.4. Here, with the abscissa taken as E_b/N_0, we see the *true power-bandwidth trade-off* at work. It can be shown [5] that well-designed systems tend to operate near the "knee" of this power-bandwidth trade-off curve for the ideal ($R = C$) channel. Actual systems are frequently within 10 dB or less of the performance of the ideal. The existence of the knee means that systems seeking to reduce the channel bandwidth they occupy or to reduce the signal power they require must make an increasingly unfavorable exchange in the other parameter. For example, from Figure 9.4, an ideal system operating at an E_b/N_0 of 1.8 dB and using a normalized bandwidth of 0.5 Hz/bits/s would have to increase E_b/N_0 to 20 dB to reduce the bandwidth occupancy to 0.1 Hz/bits/s. Trade-offs in the other direction are similarly inequitable.

Using Equation (9.6c), we an also plot C/W versus E_b/N_0. This relationship is shown plotted on the R/W versus E_b/N_0 plane in Figure 9.6. We shall denote this plane as the *bandwidth-efficiency plane.* The ordinate R/W is a measure of how much data can be communicated in a specified bandwidth within a given time; it therefore reflects how efficiently the bandwidth resource is utilized. The abscissa is E_b/N_0, in units of decibels. For the case in which $R = C$ in Figure 9.6, the curve represents a boundary that separates a region characterizing practical communication systems from a region where such communication systems are not theoretically possible. Like Figure 9.2, the bandwidth-efficiency plane in Figure 9.6 sets the limiting performance that can be achieved by practical systems. Since the abscissa in Figure 9.6 is E_b/N_0 rather than SNR, Figure 9.6 is more useful for comparing digital communication modulation and coding trade-offs than is Figure 9.2. Note that Figure 9.6 illustrates bandwidth efficiency versus E_b/N_0 for single-carrier systems. For multiple-carrier systems, bandwidth efficiency is also a function of carrier spacing (which depends on the modulation type). The trade-off becomes how closely can the carriers be spaced (thereby improving bandwidth efficiency) without suffering an unacceptable amount of adjacent channel interference (ACI).

Bandwidth Efficiency

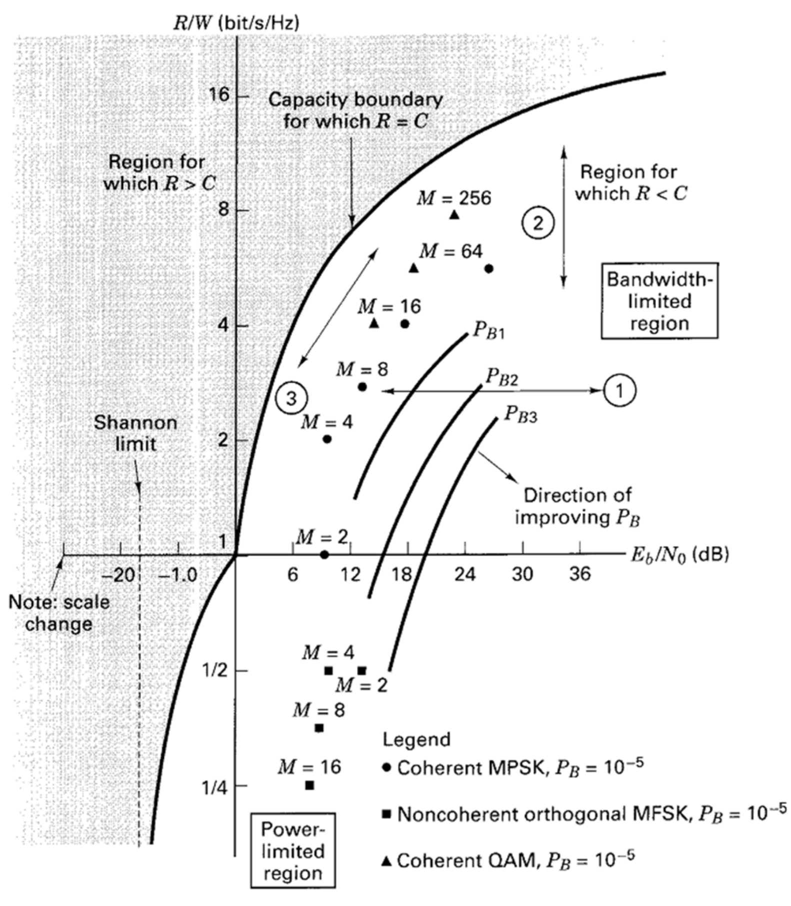


Figure 9.6 Bandwidth-efficiency plane.

9.5.1 Bandwidth Efficiency of MPSK and MFSK Modulation

On the bandwidth-efficiency plane of Figure 9.6 are plotted the operating points for coherent MPSK modulation at a bit error probability of 10^{-5}. We assume Nyquist (ideal rectangular) filtering at baseband, so that the minimum

double-sideband (DSB) bandwidth at an intermediate frequency (IF) is $W_{IF} = 1/T$, where T is the symbol duration. Thus using (Equation (9.1), the bandwidth efficiency is $R/W = \log_2 M$, where M is the symbol set size. For realistic channels and waveforms, the performance must be reduced to account for the bandwidth increase required to implement realizable filters. Notice that for MPSK modulation, R/W increases with increasing M. Notice also that the location of the MPSK points indicates that BPSK ($M = 2$) and quaternary PSK or QPSK ($M = 4$) require the same E_b/N_0. That is, for the same value of E_b/N_0, QPSK has a bandwidth efficiency of 2 bits/s/Hz, compared to 1 bit/s/Hz for BPSK. This unique features stems from the fact that QPSK is effectively a composite of two BPSK signals transmitted on orthogonal components of the carrier.

Also plotted on the bandwidth-efficiency plane of Figure 9.6 are the operating points for noncoherent orthogonal MFSK modulation, at a bit error probability of 10^{-5}. We assume that the IF transmission bandwidth is $W_{IF} = M/T$ (see Section 4.5.4.1), and thus using Equation (9.1), the bandwidth efficiency is $R/W = (\log_2 M)/M$. Notice that for MFSK modulation, R/W decreases with increasing M. Notice also that the position of the MFSK points indicates that BFSK ($M = 2$) and quaternary FSK ($M = 4$) have the same bandwidth efficiency, even though the former requires greater E_b/N_0 for the same error probability. The bandwidth efficiency varies with the modulation index (tone spacing in hertz divided by bit rate). Under the assumption that an equal increment of bandwidth is required for each MFSK tone the system uses, it can be seen that for $M = 2$, the bandwidth efficiency is 1 bit/s/2 Hz or $\frac{1}{2}$, and for $M = 4$, similarly, the R/W is 2 bits/s/4 Hz or $\frac{1}{2}$. Thus binary and 4-ary orthogonal FSK are curiously characterized by the same value of R/W.

Operating points for coherent quadrature amplitude modulation (QAM) are also plotted in Figure 9.6. Of the modulations shown, QAM is clearly the most bandwidth efficient; it is treated in greater detail in Section 9.8.3.

Bandwidth Efficiency

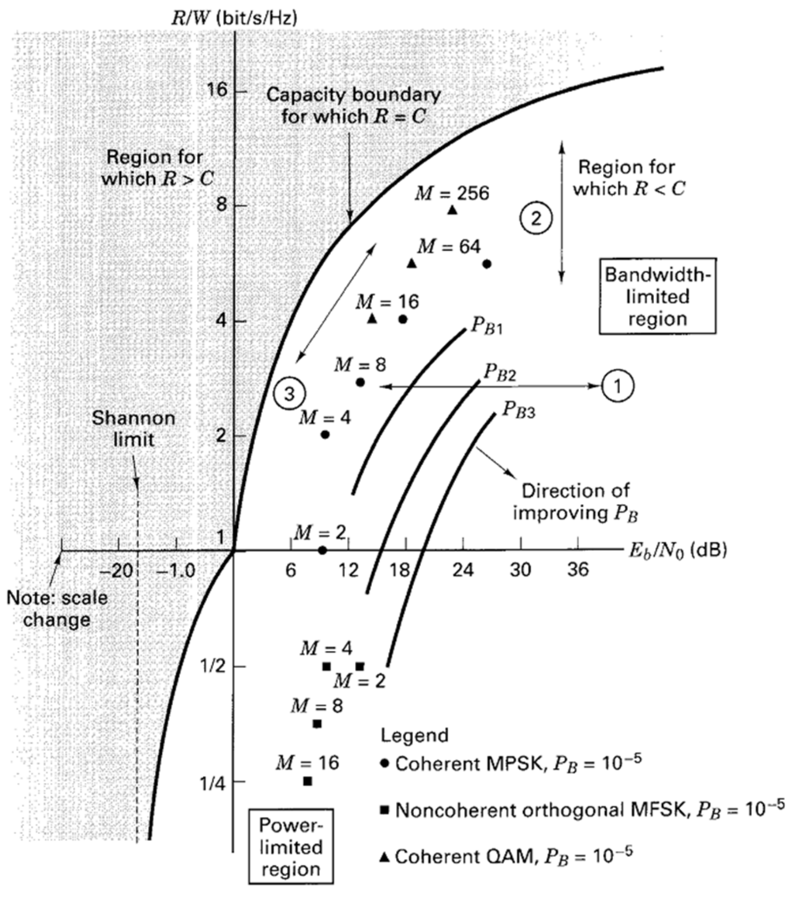


Figure 9.6 Bandwidth-efficiency plane.

9.5.2 Analogies Between Bandwidth-Efficiency and Error-Probability Planes

The bandwidth-efficiency plane in Figure 9.6 is analogous to the error-probability plane in Figure 9.1. The Shannon limit of the Figure 9.1 plane is analogous to the capacity boundary of the Figure 9.6 plane. The curves in Figure 9.1 were referred to as equibandwidth curves. In Figure 9.6, we can analogously describe equi-error-probability curves for various modulation and coding schemes. The curves, labeled P_{B1}, P_{B2}, and P_{B3}, are hypothetical constructions for some arbitrary modulation and coding scheme; the P_{B1} curve represents the largest error probability of the three curves, and the P_{B3} curve represents the smallest. The general direction in which the curves move for improved P_B is indicated on the figure.

Just as potential trade-offs among P_B, E_b/N_0, and W were considered for the error-probability plane, the same trade-offs can be considered on the bandwidth efficiency plane. The potential trade-offs are seen in Figure 9.6 as changes in operating point in the direction shown by the arrows. Movement of the operating point along line 1 can be viewed as trading P_B versus E_b/N_0, with R/W fixed. Similarly, movement along line 2 is seen as trading P_B versus W (or R/W), with E_b/N_0 fixed. Finally, movement along line 3 illustrates trading W (or R/W) versus E_b/N_0, with P_B fixed. In Figure 9.6, as in Figure 9.1, movement along line 1 can be effected by increasing or decreasing the available E_b/N_0. However, movement along line 2 or line 3 requires changes in the system modulation or coding scheme.

The two primary communications resources are the transmitted power and the channel bandwidth. In many communication systems, one of these resources may be more precious than the other, and hence most systems can be classified as either power limited or bandwidth limited. In *power-limited systems,* coding schemes can be used to save power at the expense of bandwidth, whereas in *bandwidth-limited systems,* spectrally efficient modulation techniques can be used to save bandwidth at the expense of power.

Modulation and Coding

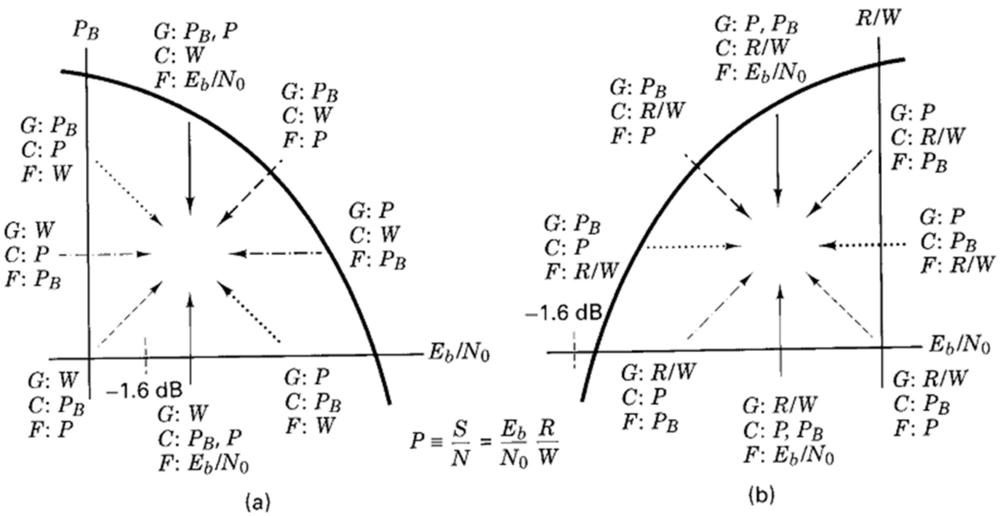


Figure 9.7 Modulation/coding trade-offs. (a) Error probability plane. (b) Bandwidth-efficiency plane.

Figure 9.7 is useful in pointing out analogies between the two performance planes, the error-probability plane of Figure 9.1 and the bandwidth-efficiency plane of Figure 9.6. Figure 9.7a and b represent the same planes as Figures 9.1 and 9.6, respectively. They have been redrawn as symmetrical by choosing appropriate scales. In each case the arrows and their labels describe the general effect of moving an operating point in the direction of the arrow by means of appropriate modulation and coding techniques. The notations *G, C,* and *F* stand for the trade-off considerations "*G*ained or achieved," "*C*ost or expended," and "*F*ixed or unchanged," respectively. The parameters being traded are P_B, *W*, *R/W*, and *P* (power or *S/N*). Just as the movement of an operating point toward the Shannon limit in Figure 9.7a can achieve improved P_B or reduced required transmitter power at the cost of bandwidth, so too movement toward the capacity boundary in Figure 9.7b can improve bandwidth efficiency at the cost of increased required power or degraded P_B.

Most often, these trade-offs are examined with a fixed P_B (constrained by the system requirement) in mind. Therefore, the most interesting arrows are those having bit error probability (marked F: P_B). There are four such arrows on Figure 9.7, two on the error probability plane and two on the bandwidth-efficiency plane. Arrows marked with the same pattern indicate correspondence between the two planes. System operation can be characterized by either of these two planes. The planes represent two ways of looking at some of the key system parameters; each plane highlights slightly different aspects of the overall design problem. The error probability plane tends to be most useful with *power-limited systems,* whereas when we move from curve to curve, the bandwidth requirements are only inferred, while the bit error probability is clearly displayed. The bandwidth-efficiency plane is generally more useful for examining *bandwidth-limited systems;* here, as we move from curve to curve, the bit-error probability is only inferred, but the bandwidth requirements are explicit.

The two system trade-off planes, error probability and bandwidth efficiency, have been presented *heuristically* with simple examples (orthogonal and multiple phase signaling) to provide some insight into the design issues of trading-off error probability, bandwidth, and power. The ideas are useful for *most modulation and coding schemes,* with the following caveat. For *some* codes or combined modulation and coding schemes, the performance curves *do not move as predictably* as those for the examples chosen here. The reason has to do with the error-correcting capability and bandwidth expansion features of the particular code. For example, the performance of coherent PSK combined with several codes was illustrated in Figure 6.22. Examine the curves characterizing the two BCH codes, (127, 64) and (127, 36). It should be clear from their relative positions that the (127, 64) code manifests *greater coding gain* than the (127, 36) code. This violates our expectations since, within the same block size, the latter code has greater redundancy (requires more bandwidth expansion) than the former. Also, in the area of trellis-coded modulation covered in Section 9.10, we consider codes that provide coding gain without any bandwidth expansion. Performance curves for such coding schemes will also behave differently from the curves of most modulation and coding schemes discussed so far.

Defining, Designing, and Evaluating Digital Communication Systems

This section is intended to serve as a "road map" for outlining typical steps that need to be considered in meeting the bandwidth, power, and error-performance requirements of a digital communication system. The criteria for choosing modulation and coding schemes, based on whether a system is bandwidth limited or power limited, are reviewed for several system examples. We will emphasize the subtle but straightforward relationships that exist when transforming from data-bits to channel-bits to symbols to chips.

The design of any digital communication system begins with a description of the channel (received power, available bandwidth, noise statistics and other impairments, such as fading), and a definition of the system requirements (data rate and error performance). Given the channel description, we need to determine design choices that best match the channel and meet the performance requirements. An orderly set of transformations and computations has evolved to aid in characterizing a system's performance. Once this approach is understood, it can serve as the format for evaluating most communication systems. In subsequent sections, we examine three system examples, chosen to provide a representative assortment: a bandwidth-limited uncoded system, a power-limited uncoded system, and a bandwidth-limited and power-limited coded system. In this section, we deal with real-time communication systems, where the term *coded* (or *uncoded*) refers to the presence (or absence) of error-correction coding schemes involving the use of *redundant bits* and expanded bandwidth.

Two primary communications resources are the *received power* and the *available transmission bandwidth.* In many communication systems, one of these resources may be more precious than the other, and hence most systems can be classified as either bandwidth limited or power limited. In bandwidth-limited systems, spectrally efficient modulation techniques can be used to save bandwidth at the expense of power, whereas in power-limited systems, power-efficient modulation techniques can be used to save power at the expanse of bandwidth. In both bandwidth- and power-limited systems, error-correction coding (often called *channel coding*) can be used to save power or to improve error performance at the expense of bandwidth. Trellis-coded modulation (TCM) schemes have been used to improve the error performance of bandwidth-limited channels without *any* increase in bandwidth [6]. These methods are considered in Section 9.10.

Defining, Designing, and Evaluating Digital Communication Systems

9.7.1 *M*-ary Signaling

For signaling schemes that process k bits at a time, the signaling is called M-ary (see Section 3.8). Each symbol in an M-ary alphabet can be related to a unique sequence of k bits, where

$$M = 2^k \quad \text{or} \quad k = \log_2 M \tag{9.13}$$

and where M is the size of the alphabet. In the case of digital transmission, the term *symbol* refers to the member of the M-ary alphabet that is transmitted during each symbol duration T_s. In order to transmit the symbol, it must be mapped onto an electrical voltage or current waveform. Because the waveform represents the symbol, the terms *symbol* and *waveform* are sometimes used interchangeably. Since one of M symbols or waveforms is transmitted during each symbol duration T_s, the date rate R can be expressed as

$$R = \frac{k}{T_s} = \frac{\log_2 M}{T_s} \quad \text{bit/s} \tag{9.14}$$

From Equation (9.14), we write that the *effective* duration T_b of each bit in terms of the symbol duration T_s or the symbol rate R_s is

$$T_b = \frac{1}{R} = \frac{T_s}{k} = \frac{1}{kR_s} \tag{9.15}$$

Then, using Equations (9.13) and (9.15), we can express the symbol rate R_s in terms of the bit rate R, as was presented earlier:

$$R_s = \frac{R}{\log_2 M} \tag{9.16}$$

From Equations (9.14) and (9.15), it is seen that any digital scheme that transmits $k = (\log_2 M)$ bits in T_s seconds, using a bandwidth of W Hz, operates at a bandwidth efficiency of

$$\frac{R}{W} = \frac{\log_2 M}{WT_s} = \frac{1}{WT_b} \quad \text{bits/s/Hz} \tag{9.17}$$

where T_b is the effective time duration of each data bit.

Defining, Designing, and Evaluating Digital Communication Systems

9.7.2 Bandwidth-Limited Systems

From Equation (9.17), it can be seen that any digital communication system will become more bandwidth efficient as its WT_b product is decreased. Thus, signals with small WT_b products are often used with bandwidth-limited systems. For example, the Global System for Mobile (GSM) Communication uses Gaussian minimum shift keying (GMSK) modulation having a WT_b product equal to 0.3 Hz/bit/s [7], where W is the 3-dB bandwidth of a Gaussian filter.

For uncoded bandwidth-limited systems, the objective is to maximize the transmitted information rate within the allowable bandwidth, at the expense of E_b/N_0 (while maintaining a specified value of bit-error probability P_B). On the bandwidth-efficiency plane of Figure 9.6 are plotted the operating points for coherent M-ary PSK (MPSK) at $P_B = 10^{-5}$. We shall assume Nyquist (ideal rectangular) filtering at baseband [2], so that, for MPSK, the required double-sideband (DSB) bandwidth at an intermediate frequency (IF) is related to the symbol rate by

$$W = \frac{1}{T_s} = R_s \qquad (9.18)$$

where T_s is the symbol duration and R_s is the symbol rate. The use of Nyquist filtering results in the *minimum* required transmission bandwidth that yields zero intersymbol interference; such ideal filtering gives rise to the name *Nyquist minimum bandwidth*. Note that the bandwidth of nonorthogonal signaling, such as MPSK or MQAM, does not depend on the density of the signaling points in the constellation but only on the speed of signaling. When a phasor is transmitted, the system cannot distinguish as to whether that signal arose from a sparse alphabet set or a dense alphabet set. It is this aspect of nonorthogonal signals that allows us to pack the signaling space densely and thus achieve improved bandwidth efficiency at the expense of power. From Equations (9.17) and (9.18), the bandwidth efficiency of MPSK modulated signals using Nyquist filtering can be expressed as

$$\frac{R}{W} = \log_2 M \quad \text{bits/s/Hz} \qquad (9.19)$$

The MPSK points plotted in Figure 9.6 confirm the relationship shown in Equation (9.19). Note that MPSK modulation is a bandwidth-efficient scheme. As M increases in value, R/W also increases. From Figure 9.6, it can be verified that MPSK modulation can achieve improved bandwidth efficiency at the cost of increased E_b/N_0. Many highly bandwidth-efficient modulation schemes have been investigated [8], but such schemes are beyond the scope of this book.

Two regions, the bandwidth-limited region and the power-limited region, are shown on the bandwidth-efficiency plane of Figure 9.6. Notice that the desirable trade-offs associated with each of these regions are not equitable. For the bandwidth-limited region, large R/W is desired; however, as E_b/N_0 is increased, the capacity boundary curve flattens out and ever-increasing amounts of additional E_b/N_0 are required to achieve improvement in R/W. A similar relationship is at work in the power-limited region. Here a savings in E_b/N_0 is desired, but the capacity boundary curve is steep; to achieve a small reduction in required E_b/N_0, requires a large reduction in R/W.

Defining, Designing, and Evaluating Digital Communication Systems

9.7.3 Power-Limited Systems

For the case of power-limited systems in which power is scarce but system bandwidth is available (e.g., a space communication link), the following trade-offs, which can be seen in Figure 9.1a, are possible: (1) improved P_B at the expense of bandwidth for a fixed E_b/N_0; or (2) reduction in E_b/N_0 at the expense of bandwidth for a fixed P_B. A "natural" modulation choice for a power-limited system is M-ary FSK (MFSK). Plotted on Figure 9.6 are the operating points for noncoherent orthogonal MFSK modulation at $P_B = 10^{-5}$. For such MFSK, the IF minimum bandwidth, assuming minimum tone spacing, is given by (see Section 4.5.4.1)

$$W = \frac{M}{T_s} = MR_s \tag{9.20}$$

where T_s is the symbol duration, and R_s is the symbol rate. With M-ary FSK, the required transmission bandwidth is expanded M-fold over binary FSK since there are M different orthogonal waveforms, each requiring a bandwidth of $1/T_s$. Thus, from Equations (9.17) and (9.20), the bandwidth efficiency of noncoherent MFSK signals can be expressed as

$$\frac{R}{W} = \frac{\log_2 M}{M} \quad \text{bits/s/Hz} \tag{9.21}$$

Notice the important difference between the *bandwidth efficiency* (R/W) of MPSK expressed in Equation (9.19) and that of MFSK expressed in Equation (9.21). With MPSK, R/W increases as the signal dimensionality M increases. With MFSK there are two mechanisms at work. The numerator shows the same increase in R/W with larger M, as in the case of MPSK. But the denominator indicates a decrease in R/W with larger M. As M grows larger, the denominator grows faster than the numerator, and thus R/W decreases. The MFSK points plotted in Figure 9.6 confirm the relationship shown in Equation (9.21), that orthogonal signaling such as MFSK is a bandwidth-expansive scheme. From Figure 9.6, it can be seen that MFSK modulation can be used for realizing a reduction in required E_b/N_0, at the cost of increased bandwidth.

It is important to emphasize that in Equations (9.18) and (9.19) for MPSK, and for all the MPSK points plotted in Figure 9.6, Nyquist (ideal rectangular) filtering has been assumed. Such filters are not realizable. For *realistic* channels and waveforms, the required transmission bandwidth must be *increased* in order to account for *realizable* filters.

In each of the examples that follow, we consider radio channels, disturbed *only* by additive white Gaussian noise (AWGN) and having no other impairments. For simplicity, the modulation choice is limited to *constant-envelope types*—either MPSK or noncoherent orthogonal MFSK. Thus, for an *uncoded* system, if the channel is bandwidth limited, MPSK is selected, and if the channel is power limited, MFSK is selected. Note that, *when error-correction coding is considered*, modulation selection is not so simple, because there exist coding techniques [9] that can provide power-bandwidth trade-offs more effectively than would be possible through the use of any M-ary modulation scheme.

Note that in the most general sense, M-ary signaling can be regarded as a *waveform-coding* procedure. That is, whenever we select an M-ary modulation technique instead of a binary one, we *in effect* have replaced the binary waveforms with *better* waveforms—either better for bandwidth performance (MPSK), or better for power performance (MFSK). Even though orthogonal MFSK signaling can be thought of as being a coded system (it can be described as a first-order Reed–Muller code [10]), we shall here restrict our use of the term *coded system* to refer only to those traditional error-correction codes using redundancies, such as block codes or convolutional codes.

Defining, Designing, and Evaluating Digital Communication Systems

9.7.4 Requirements for MPSK and MFSK Signaling

The basic relationship between the symbol (or waveform) transmission rate R_s and the data rate R was shown in Equation (9.16) to be

$$R_s = \frac{R}{\log_2 M}$$

Using this relationship together with Equations (9.18) through (9.21), and a given data rate of $R = 9600$ bit/s, Table 9.1 has been compiled [11]. The table is a summary of symbol rate, minimum bandwidth, and bandwidth efficiency for MPSK and noncoherent orthogonal MFSK, for the values of $M = 2, 4, 8, 16$, and 32. Also included in Table 9.1 are the required values of E_b/N_0 to achieve a bit-error probability of 10^{-5} for MPSK and MFSK for each value of M shown. These E_b/N_0 entries were computed using relationships that are presented later. The E_b/N_0 entries corroborate the trade-offs shown in Figure 9.6. As M increases, MPSK signaling provides more bandwidth efficiency at the cost of increased E_b/N_0, while MFSK signaling allows for a reduction in E_b/N_0 at the cost of increased bandwidth. The next three sections are presented in the context of examples taken from Table 9.1.

TABLE 9.1 Symbol Rate, Minimum Bandwidth, Bandwidth Efficiency, and Required E_b/N_0 for MPSK and Noncoherent Orthogonal MFSK Signaling at 9600 bit/s

M	k	R (bit/s)	R_s (symb/s)	MPSK Minimum Bandwidth (Hz)	MPSK R/W	MPSK E_b/N_0 (dB) $P_B = 10^{-5}$	Noncoherent Orthog MFSK Min Bandwidth (Hz)	MFSK R/W	MFSK E_b/N_0 (dB) $P_B = 10^{-5}$
2	1	9600	9600	9600	1	9.6	19,200	1/2	13.4
4	2	9600	4800	4800	2	9.6	19,200	1/2	10.6
8	3	9600	3200	3200	3	13.0	25,600	1/3	9.1
16	4	9600	2400	2400	4	17.5	38,400	1/4	8.1
32	5	9600	1920	1920	5	22.4	61,440	5/32	7.4

Defining, Designing, and Evaluating Digital Communication Systems

9.7.5 Bandwidth-Limited Uncoded System Example

Suppose we are given a bandwidth-limited AWGN radio channel with an available bandwidth of $W = 4000$ Hz. Also, consider that the link constraints (transmitter power, antenna gains, path loss, etc.) result in the ratio of received signal power to noise-power spectral density (P_r/N_0) being equal to 53 dB-Hz. Let the required data rate R be equal to 9600 bits/s, and let the required bit-error performance P_B be *at most* 10^{-5}. The goal is to choose a modulation scheme that meets the required performance. In general, an error-correction coding scheme may be needed if none of the allowable modulation schemes can meet the requirements. However, in this example, we will see that the use of error-correction coding is not necessary.

For any digital communication system, the relationship between received power to noise-power spectral density (P_r/N_0) and received bit-energy to noise-power spectral density (E_b/N_0) was shown in Equation (5.20c) to be

$$\frac{P_r}{N_0} = \frac{E_b}{N_0} R \tag{9.22}$$

Solving for E_b/N_0, in decibels, we obtain

$$\frac{E_b}{N_0}(\text{dB}) = \frac{P_r}{N_0} \ (\text{dB-Hz}) \ -R \ (\text{dB-bit/s}) \tag{9.23}$$

$$= 53 \text{ dB-Hz} - (10 \times \log_{10} 9600) \text{ dB-bit/s} = 13.2 \text{ dB (or 20.89)}$$

Since the required data rate of 9600 bits/s is much larger than the available bandwidth of 4000 Hz, the channel can be described as *bandwidth limited*. We therefore select MPSK as our modulation scheme. Recall that we have confined the possible modulation choices to be constant-envelope types; without such a restriction, it would be possible to select a modulation type with greater bandwidth efficiency. In an effort to conserve power, we next compute the *smallest possible* value of M, such that the symbol rate is *at most* equal to the available bandwidth of 4000 Hz. From Table 9.1, it is clear that the smallest value of M meeting this requirement is $M = 8$. Our next task is to determine whether the required bit-error performance of $P_B \leq 10^{-5}$ can be met by using 8-PSK modulation alone, or whether it is necessary to additionally use an error-correction coding scheme. It can be seen from Table 9.1, that 8-PSK *alone* will meet the requirements, since the required E_b/N_0 listed for 8-PSK is less then the received E_b/N_0 that was derived in Equation (9.23). However, imagine that we do not have Table 9.1. Let us demonstrate how to evaluate whether or not error-correction coding is necessary.

Figure 9.8 shows the basic modulator/demodulator (MODEM) block diagram summarizing the functional details of this design. At the modulator, the transformation from data bits to symbols yields an output symbol rate R_s that is a factor ($\log_2 M$) smaller than the input data-bit rate R, as can be seen in Equation (9.16). Similarly, at the input to the demodulator, the symbol-energy to noise-power spectral density E_s/N_0 is a factor ($\log_2 M$) larger than E_b/N_0, since each symbol is made up of ($\log_2 M$) bits. Because E_s/N_0 is larger than E_b/N_0 by the same factor that R_s is smaller than R, we can expand Equation (9.22), as follows:

$$\frac{P_r}{N_0} = \frac{E_b}{N_0} R = \frac{E_s}{N_0} R_s \tag{9.24}$$

The demodulator receives a waveform (in this example, one of $M = 8$ possible phase shifts) during each time interval T_s. The probability that the demodulator makes a symbol error $P_E(M)$ is well approximated by [12], and we write

$$P_E(M) \approx 2Q\left[\sqrt{\frac{2E_s}{N_0}} \sin\left(\frac{\pi}{M}\right)\right] \quad \text{for } M > 2 \tag{9.25}$$

Defining, Designing, and Evaluating Digital Communication Systems

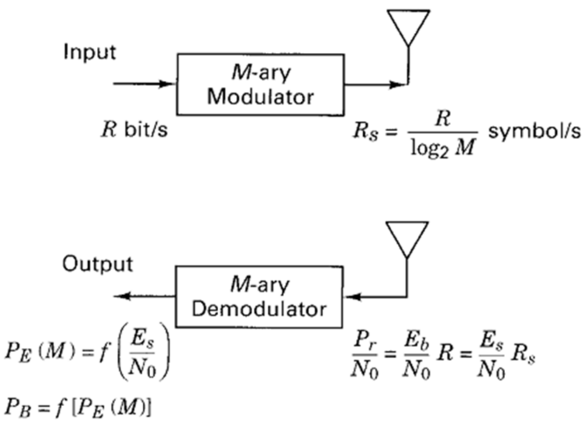


Figure 9.8: Basic modulator/demodulator (MODEM) without channel coding.

where $Q(x)$, the *complementary error function*, was defined in Equation (3.43) as

$$Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty \exp\left(-\frac{u^2}{2}\right) du$$

In Figure 9.8 and all the figures that follow, rather than show explicit probability relationships, the generalized notation $f(x)$ has been used to indicate some functional dependence on x.

A traditional way of characterizing communication (power) efficiency or error performance in digital systems is in terms of the received E_b/N_0 in decibels. This E_b/N_0 description has become standard practice. However, recall that at the input to the demodulator/detector, there are no bits; there are only waveforms that have been assigned bit meanings. Thus, the received E_b/N_0 value represents a bit-apportionment of the arriving waveform energy. A more precise (but unwieldy) name would be the energy per *effective bit* versus N_0. To solve for $P_E(M)$ in Equation (9.25), we first need to compute the ratio of received symbol-energy to noise-power spectral density, E_s/N_0. Since, from Equation (9.23), E_b/N_0 = 13.2 dB (or 20.89), and because each symbol is made up of ($\log_2 M$) bits, we compute, with $M = 8$,

$$\frac{E_s}{N_0} = (\log_2 M)\frac{E_b}{N_0} = 3 \times 20.89 = 62.67 \tag{9.26}$$

Using the results of Equation (9.26) in Equation (9.25), yields the symbol-error probability, $P_E = 2.2 \times 10^{-5}$. To transform this to bit-error probability, we need to use the relationship between bit-error probability P_B and symbol-error probability P_E for multiple-phase signaling [10]. We write

$$P_B \approx \frac{P_E}{\log_2 M} \quad (\text{for } P_E \ll 1) \tag{9.27}$$

which is a good approximation, when Gray coding [12] is used for the bit-to-symbol assignment. This last computation yields $P_B = 7.3 \times 10^{-6}$, which meets the required bit-error performance. Thus, in this example, no error-correction coding is necessary and 8-PSK modulation represents the design choice to meet the requirements of the bandwidth-limited channel (which we had predicted by examining the required E_b/N_0 values in Table 9.1).

Defining, Designing, and Evaluating Digital Communication Systems

9.7.6 Power-Limited Uncoded System Example

Now, suppose that we have exactly the same data rate and bit-error probability requirements as in the example of Section 9.7.5. However, in this example, let the available bandwidth W be equal to 45 kHz, and let the available P_r/N_0 be equal to 48 dB-Hz. As before, the goal is to choose a modulation or modulation/coding scheme that yields the required performance. In this example, we shall again find that error-correction coding is not required.

The channel in this example is clearly not bandwidth limited since the available bandwidth of 45 kHz is more than adequate for supporting the required data rate of 9600 bits/s. The received E_b/N_0 is found from Equation (9.23), as follows:

$$\frac{E_b}{N_0}(\text{dB}) = 48 \text{ db-Hz} - (10 \times \log_{10} 9600) \text{ dB-bit/s} = 8.2 \text{ dB (or 6.61)} \quad (9.28)$$

Since there is abundant bandwidth but a relatively small amount of E_b/N_0 for the required bit-error probability, this channel may be referred to as *power limited*. We therefore choose MFSK as the modulation scheme. In an effort to conserve power, we next search for the *largest possible M* such that the MFSK minimum bandwidth is not expanded beyond our available bandwidth of 45 kHz. From Table 9.1, we see that such a search results in the choice of $M = 16$. Out next task is to determine whether the required error performance of $P_B \leq 10^{-5}$ can be met by using 16-FSK alone, without the use of any error-correction coding. Similar to the previous example, it can be seen from Table 9.1, that 16-FSK *alone* will meet the requirements, since the required E_b/N_0 listed for 16-FSK is less than the received E_b/N_0 that was derived in Equation (9.28). However, imagine again that we do not have Table 9.1. Let us demonstrate how to evaluate whether or not error-correction coding is necessary.

As before, the block diagram in Figure 9.8 summarizes the relationship between symbol rate R_s and bit rate R, and between E_s/N_0 and E_b/N_0, which is identical to each of the respective relationships in the previous bandwidth-limited example. In this example, the 16-FSK demodulator receives a waveform (one of 16 possible frequencies) during each symbol time interval T_s. For noncoherent MFSK, the probability that the demodulator makes a symbol error is approximated by [13]

$$P_E(M) \lesssim \frac{M-1}{2} \exp\left(-\frac{E_s}{2N_0}\right) \quad (9.29)$$

To solve for $P_E(M)$ in Equation (9.29), we need to compute E_s/N_0, as we did in Example 1. Using the results of Equation (9.28) in Equation (9.26), with $M = 16$, we get

$$\frac{E_s}{N_0} = (\log_2 M)\frac{E_b}{N_0} = 4 \times 6.61 = 26.44 \quad (9.30)$$

Next, we combine the results of Equation (9.30) in Equation (9.29) to yield the symbol-error probability $P_E = 1.4 \times 10^{-5}$. To transform this to bit-error probability P_B, we need to use the relationship between P_B and P_E for orthogonal signaling [13], given by

$$P_B = \frac{2^{k-1}}{2^k - 1} \; P_E \quad (9.31)$$

This last computation yields $P_B = 7.3 \times 10^{-6}$, which meets the required bit-error performance. Thus, we can meet the given specifications for this power-limited channel by using 16-FSK modulation, without any need for error-correction coding (which we had predicted by examining the required E_b/N_0 values in Table 9.1).

Defining, Designing, and Evaluating Digital Communication Systems

9.7.7 Bandwidth-Limited and Power-Limited Coded System Example

In this example, we start with the same channel parameters as in the bandwidth-limited example of Section 9.7.5, namely, $W = 4000$ Hz, $P_r/N_0 = 53$ dB-Hz, and $R = 9600$ bits/s, with one exception. In the present example, we specify that the bit-error probability must be *at most* 10^{-9}. Since the available bandwidth is 4000 Hz, and from Equation (9.23) the available E_b/N_0 is 13.2 dB, it should be clear from Table 9.1, that the system is both bandwidth limited *and* power limited (8-PSK is the only possible choice to meet the bandwidth constraint; however, the available E_b/N_0 of 13.2 dB is certainly insufficient to meet the required bit-error probability of 10^{-9}). For such a small value of P_B, the system shown in Figure 9.8 will obviously be inadequate, and we need to consider the performance improvement that error-correction coding (within the available bandwidth) can provide. In general, one can use convolutional codes or block codes. To simplify the explanation, we shall choose a block code. The Bose, Chaudhuri, and Hocquenghem (BCH) codes form a large class of powerful error-correcting cyclic (block) codes [14]. For this example, let us select one of the codes from this family of codes. Table 9.2 presents a partial catalog of the available BCH codes in terms of n, k, and t, where k represents the number of information or data bits that the codes transforms into a longer block or n code bits (also called *channel bits* or *channel symbols*), and t represents the largest number of incorrect channel bits that the code can correct within each n-sized block. The *rate* of a code is defined as the ratio k/n; its inverse represents a measure of the code's redundancy.

TABLE 9.2 BCH Codes (Partial Catalog)

n	k	t
7	4	1
15	11	1
	7	2
	5	3
31	26	1
	21	2
	16	3
	11	5
63	57	1
	51	2
	45	3
	39	4
	36	5
	30	6
127	120	1
	113	2
	106	3
	99	4
	92	5
	85	6
	78	7
	71	9
	64	10
	57	11
	50	13
	43	14
	36	15
	29	21
	22	23
	15	27
	8	31

Defining, Designing, and Evaluating Digital Communication Systems

Since this example is represented by the same bandwidth-limited parameters that were given in Section 9.7.5, we start with the same 8-PSK modulation as before in order to meet the stated bandwidth constraint. However, we now additionally need to employ error-correction coding so that the bit-error probability can be lowered to $P_B \leq 10^{-9}$. To make the optimum code selection from Table 9.2, we are guided by the following goals:

1. The output bit-error probability of the combined modulation/coding system must meet the system error requirement.
2. The rate of the code must not expand the required transmission bandwidth beyond the available channel bandwidth.
3. The code should be as simple as possible. Generally, the shorter the code, the simpler will be its implementation.

The uncoded 8-PSK minimum bandwidth requirement is 3200 Hz (see Table 9.1), and the allowable channel bandwidth is specified as 4000 Hz. Therefore, the uncoded signal bandwidth may be increased by *no more than* a factor of 1.25 (or an expansion of 25%). Thus, the very first step in this (simplified) code selection example is to eliminate the candidates from Table 9.2 that would expand the bandwidth by more than 25%. The remaining entries in Table 9.2 form a much reduced set of "bandwidth-compatible" codes, which have been listed in Table 9.3. In Table 9.3, two columns designated Coding Gain, G, have been added, where coding gain in decibels is defined as

$$G(\text{dB}) = \left(\frac{E_b}{N_0}\right)_{\text{uncoded}} (\text{dB}) - \left(\frac{E_b}{N_0}\right)_{\text{coded}} (\text{dB}) \tag{9.32}$$

From Equation (9.32), coding gain can be described as a measure of the *reduction* in the required E_b/N_0 (in decibels) that needs to be provided, due to the error-performance properties of the channel coding. Coding gain is a function of the modulation type and bit-error probability. In Table 9.3, the coding gain G has been computed for MPSK at $P_B = 10^{-5}$ and 10^{-9}. For MPSK modulation, G is relatively independent of the value of M. Thus, for a particular bit-error probability, a given code will provide approximately the same coding gain when used with any of the MPSK modulation schemes. The coding gains in Table 9.3 were calculated using a procedure outlined under the section below, entitled "Calculating Coding Gain."

TABLE 9.3 Bandwidth-Compatible BCH Codes

			Coding Gain, G (dB), with MPSK	
n	k	t	$P_B = 10^{-5}$	$P_B = 10^{-9}$
31	26	1	1.8	2.0
63	57	1	1.8	2.2
	51	2	2.6	3.2
127	120	1	1.7	2.2
	113	2	2.6	3.4
	106	3	3.1	4.0

Defining, Designing, and Evaluating Digital Communication Systems

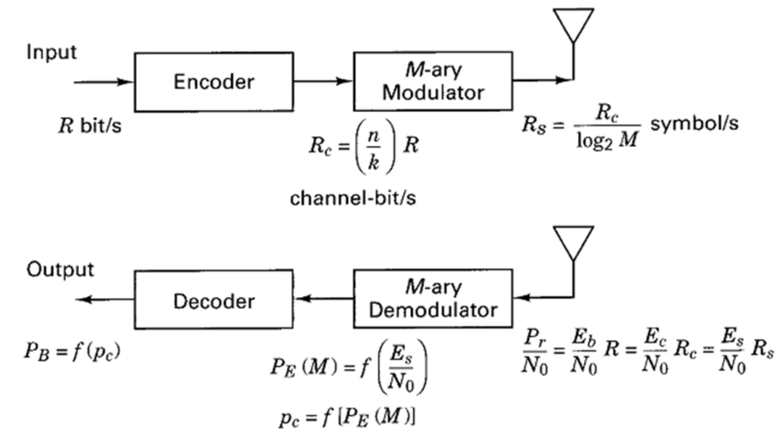


Figure 9.9: MODEM with channel coding.

Figure 9.9 illustrates a block diagram that summarizes the details of this system containing a modulator/demodulator (MODEM) and coding. When comparing Figure 9.9 with Figure 9.8, we see that the introduction of the encoder/decoder blocks has brought about additional transformations. In Figure 9.9, at the encoder/modulator are shown the relationships that exist when transforming from R bit/s to R_c channel-bit/s to R_s symbol/s.

We assume that our communication system is a real-time system and thus cannot tolerate any message delay. Therefore, the channel-bit rate R_c must *exceed* the data-bit rate R by the factor n/k. Further, each transmission symbol is made up of ($\log_2 M$) channel bits, so the symbol rate R_s is *less* than R_c by the factor ($\log_2 M$). For a system containing both modulation and coding, we summarize the rate transformations, as follows:

$$R_c = \left(\frac{n}{k}\right) R \tag{9.33}$$

$$R_s = \frac{R_c}{\log_2 M} \tag{9.34}$$

At the demodulator/decoder shown in Figure 9.9, the transformations amongst data-bit energy, channel-bit energy, and symbol energy are related (in a reciprocal fashion) by the same factors as shown amongst the rate transformations in Equations (9.33) and (9.34). Since the encoding transformation has replaced k data bits with n channel bits, then the ratio of channel-bit energy to noise-power spectral density, E_c/N_0, is computed by decrementing the value of E_b/N_0 by the factor k/n. Also, since each transmission symbol is made up of ($\log_2 M$) channel bits, then E_s/N_0, which is needed in Equation (9.25) to solve for P_E, is computed by incrementing E_c/N_0 by the factor ($\log_2 M$). For a system containing both modulation and coding, we summarize the energy-to-noise-power spectral density transformations, as follows:

$$\frac{E_c}{N_0} = \left(\frac{k}{n}\right) \frac{E_b}{N_0} \tag{9.35}$$

$$\frac{E_s}{N_0} = (\log_2 M) \frac{E_c}{N_0} \tag{9.36}$$

Therefore, using Equations (9.33) through (9.36), we can now expand the expression for P_r/N_0 in Equation (9.24), as follows:

$$\frac{P_r}{N_0} = \frac{E_b}{N_0} R = \frac{E_c}{N_0} R_c = \frac{E_s}{N_0} R_s \tag{9.37}$$

Defining, Designing, and Evaluating Digital Communication Systems

As before, a standard way of describing the link is in terms of the received E_b/N_0 in decibels. However, *there are no data bits* at the input to the demodulator/detector: *neither are there any channel bits.* There are *only* waveforms (transmission symbols) that have bit meanings, and thus the waveforms can be described in terms of bit-energy apportionments. Equation (9.37) illustrates that the predetection point in the receiver is a useful reference point at which we can relate the *effective* energy and the *effective* speed of various parameters of interest. We use the word "effective" because the only type of signals that actually appear at the predetection point are waveforms (transformed to baseband pulses) that we call symbols. Of course, these symbols are related to channel bits, which in turn are related to data bits. To emphasize the point that Equation (9.37) represents a useful kind of "bookkeeping," consider a system wherein a stream of some number of bits, say 273 bits, appears so repeatedly as a module, that we give this group 273 bits a name; we all it a "chunk." Engineers do that all the time—e.g., eight bits are referred to as a byte. The moment we identify this new entity, the chunk, it can immediately be related to the parameters in Equation (9.37), since P_r/N_0 will now also equal the energy in a chunk over N_0 times the chunk rate. In Chapter 12, we will in fact do something similar when we extend Equation (9.37) to include spread-spectrum *chips.*

Since P_r/N_0 and R were given as 53 dB-Hz and 9600 bits/s, respectively, we find as before from Equation (9.23) that the received $E_b/N_0 = 13.2$ dB. Note that the received E_b/N_0 is fixed and independent of the code parameters n and k, and the modulation parameter M. As we search, in Table 9.3, for the ideal code that will meet the specifications, we can iteratively repeat the computations that are summarized in Figure 9.9. It might be useful to program on a PC (or calculator) the following four steps as a function of n, k, and t. Step 1 starts by combining Equations (9.35) and (9.36), as follows:

Step 1:
$$\frac{E_s}{N_0} = (\log_2 M)\frac{E_c}{N_0} = (\log_2 M)\left(\frac{k}{n}\right)\frac{E_b}{N_0} \tag{9.38}$$

Step 2:
$$P_E(M) \approx 2Q\left[\sqrt{\frac{2\,E_s}{N_0}}\sin\left(\frac{\pi}{M}\right)\right] \tag{9.39}$$

The expression in step 2 is the approximation (for M-ary PSK) for symbol-error probability P_E, rewritten from Equation (9.25). At each symbol-time interval, the demodulator makes a symbol decision, but it delivers to the decoder a channel-bit sequence representing that symbol. When the channel-bit output of the demodulator is quantized to two levels, denoted by 1 and 0, the demodulator is said to make *hard decisions.* When the output is quantized to more than two levels, the demodulator is said to make *soft decisions.* Throughout this section, hard-decision demodulation is assumed.

Now that a decoder block is present in the system, we designate the channel-bit-error probability out of the demodulator and into the decoder as p_c, and reserve the notation P_B for the bit-error probability *out of the decoder* (the decoded bit-error probability). Equation (9.27) is rewritten in terms of p_c as follows:

Step 3:
$$p_c \approx \frac{P_E}{\log_2 M} \quad (\text{for } P_E \ll 1) \tag{9.40}$$

Step 3 relates the channel-bit-error probability to the symbol-error probability out of the demodulator, assuming Gray coding, as referenced in Equation (9.27).

For a real-time communication system, using traditional channel-coding schemes, and a given value of received P_r/N_0, the value of E_s/N_0 with coding will *always be less* than the value of E_s/N_0 without coding. Since the demodulator, with coding, receives less E_s/N_0 it makes more errors! However, when coding is used, the system error-performance doesn't only depend on the performance of the demodulator, it also depends on the performance of the decoder. Thus, for error-performance improvement due to coding, we require that the decoder provides enough error correction to *more than compensate* for the poor performance of the demodulator. The final output decoded bit-error probability P_B depends on the particular code, the decoder, and the channel-bit-error probability p_c. It can be expressed [15] by the following approximation:

Defining, Designing, and Evaluating Digital Communication Systems

Step 4: $$P_B \approx \frac{1}{n} \sum_{j=t+1}^{n} j \binom{n}{j} p_c^j (1 - p_c)^{n-j} \quad (9.41)$$

In Step 4, t is the largest number of channel bits that the code can correct within each block of n bits. Using Equations (9.38) through (9.41) in the above four steps, the decoded bit-error probability P_B can be computed as a function of n, k, and t for each of the codes listed in Table 9.3. The entry that meets the stated error requirement with the *largest possible* code rate and the *smallest* value of n is the double-error correcting (63, 51) code. The computations are as follows:

Step 1: $$\frac{E_s}{N_0} = 3\left(\frac{51}{63}\right) 20.89 = 50.73$$

where $M = 8$, and the received $E_b/N_0 = 13.2$ dB (or 20.89)

Step 2: $$P_E \approx 2Q\left[\sqrt{101.5} \times \sin\left(\frac{\pi}{8}\right)\right] = 2Q(3.86) = 1.2 \times 10^{-4}$$

Step 3: $$p_c \approx \frac{1.2 \times 10^{-4}}{3} = 4 \times 10^{-5}$$

Step 4: $$P_B \approx \frac{3}{63}\binom{63}{3}(4 \times 10^{-5})^3 (1 - 4 \times 10^{-5})^{60} + \frac{4}{63}\binom{63}{4}(4 \times 10^{-5})^4 (1 - 4 \times 10^{-5})^{59} + \cdots$$
$$= 1.2 \times 10^{-10}$$

In Step 4, the bit-error-correcting capability of the code is $t = 2$. For the computation of P_B in Step 4, only the first two terms in the summation of Equation (9.41) have been used, since the other terms have a vanishingly small effect on the result whenever p_c is small or E_b/N_0 reasonably large. It is important to note that when performing this computation with a computer, it is advised (for being safe) to *always* include all of the summation terms in Equation (9.41), since a truncated solution can be very erroneous whenever E_b/N_0 is small. Now that we have selected the (63, 51) code, the values of channel-bit rate R_c and symbol rate R_s are computed using Equations (9.33) and (9.34), with $M = 8$:

$$R_c = \left(\frac{n}{k}\right) R = \left(\frac{63}{51}\right) 9600 \approx 11{,}859 \text{ channel-bits/s}$$

$$R_s = \frac{R_c}{\log_2 M} = \frac{11{,}859}{3} = 3953 \text{ symbols/s}$$

Defining, Designing, and Evaluating Digital Communication Systems

9.7.7.1 Calculating Coding Gain

A more direct way to find the simplest code that meets the specified error performance for the example in Section 9.7.7 is to first compute, for the *uncoded* 8-PSK, how much more E_b/N_0 beyond the available 13.2 dB would be required to yield $P_B = 10^{-9}$. This additional E_b/N_0 is the required coding gain. Next, we simply choose, from Table 9.3, the code that provides this coding gain. The *uncoded* E_s/N_0 that will yield an error probability of $P_B = 10^{-9}$ is found by writing, from Equations (9.27) and (9.39),

$$P_B \approx \frac{P_E}{\log_2 M} \approx \frac{2Q\left[\sqrt{\frac{2E_s}{N_0}}\sin\left(\frac{\pi}{M}\right)\right]}{\log_2 M} = 10^{-9} \tag{9.42}$$

At this low value of bit-error probability, it is valid to use Equation (3.44) to approximate $Q(x)$ in Equation (9.42). By trial-and-error (on a programmable calculator), we find that the *uncoded* $E_s/N_0 = 120.67 = 20.8$ dB, and since each symbol is made up of $(\log_2 8) = 3$ bits, the required $(E_b/N_0)_{\text{uncoded}} = 120.67/3 = 40.22 = 16$ dB.

We know from the given parameters in this example and Equation (9.23), that the received $(E_b/N_0)_{\text{coded}} = 13.2$ dB. Therefore, using Equation (9.32), we see that the required coding gain to meet the bit-error performance of $P_B = 10^{-9}$ is

$$G(\text{dB}) = \left(\frac{E_b}{N_0}\right)_{\text{uncoded}}(\text{dB}) - \left(\frac{E_b}{N_0}\right)_{\text{coded}}(\text{dB}) = 16\ \text{dB} - 13.2\ \text{dB} = 2.8\ \text{dB}$$

To be precise in this computation, each of the E_b/N_0 values in the above computation must correspond to exactly the same value of bit-error probability (which they do not). They correspond to $P_B = 10^{-9}$ and $P_B = 1.2 \times 10^{-10}$, respectively. However, at these low probability values, even with such a discrepancy, this computation still provides a good approximation of the required coding gain. In searching Table 9.3 for the simplest code that will yield a coding gain of *at least* 2.8 dB, we se that the choice is the (63, 51) code, which corresponds to the same code choice that was made earlier. Note that coding gain must always be specified for a particular error probability and modulation type, as it is in Table 9.3.

Defining, Designing, and Evaluating Digital Communication Systems

9.7.7.2 Code Selection

Consider a real-time communication system, where the specifications cause it to be power-limited, but there is ample available bandwidth, and the users require a very small bit-error probability. The situation calls for error-correction coding. Suppose that we were asked to select one of the BCH codes listed in Table 9.2. Since the system is not bandwidth-limited, and it requires very good error performance, one might be tempted to simply choose the most powerful code in Table 9.2, that is the (127, 8) code, capable of correcting any combination of up to 31 flawed bits within a block of 127 code bits. Would anyone use such a code in a real-time communication system? No, they would not. Let us explain why such a choice would be *unwise*.

Whenever error-correction coding is used in a real-time communication system, there are two mechanisms at work that influence error performance. One mechanism works to improve the performance, and the other works to degrade it. The improving mechanism is the coding; the greater redundancy, the greater will be the error-correcting capability of the code. The degrading mechanism is the energy reduction per channel symbol or code bit (compared with the data bit). This reduced energy stems from the increased redundancy (giving rise to faster signaling in a real-time communication system). The reduced symbol energy causes the demodulator to make more errors. Eventually, the second mechanism wins out, and thus at very low code rates we see degradation. This is demonstrated in Example 9.4 below. Note that the degrading mechanism only applies for coding in a real-time system (where messages cannot be delayed). For systems that can endure message delays, the trade-off for getting the benefits of the code redundancy is delay (not reduced symbol energy).

Defining, Designing, and Evaluating Digital Communication Systems

Example 9.4 Choosing a Code to Meet Performance Requirements

A system is specified with the following parameters: $P_r/N_0 = 67$ dB-Hz, data rate $R = 10^6$ bits/s, available bandwidth $W = 20$ MHz, decoded bit-error probability $P_B \le 10^{-7}$, and the modulation is BPSK. Choose a code from Table 9.2 that will fulfill these requirements. Start by considering the (127, 8) code. It appears attractive because it has the greatest bit-error correcting capability on the list.

Solution

The (127, 8) code expands the transmission bandwidth by a factor of 127/8 = 15.875. Hence, the signaling rate of 1 Mbit/s (giving rise to a nominal bandwidth of 1 MHz) will be expanded by using this code to 15.875 MHz. The transmission signal is within the available bandwidth of 20 MHz, even after allowing another 25% bandwidth expansion for filtering. After choosing this code, we next evaluate the error performance, by following the steps outlined in Section 9.7.7, which yields

$$\frac{E_b}{N_0} = \frac{P_r}{N_0}\left(\frac{1}{R}\right) = 67 \text{ dB} - 60 \text{ dB} = 7 \text{ dB (or 5)}$$

$$\frac{E_s}{N_0} = \frac{E_c}{N_0} = \left(\frac{k}{n}\right)\frac{E_b}{N_0} = \left(\frac{8}{127}\right)5 = 0.314$$

Since the modulation is binary, then $p_c = P_E$, so that

$$p_c = P_E \approx Q\left(\sqrt{\frac{2E_s}{N_0}}\right) = Q(\sqrt{0.628}) = Q(0.7936) = 0.2156$$

Since the (127, 8) code is a $t = 31$ error-correcting code, we next use Equation (9.41) to find the decoded bit-error probability, as follows:

$$P_B \approx \frac{1}{n}\sum_{j=t+1}^{n} j\binom{n}{j} p_c^j (1-p_c)^{n-j} = \frac{1}{127}\sum_{j=32}^{127} j\binom{127}{j}(0.2156)^j(1-0.2156)^{127-j}$$

Whenever p_c is very small, it suffices to only use the first term, or the first few terms in the summation. But when p_c is large, as it is here, computer assistance is helpful. Solving the above with $p_c = 0.2156$ yields a decoded bit-error probability of $P_B = 0.05$, which is a far cry from the system requirement of 10^{-7}. Let us next select a code whose code rate is close to the popular rate $\frac{1}{2}$—that is, the (127, 64) code. It is not as capable as the first choice because it only corrects 10 flawed bits in a block of 127 code bits. However, watch what happens. Using the same steps as before yields

$$\frac{E_s}{N_0} = \frac{E_c}{N_0} = \left(\frac{k}{n}\right)\frac{E_b}{N_0} = \left(\frac{64}{127}\right)5 = 2.519$$

Notice how much larger the E_s/N_0 is here, compared with the case using (127, 8) coding:

$$p_c = Q(\sqrt{2 \times 2.519}) = Q(\sqrt{2.245}) = 0.0124$$

$$P_B \approx \frac{1}{127}\sum_{j=11}^{127} j\binom{127}{j}(0.0124)^j(1-0.0124)^{127-j}$$

And the result yields $P_B = 5.6 \times 10^{-8}$, which meets the system requirements. From this example, one should see that the selection of a code needs to be made in concert with the modulation choice and the available E_b/N_0. One can be guided by the fact that very high rates and very low rates generally perform poorly in a real-time communication system, as evidenced by the Figure 8.6 curves presented in Chapter 8.

9.8 BANDWIDTH-EFFICIENT MODULATION

The primary objective of spectrally efficient modulation techniques is to maximize bandwidth efficiency. The increasing demand for digital transmission channels has led to the investigation of spectrally efficient modulation techniques [8, 16] to maximize bandwidth efficiency and thus help ameliorate the spectral congestion problem.

Some systems have additional modulation requirements besides spectral efficiency. For example, satellite systems with highly nonlinear transponders require a constant envelope modulation. This is because the nonlinear transponder produces extraneous sidebands when passing a signal with amplitude fluctuations (due to a mechanism called AM-to-PM conversion). These sidebands deprive the information signals of some of their portion of transponder power, and also can interfere with nearby channels (adjacent channel interference) or with other communication systems (co-channel interference). *Offset QPSK* (OQPSK) and *Minimum shift keying* (MSK) are two examples of constant envelope modulation schemes that are attractive for systems using nonlinear transponders.

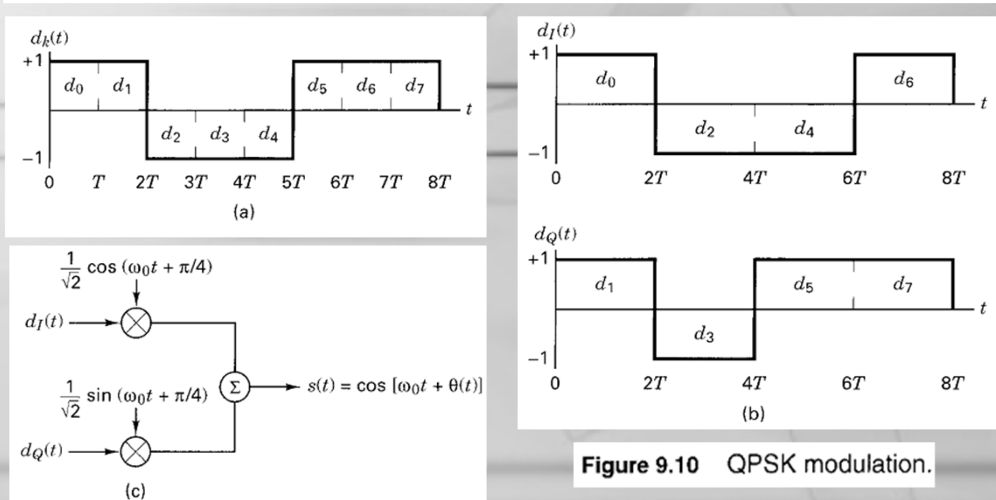


Figure 9.10 QPSK modulation.

9.8.1 QPSK and Offset QPSK Signaling

Figure 9.10 illustrates the partitioning of a typical pulse stream for QPSK modulation. Figure 9.10a shows the original data stream $d_k(t) = d_0, d_1, d_2, \ldots$ consisting of bipolar pulses; that is, the values of $d_k(t)$ are +1 or − 1, representing binary one and zero, respectively. This pulse stream is divided into an in-phase stream, $d_I(t)$, and a quadrature stream, $d_Q(t)$, illustrated in Figure 9.10b, as follows:

$$d_I(t) = d_0, d_2, d_4, \ldots \text{ (even bits)} \tag{9.43}$$

$$d_Q(t) = d_1, d_3, d_5, \ldots \text{ (odd bits)}$$

Note that $d_I(t)$ and $d_Q(t)$ each have half the bit rate of $d_k(t)$. A convenient orthogonal realization of a QPSK waveform, $s(t)$, is achieved by amplitude modulating the in-phase and quadrature data streams onto the cosine and sine functions of a carrier wave, as follows:

$$s(t) = \frac{1}{\sqrt{2}} d_I(t) \cos\left(2\pi f_0 t + \frac{\pi}{4}\right) + \frac{1}{\sqrt{2}} d_Q(t) \sin\left(2\pi f_0 t + \frac{\pi}{4}\right) \tag{9.44}$$

Using the trigonometric identities shown in Equations (D.5) and (D.6), Equation (9.44) can also be written as

$$s(t) = \cos\left[2\pi f_0 t + \theta(t)\right] \tag{9.45}$$

The QPSK modulator shown in Figure 9.10c uses the sum of cosine and sine terms, while a similar device, described in Section 4.6, uses the difference of such terms. The treatment in this section corresponds to that of Pasupathy [17]. Because a coherent receiver needs to resolve any phase ambiguities, then the use of a different phase format at the transmitter can be handled as part of that ambiguity. The pulse stream $d_I(t)$ amplitude-modulates the cosine function with an amplitude of +1 or −1. This is equivalent to shifting the phase of the cosine function by 0 or π; consequently, this produces a BPSK waveform. Similarly, the pulse stream $d_Q(t)$ modulates the sine function, yielding a BPSK waveform orthogonal to the cosine function. The summation of these two orthogonal components of the carrier yields the QPSK waveform. The value of $\theta(t)$ will correspond to one of the four possible combinations of $d_I(t)$ and $d_Q(t)$ in Equation 9.44: $\theta(t) = 0°, \pm 90°$, or $180°$; the resulting signal vectors are seen in the signal space illustrated in Figure 9.11. Because $\cos(2\pi f_0 t + \pi/4)$ and $\sin(2\pi f_0 t + \pi/4)$ are orthogonal, the two BPSK signals can be detected separately.

Defining, Designing, and Evaluating Digital Communication Systems

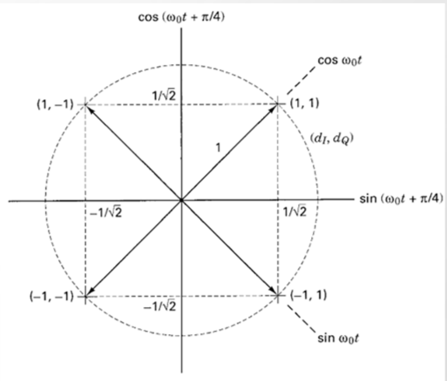


Figure 9.11 Signal space for QPSK and OQPSK.

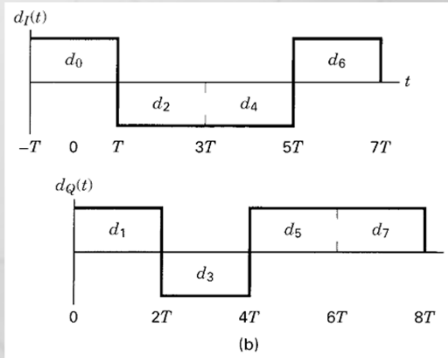


Figure 9.12 Offset QPSK (OQPSK) data streams.

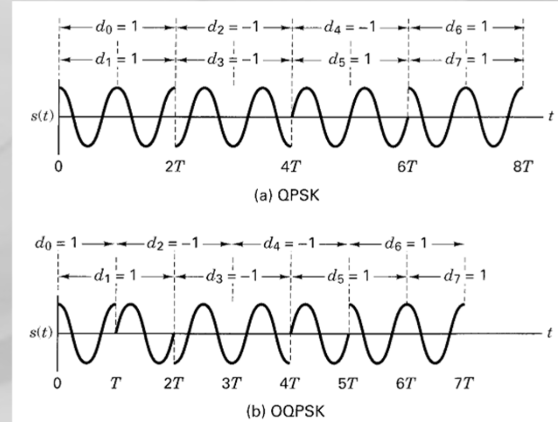


Figure 9.13 (a) QPSK and (b) OQPSK waveforms. (Reprinted with permission from S. Pasupathy, "Minimum Shift Keying: A Spectrally Efficient Modulation," *IEEE Commun. Mag.*, July 1979, Fig. 4, p. 17. © 1979 IEEE.)

Offset QPSK (OQPSK) signaling can also be represented by Equations (9.44) and (9.45); the difference between the two modulation schemes, QPSK and OQPSK, is only in the *alignment* of the two baseband waveforms. As shown in Figure 9.10, the duration of each original pulse is T (Figure 9.10a), and hence in the partitioned streams of Figure 9.10b, the duration of each pulse is $2T$. In standard QPSK, the odd and even pulse streams are both transmitted at the rate of $1/2T$ bit/s and are synchronously aligned, such that their transitions coincide, as shown in Figure 9.10b. In OQPSK, sometimes called *staggered QPSK* (SQPSK), there is the same data stream partitioning and orthogonal transmission; the difference is that the timing of the pulse stream $d_I(t)$ and $d_Q(t)$ is shifted such that the alignment of the two streams is offset by T. Figure 9.12 illustrates this offset.

In standard QPSK, due to the coincident alignment of $d_I(t)$ and $d_Q(t)$, the carrier phase can change only once every $2T$. The carrier phase during any $2T$ interval can be any one of the four phases shown in Figure 9.11, depending on the values of $d_I(t)$ and $d_Q(t)$ during that interval. During the next $2T$ interval, if neither pulse stream changes sign, the carrier phase remains the same. If only one of the pulse streams change sign, a phase shift of $\pm 90°$ occurs. A change in both streams results in a carrier phase shift of 180°. Figure 9.13a shows a typical QPSK waveform for the sample sequence $d_I(t)$ and $d_Q(t)$ shown in Figure 9.10.

If a QPSK modulated signal undergoes filtering to reduce the spectral side-lobes, the resulting waveform will no longer have a constant envelope and in fact, the occasional 180° phase shifts will cause the envelope to go to zero momentarily (see Figure 9.13a). When these signals are used in satellite channels employing highly nonlinear amplifiers, the constant envelope will tend to be restored. However, at the same time, all of the *undesirable* frequency side-lobes, which can interfere with nearby channels and other communication systems, are also restored.

In OQPSK, the pulse streams $d_I(t)$ and $d_Q(t)$ are staggered and thus do not change states simultaneously. The possibility of the carrier changing phase by 180° is eliminated, since only one component can make a transition at one time. Changes are limited to 0° and $\pm 90°$ every T seconds. Figure 9.13b shows a typical OQPSK waveform for the sample sequence in Figure 9.12. When an OQPSK signal undergoes bandlimiting, the resulting intersymbol interference causes the envelope to droop slightly in the region of $\pm 90°$ phase transition, but since the phase transitions of 180° have been avoided in OQPSK, the envelope will not go to zero as it does with QPSK. When the bandlimited OQPSK goes through a nonlinear transponder, the envelope droop is removed; however, the high-frequency components associated with the collapse of the envelope are not reinforced. Thus out-of-band interference is avoided [17].

Defining, Designing, and Evaluating Digital Communication Systems

9.8.2 Minimum Shift Keying

The main advantage of OQPSK over QPSK, that of suppressing out-of-band interference, suggests that further improvement is possible if the OQPSK format is modified to avoid discontinuous phase transitions. This was the motivation for designing continuous phase modulation (CPM) schemes. *Minimum shift keying* (MSK) is one such scheme [17–20]. MSK can be viewed as either a special case of *continuous-phase frequency shift keying* (CPFSK), or a special case of OQPSK with sinusoidal symbol weighting. When viewed as CPFSK, the MSK waveform can be expressed as [18]

$$s(t) = \cos\left[2\pi\left(f_0 + \frac{d_k}{4T}\right)t + x_k\right] \qquad kT < t < (k+1)T \tag{9.46}$$

where f_0 is the carrier frequency, $d_k = \pm 1$ represents the bipolar data being transmitted at a rate $R = 1/T$, and x_k is a phase constant which is valid over the kth binary data interval. Notice that for $d_k = 1$, the frequency transmitted is $f_0 + 1/4T$, and for $d_k = -1$, the frequency transmitted is $f_0 - 1/4T$. The tone spacing in MSK is thus one-half that employed for noncoherently demodulated orthogonal FSK, giving rise to the name *minimum* shift keying. During each T-second data interval, the value of x_k is a constant, that is, $x_k = 0$ or π, determined by the requirement that the phase of the waveform be continuous at $t = kT$. This requirement results in the following recursive phase constraint for x_k:

$$x_k = \left[x_{k-1} + \frac{\pi k}{2}(d_{k-1} - d_k)\right] \text{ modulo } 2\pi \tag{9.47}$$

Equation (9.46) can be expressed in a quadrature representation, using the identities in Equation (D.5) and (D.6), and we write

$$s(t) = a_k \cos\frac{\pi t}{2T}\cos 2\pi f_0 t - b_k \sin\frac{\pi t}{2T}\sin 2\pi f_0 t \qquad kT < t < (k+1)T \tag{9.48}$$

where

$$a_k = \cos x_k = \pm 1$$
$$b_k = d_k \cos x_k = \pm 1 \tag{9.49}$$

The in-phase (I) component is identified as $a_k \cos(\pi t/2T)\cos 2\pi f_0 t$, where $\cos 2\pi f_0 t$ is the carrier, $\cos(\pi t/2T)$ can be regarded as a *sinusoidal symbol weighting*, and a_k is a data-dependent term. Similarly, the quadrature (Q) component is identified as $b_k \sin(\pi t/2T)\sin 2\pi f_0 t$, where $\sin 2\pi f_0 t$ is the quadrature carrier term, $\sin(\pi t/2T)$ can be regarded as a sinusoidal symbol weighting, and b_k is a data-dependent term. It might appear that the a_k and b_k terms can change every T seconds, since the source data d_k can change every T seconds. However, because of the continuous phase constraint, the a_k term can only change value at the zero crossings of $\cos(\pi t/2T)$ and the b_k term can only change value at the zero crossings of $\sin(\pi t/2T)$. Thus, the symbol weighting in either the I- or Q-channel is a half-cycle sinsuoidal pulse of duration $2T$ seconds with alternating sign. As in the case of OQPSK, the I and Q components are offset T seconds with respect to one another.

Notice that x_k in Equation (9.46) is a function of the difference between the prior data bit and the present data bit (differential encoding). Hence the a_k and b_k terms in Equation (9.48) can be viewed as *differentially encoded* components of the d_k source data. However, for bit-to-bit independent data d_k, the signs of successive I- or Q-channel pulses are also random from one $2T$-second pulse interval to the next. Thus when viewed as a special case of OQPSK, Equation (9.48) can be rewritten with more straightforward (nondifferential) data encoding [18] as follows:

$$s(t) = d_I(t)\cos\frac{\pi t}{2T}\cos 2\pi f_0 t + d_Q(t)\sin\frac{\pi t}{2T}\sin 2\pi f_0 t \qquad (9.50)$$

where $d_I(t)$ and $d_Q(t)$ have the same in-phase and quadrature data stream interpretation as in Equation (9.43). This MSK format in Equation (9.50) is sometimes referred to as *precoded MSK.* Figure 9.14 illustrates Equation (9.50) pictorially. Figure 9.14a and c show the sinusoidal weighting of the I- and Q-channel pulses. These sequences represent the same data sequences as in Figure 9.12, but here, multiplication by a sinusoid results in more gradual phase transitions compared to those of the original data representation. Figure 9.14b and d illustrate the modulation of the orthogonal components $\cos 2\pi f_0 t$ and $\sin 2\pi f_0 t$, respectively, by the sinusoidally shaped data streams. Figure 9.14e illustrates the summation of the orthogonal components from Figure 9.14b and d. In summary, the following properties of MSK modulation can be deduced from Equation (9.50) and Figure 9.14:

(1) the waveform $s(t)$ has constant envelope; (2) there is phase continuity in the RF carrier at the bit transitions; and (3) the waveform $s(t)$ can be regarded as an FSK waveform with signaling frequencies $f_0 + 1/4T$ and $f_0 - 1/4T$. Therefore, the minimum tone separation required for MSK modulation is

$$\left(f_0 + \frac{1}{4T}\right) - \left(f_0 - \frac{1}{4T}\right) = \frac{1}{2T} \qquad (9.51)$$

which is equal to half the bit rate. Notice that the required tone spacing for MSK is one-half the spacing, $1/T$, required for the noncoherent detection of FSK signals (see Section 4.5.4). This is because the carrier phase is known and continuous, enabling the signal to be coherently demodulated.

The power spectral density $G(f)$ for QPSK and OQPSK is given by [18]

$$G(f) = 2PT\left(\frac{\sin 2\pi fT}{2\pi fT}\right)^2 \qquad (9.52)$$

where P is the average power in the modulated waveform. For MSK, $G(f)$ is given by [18]

$$G(f) = \frac{16PT}{\pi^2}\left(\frac{\cos 2\pi fT}{1 - 16f^2T^2}\right)^2 \qquad (9.53)$$

The normalized power spectral density ($P = 1$ W) for QPSK, OQPSK, and MSK are sketched in Figure 9.15. A spectral plot of BPSK is included for comparison. The fact that BPSK requires more bandwidth than the others for a given level of spectral density should come as no surprise. In Section 9.5.1 and Figure 9.6 we saw that the theoretical bandwidth efficiency of BPSK is half that of QPSK. It is seen from Figure 9.15 that MSK has lower sidelobes than QPSK or OQPSK. This is a consequence of multiplying the data stream with a sinusoid, yielding more *gradual phase transitions.* The more gradual the transition, the faster the spectral tails drop to zero. MSK is *spectrally more efficient* than QPSK or OQPSK; however, as can be seen from Figure 9.15, the MSK spectrum has a wider mainlobe than QPSK and OQPSK. Therefore, MSK may not be the preferred method for narrowband links. However, MSK might be the preferred choice for multiple-carrier systems, because its relatively low spectral sidelobes help to avoid excessive adjacent channel interference (ACI). The reason for the QPSK spectrum having a narrower mainlobe than MSK is that, for a given bit rate, the QPSK symbol rate is half the MSK symbol rate.

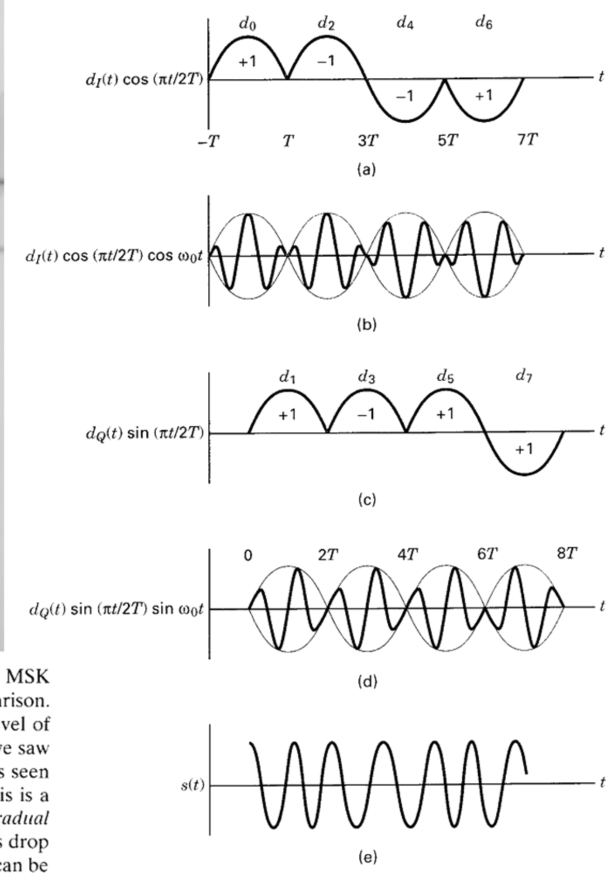


Figure 9.14 Minimum shift keying (MSK). (a) Modified I bit stream. (b) I bit stream times carrier. (c) Modified Q bit stream. (d) Q bit stream times carrier. (e) MSK waveform. (Reprinted with permission from S. Pasupathy, "Minimum Shift Keying: A Spectrally Efficient Modulation," *IEEE Commun. Mag.*, July 1979, Fig. 5, p. 18. © 1979 IEEE.)

Defining, Designing, and Evaluating Digital Communication Systems

9.8.2.1 Error Performance of OQPSK and MSK

We have seen that BPSK and QPSK have the same bit-error probability because QPSK is configured as two BPSK signals modulating orthogonal components of the carrier. Since staggering the bit streams does not change the orthogonality of the carriers, OQPSK has the same theoretical bit error performance as BPSK and QPSK.

Minimum shift keying uses antipodal symbol shapes, $\pm \cos(\pi t/2T)$ and $\pm \sin(\pi t/2T)$, over $2T$ to modulate the two quadrature components of the carrier. Thus

when a matched filter is used to recover the data from each of the quadrature components independently, MSK, as defined in Equation (9.50), has the same error performance properties as BPSK, QPSK, and OQPSK [17]. However, if MSK is coherently detected as an FSK signal over an observation interval of T seconds, it would be poorer than BPSK by 3 dB [17]. MSK, with differentially encoded data, as defined in Equation (9.46), has the same error probability performance as the coherent detection of differentially encoded PSK. MSK can also be noncoherently detected [19]. This permits inexpensive demodulation when the value of received E_b/N_0 permits.

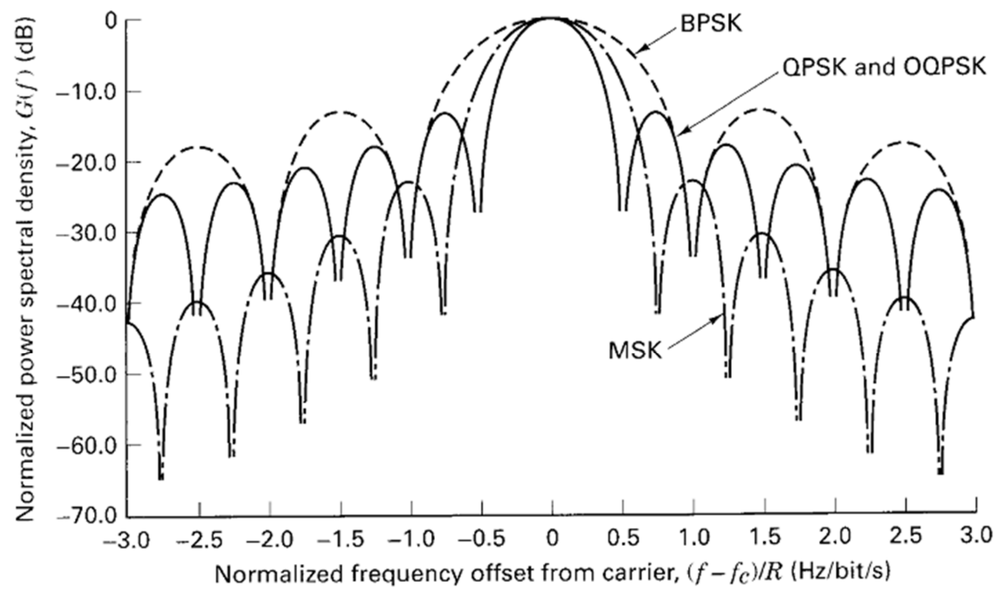


Figure 9.15 Normalized power spectral density for BPSK, QPSK, OQPSK, and MSK. (Reprinted with permission from F. Amoroso, "The Bandwidth of Digital Data Signals," *IEEE Commun. Mag.*, vol. 18, no. 6, Nov. 1980, Fig. 2A, p. 16. © 1980 IEEE.)

Defining, Designing, and Evaluating Digital Communication Systems

9.8.3 Quadrature Amplitude Modulation

Coherent M-ary phase shift keying (MPSK) modulation is a well-known technique for achieving bandwidth reduction. Instead of using a binary alphabet with 1 bit of information per channel symbol period, an alphabet with M symbols is used, permitting the transmission of $k = \log_2 M$ bits during each symbol period. Since the use of M-ary symbols allows a k-fold increase in the data rate within the same bandwidth, then for a fixed data rate, use of M-ary PSK reduces the required bandwidth by a factor k. (See Section 4.8.3.)

From Equation (9.44) it can be seen that QPSK modulation consists of two independent streams. One stream amplitude-modulates the cosine function of a carrier wave with levels +1 and −1, and the other stream similarly amplitude-modulates the sine function. The resultant waveform is termed a double-sideband suppressed-carrier (DSB-SC) wave, since the RF bandwidth is twice the baseband bandwidth (see Section 1.7.1) and there is no isolated carrier term. *Quadrature amplitude modulation* (QAM) can be considered a logical extension of QPSK, since QAM also consists of two independently amplitude-modulated carriers in quadrature. Each block of k bits (k assumed even) can be split into two ($k/2$)-bit blocks which use ($k/2$)-bit digital-to-analog (D/A) converters to provide the required modulating voltages for the carriers. At the receiver, each of the two signals is independently detected using matched filters. QAM signaling can also be viewed as a combination of amplitude shift keying (ASK) and phase shift keying (PSK), giving rise to the alternative name, *amplitude phase keying* (APK). Finally, it can also be viewed as amplitude shift keying in two dimensions, giving rise to the name *quadrature amplitude shift keying* (QASK).

Figure 9.16a illustrates a two-dimensional signal space and a set of 16-ary QAM signal vectors or points arranged in a rectangular constellation. A canonical QAM modulator is shown in Figure 9.16b. Assuming that Gaussian noise is the only channel disturbance, the simple channel model of Figure 9.16c applies. Signals are sent in pairs (x, y). The model indicates that the signal point coordinates (x, y) are transmitted over separate channels and independently perturbed by Gaussian noise variables (n_x, n_y), each with zero mean and variance N. Or we can say that the two-dimensional signal point is perturbed by a two-dimensional Gaussian noise variable. If the average signal energy (mean-square value of the signal coordinates) is S, then the signal-to-noise ratio is S/N. The simplest method of digital signaling through such a system is to use one-dimensional pulse amplitude modulation (PAM) independently for each signal coordinate. In PAM, to send k bits/dimension over a Gaussian channel, each signal point coordinate takes on one of 2^k equally likely equispaced amplitudes. By convention, the signal points are grouped about the center of the space at amplitudes $\pm 1, \pm 3, \ldots, \pm(2^k - 1)$.

9.8.3.1 QAM Probability of Bit Error

For a rectangular constellation, a Gaussian channel, and matched filter reception, the bit-error probability for M-QAM, where $M = 2^k$ and k is even, is {12}

$$P_B \approx \frac{2(1 - L^{-1})}{\log_2 L} Q\left[\sqrt{\left(\frac{3\log_2 L}{L^2 - 1}\right)\frac{2E_b}{N_0}}\right] \tag{9.54}$$

where $Q(x)$ is as defined in Equation (3.43) and $L = \sqrt{M}$ represents the number of amplitude levels in one dimension. In the context of L-PAM, a sequence of $k/2 = \log_2 L$ bits are assigned to an L-ary symbol using a Gray code (defined in Section 4.9.4).

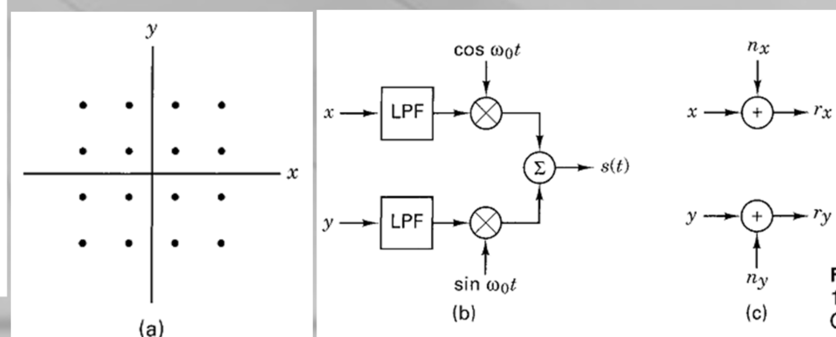


Figure 9.16 QAM modulation. (a) 16-ary signal space. (b) Canonical QAM modulator. (c) QAM channel model.

Defining, Designing, and Evaluating Digital Communication Systems

9.8.3.2 Bandwidth–Power Trade-Off

The bandwidth–power trade-off of M-ary QAM at a bit error probability of 10^{-5} is displayed on the bandwidth-efficiency plane in Figure 9.6, with the abscissa measured in average E_b/N_0. We assume Nyquist filtering of the baseband pluses so that the DSB transmission bandwidth at IF is $W_{IF} = 1/T$, where T is the symbol duration. Thus the bandwidth efficiency is $R/W = \log_2 M$, where M is the symbol set size. For realistic channels and waveforms, the performance must be reduced to account for the increased bandwidth necessary to implement realizable filters. From Figure 9.6 it can be seen that QAM represents a method of reducing the bandwidth required for the transmission of digital data. As with M-ary PSK, bandwidth efficiency can be exchanged for power or E_b/N_0; however, in the case of QAM, a *much more efficient exchange* is possible than in the case of M-ary PSK.

Example 9.5 Waveform Design

Assume that a data stream with data rate $R = 144$ Mbits/s is to be transmitted on an RF channel using a DSB modulation scheme. Assume Nyquist filtering and an allowable DSB bandwidth of 36 MHz. Which modulation technique would you choose for this requirement? If the available E_b/N_0 is 20, what would be the resulting probability of bit error?

Solution

The required spectral efficiency is

$$\frac{R}{W} = \frac{144 \text{ Mbits/s}}{36 \text{ MHz}} = 4 \text{ bits/s/Hz}$$

From Figure 9.6 we note that 16-ary QAM, with a theoretical spectral efficiency of 4 bits/s/Hz, requires a lower E_b/N_0 than that of 16-ary PSK for the same P_B. Based on these considerations we choose a 16-ary QAM modem.

With the available E_b/N_0 given as 20, we use Equation (9.54) to calculate the expected bit error probability as

$$P_B \approx \frac{3}{4} Q\left(\sqrt{\frac{4}{5}\frac{E_b}{N_0}}\right) = 2.5 \times 10^{-5}$$

Example 9.6 Spectral Efficiency

(a) Explain the computation of the QAM spectral efficiency in Example 9.5, considering that QAM is transmitted on orthogonal components of a carrier wave.

(b) Since the DSB bandwidth is 36 MHz in Example 9.5, consider using half that amount at baseband to transmit the 144-Mbits/s data stream, using multilevel PAM. What is the spectral efficiency needed to accomplish this, and how many levels of PAM would be required? Assume Nyquist filtering.

Solution

(a) *Bandpass channel using QAM:* The 144-Mbits/s data stream is partitioned into a 72-Mbits/s in-phase and a 72-Mbits/s quadrature stream; one stream amplitude-modulates the cosine component of a carrier over a bandwidth of 36 MHz, and the other stream amplitude-modulates the sine component of the carrier wave over the same 36-MHz bandwidth. Since each 72-Mbits/s stream modulates an orthogonal component of the carrier, the 36 MHz suffices for both streams, or for the full 144 Mbits/s. Thus the spectral efficiency is (144 Mbits/s)/36 MHz = 4 bits/s/Hz.

(b) *Required spectral efficiency at baseband*

$$\frac{R}{W} = \frac{144 \text{ Mbits/s}}{18 \text{ MHz}} = 8 \text{ bits/s/Hz}$$

Assuming Nyquist filtering, a bandwidth of 18 MHz can support a maximum symbol rate of $R_s = 2W = 36$ megasymbols/s [see Equation (3.80)]. Each PAM pulse must therefore have an ℓ-bit meaning, such that

$$R = \ell R_s$$

Hence,

$$\ell = \frac{144 \text{ Mbits/s}}{36 \text{ megapulses/s}} = 4 \text{ bits/pulse}$$

where $\ell = \log_2 L$, and $L = 16$ levels.

Defining, Designing, and Evaluating

9.9 MODULATION AND CODING FOR BANDLIMITED CHANNELS

The channel coding techniques of Chapters 6–8 have generally *not* been associated with voice-grade telephone channels (although the first field test of sequential decoding of convolutional codes was on a telephone line). Recently, however, there has been considerable interest in techniques that can provide coding gain for bandlimited channels. The motivation is to enable the reliable transmission of *higher data rates* over voice-grade channels. The potential gain is about 3 bits/symbol (for a given signal-to-noise ratio) [21] or, alternatively, a given error performance could be achieved with a power savings of 9 dB [21].

The greatest interest is in the following three separate coding research areas:

1. Optimum signal constellation boundaries (choosing a closely packed signal subset from any regular array or lattice of candidate points)
2. Higher-density lattice structures (adding improvement to the signal subset choice by starting with the densest possible lattice for the space)
3. Trellis-coded modulation (combined modulation and coding techniques for obtaining coding gain for bandlimited channels)

The first two areas are not "true" error control coding schemes. By "true error control coding" we refer to those techniques that employ some structured redundancy to improve the error performance. Only the third technique, trellis-coded modulation, involves redundancy. Each of these coding research areas and their expected performance improvements are discussed below.

9.9.1 Commercial Telephone Modems

The use of efficient modulation techniques has traditionally been spearheaded by the telecommunications industry, since the telephone company's foremost resource consists of sharply bandlimited voice-grade channels. The typical telephone channel is characterized by a high signal-to-noise ratio (SNR) of approximately 30 dB and a bandwidth of approximately 3 kHz. Table 9.4 lists the evolution of leased-line telephone modems, and Table 9.5 lists the evolution of dial-line telephone modem standards.

TABLE 9.4 Evolution of Leased-Line Telephone Modems

Year	Name	Maximum Bit Rate (bits/s)	Signaling Rate (symbols/s)	Modulation Technique	Signaling Efficiency (bits/symbol)
1962	Bell 201	2400	1200	4-PSK	2
1967	Milgo 4400/48	4800	1600	8-PSK	3
1971	Codex 9600C	9600	2400	16-QAM	4
1980	Paradyne MP14400	14,400	2400	64-QAM	6
1981	Codex SP14.4	14,400	2400	64-QAM	6
1984	Codex 2660	16,800	2400	Trellis-coded 256-QAM	7
1985	Codex 2680	19,200	2743	8-D Trellis-coded 160-QAM	7

TABLE 9.5 Evolution of Dial-Line Telephone Modem Standards

Year	Name	Maximum Bit Rate (bits/s)	Signaling Rate (symbols/s)	Modulation Technique	Signaling Efficiency (bits/symbol)
1984	V.32	9600	2400	2-D Trellis Coded 32-QAM	4
1991	V.32bis	14,400	2400	2-D Trellis Coded 128-QAM	6
1994	V.34	28,800	2400, 2743, 2800, 3000, 3200, 3429	4-D Trellis Coded 960-QAM	≈ 9
1996	V.34	33,600	2400, 2743, 2800, 3000, 3200, 3429	4-D Trellis Coded 1664-QAM	≈ 10
1998	4.90	downstream: 56,000 upstream: 33,600	8000 as in V.34	PCM* (M-PAM) as in V.34	7 ≈ 10
2000	V.92	downstream: 56,000 upstream: 48,000	8000 8000	PCM* (M-PAM) Trellis Coded PCM*	7 6

*In the G.711 ITU-T Recommendation, PCM is the term used for M-ary PAM signaling.

Defining, Designing, and Evaluating Digital Communication Systems

9.9.2 Signal Constellation Boundaries

Several researchers [22–26] have examined large numbers of possible QAM signal constellations in a search for designs that result in the best error performance for a given average signal-to-noise ratio. Figure 9.17 illustrates some examples of symbol constellations for $M = 4, 8$, and 16 that have been considered [22]. The circular sets are designated by the notation $(a, b, \ldots)$, where there are a quantity of a signals on the inner circle, b signals on the next circle, and so on. In general, the constellation rule, known as the Campopiano-Glazer construction rule [24], that yields optimum signal set performance can be summarized as follows: From an infinite array of points closely packed in a *regular array or lattice*, select a closely packed subset of 2^k points as a signal constellation. In this case "optimum" means minimum average or peak power for a given error probability. In a two-dimensional signal space the optimum boundary surrounding an array of points tends toward a circle. Figure 9.18 illustrates examples of 64-ary ($k = 6$) and 128-ary ($k = 7$) signal sets from a rectangular array. The cross-shaped boundaries are a compromise to the optimum circle. The $k = 6$ constellation was used in the Paradyne 14.4-kbits/s modem. Compared with a square, the performance improvement resulting from a circular boundary is only a modest 0.2 dB [21].

9.9.3 Higher-Demensional Signal Constellations

For any particular information rate and channel-noise process that is independent and identically distributed in two dimensions, signaling in a two-dimensional space can provide the same error performance with less average (or peak) power than signaling in a one-dimensional pulse-amplitude (PAM) space. This is accomplished by choosing signaling points on a two-dimensional lattice from within a circular rather than a rectangular boundary. In the same way, by going to a higher number, N dimensions, and choosing points on an n-dimensional lattice from within an N-sphere rather than an N-cube, further energy savings are possible [27–30]. The goal of this constellation shaping is to make the required average energy of signal points from the N-sphere less than that from the N-cube; such reduction in required energy for a given error performance is referred to as a *shaping gain* [16]. Table 9.6 gives the energy savings possible in N dimensions. As N goes to infinity, the gain goes to 1.53 dB; it is not difficult to achieve shaping gain on the order of 1 dB [16, 21].

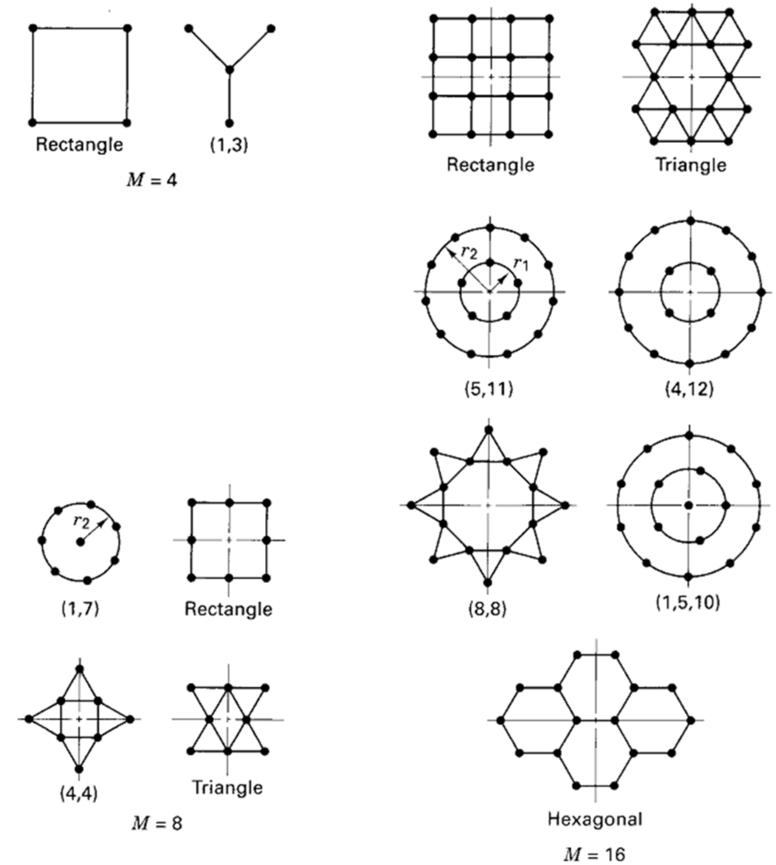


Figure 9.17 *M*-ary symbol constellations. (Reprinted with permission from C. M. Thomas, M. Y. Weidner, and S. H. Durrani, "Digital Amplitude-Phase Keying with *M*-ary Alphabets," *IEEE Trans. Commun.*, vol. COM22, no. 2, Feb. 1974, Figs. 2 and 3, p. 170. © 1974 IEEE.)

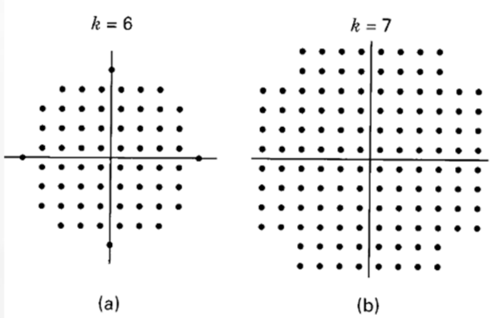


Figure 9.18 Examples of M-ary constellations using a rectangular array.

TABLE 9.6 Energy Savings from N-Sphere Mapping versus N-Cube Mapping (Shaping Gain)

Dimensions (N)	N-Sphere Mapping Gain (dB)
2	0.20
4	0.45
8	0.73
16	0.98
24	1.10
32	1.17
48	1.26
64	1.31

Source: G. D. Forney, Jr., et al., "Efficient Modulation for Bandlimited Channels," *IEEE J. Sel. Areas Commun.*, vol. SAC2, no. 5, September 1984, pp. 632–647.

The channel is essentially two-dimensional, since symbols represented as points on a two-dimensional plane are transmitted in quadrature fashion. Multidimensional signaling is generally taken to mean signaling with two or more such planes. For transmitting n bits/symbol with N-dimensional (N even and greater than 2) signaling, the incoming bits are grouped into blocks of $nN/2$. A mapping must then be made that assigns data bits to $2^{nN/2}$ N-dimensional vectors that have the least energy among all such vectors. A corresponding inverse mapping must be made at the receiver.

Consider an example of mapping signals from a two-dimensional to a four-dimensional space. We start with a two-dimensional M-ary constellation, such as M-QAM with $M = 16$. Here the transmitted symbol, viewed as a point on a plane, is represented by $n = 4$ bits (two 4-ary amplitudes, and two bits per amplitude). Each symbol transmission consists of sending a vector from a space of 16 possible vectors. With four-dimensional signaling, the transmitted symbol, viewed as two points, one from each of two planes, is represented by 8 bits. Then, each (two-point) transmission consists of sending a vector from a space of $16 \times 16 = 256$ vectors. In general, the source-data bits are grouped into blocks of $nN/2$. In this example for four-dimensional signaling, we group the source-data bits into blocks of 8 bits (2 planes $\times$ n = 4 bits/plane). Such an 8-bit transmission can be viewed as a mapping from a space of 2^n two-dimensional vectors to a space of $2^{nN/2}$ four-dimensional vectors. For the four-dimensional system depicted in Figure 9.19, a given source produces one of 256 four-dimensional vectors m_i ($i = 1, 2, \ldots, 256$) by grouping two 16-ary symbols (two planes) at a time and transmitting waveforms, $a_js(t)$, $b_js(t)$, $c_js(t)$, $d_js(t)$, where $j = 1, \ldots, 4$ represents one of 4-ary amplitude values. These baseband or bandpass waveforms are transmitted on separate noninterfering channels. In each channel, the waveforms are distorted by independent AWGN, and at the receiver they are demodulated with matched filters. We may choose to transmit the N-dimensional signal in a number of ways:

1. Using four separate wires representing four baseband channels.
2. Using two bandpass channels, each with separately modulated inphase and quadrature components.
3. Using time- or frequency-division multiplexing to carry the baseband or bandpass channels on a common transmission line.
4. Using orthogonal electromagnetic wave polarization.

Thus, if the Figure 9.19 example represents a radio system, we could follow method 2 above and modulate waveforms $a_js(t)$, and $b_js(t)$ in quadrature fashion onto a particular carrier wave, while modulating waveforms $c_js(t)$, and $d_js(t)$ onto a second carrier wave. Thus, during each $2T$ second interval, one would transmit four 4-ary numbers

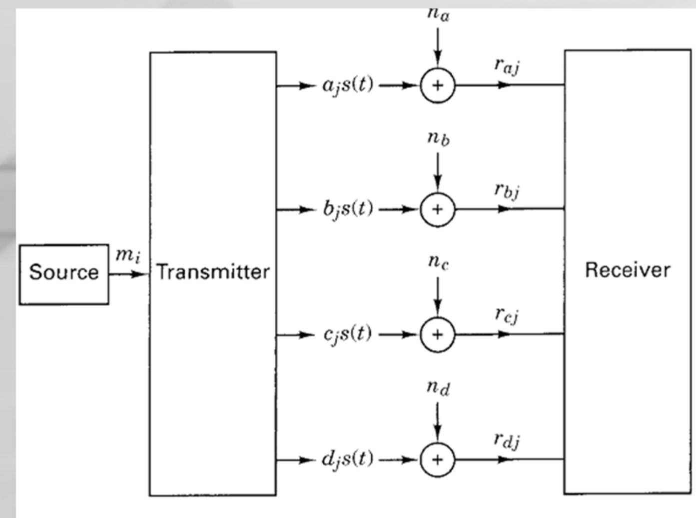
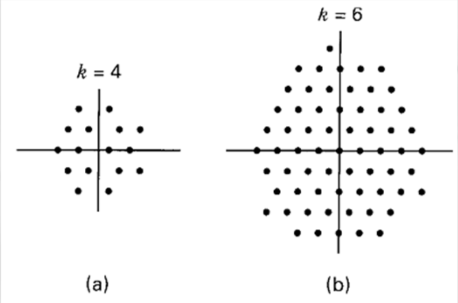


Figure 9.19 Four-dimensional system configuration.

representing 8 bits or a vector from a 256-ary space. Further shaping gain can be similarly achieved for the delivery of 16-ary symbols per plane with six-dimensional signaling, where every $3T$ seconds, a 16-ary symbol from each of three planes is transmitted. Thus, each six-dimensional signal constitutes three 16-ary values representing 12 bits or a point on a 4096-ary signal space. It is important to emphasize that it is not the mere grouping of the 16-ary symbols that brings about the shaping gain. The gain comes about because detection performed over a larger signal space can achieve a given error performance with a smaller E_b/N_0. In the case of sending 16-ary symbols with six-dimensional signaling, a 12-bit sequence is detected every $3T$ seconds (not a 4-bit sequence every T seconds). Detection in a higher dimensional space entails greater (signal-mapping) complexity. Compromises are generally used to simplify the mapping complexity at the cost of some suboptimality in energy efficiency.

Defining, Designing, and Evaluating Digital Communication Systems



9.9.4 Higher-Density Lattice Structures

In Section 9.9.3, we discussed the selection of a closely packed subset of points from any regular array or lattice. Here we consider the added improvement by starting with the *densest possible lattice* in the space. In a two-dimensional signal space, the densest lattice is the hexagonal lattice (try penny packing). The result of employing a hexagonal lattice instead of a rectangular one, such as those shown in Figure 9.18, can be a 0.6-dB savings in average energy. Figure 9.20 illustrates some examples of hexagonal packing. The strange-looking $k = 4$ constellation in Figure 9.20a was discovered by Foschini et al. [26] and is still the best 16-ary constellation known. The $k = 6$ constellation in Figure 9.20b was used in the Codex SP14.4 modem.

The hexagonal lattice is optimum for two dimensions. For higher dimensions there are other lattice structures that provide the densest packing. Table 9.7 gives the gain over the rectangular lattice, in decibels, due to the densest packings currently known for various dimensions.

9.9.5 Combined Gain: *N*-Sphere Mapping and Dense Lattice

It is possible to combine the benefits of the Campopiano–Glazer boundary construction in N dimensions with the gain from the densest lattice in N-space. The resulting gain is a combination of N-sphere versus N-cube boundary gain of Table 9.6 and the lattice packing density gain of Table 9.7. The combined energy savings are shown in Table 9.8.

TABLE 9.7 Energy Savings from Dense Lattices versus the Rectangular Lattice

Dimensions (N)	Dense Lattice Gain (dB)
2	0.62
4	1.51
8	3.01
16	4.52
24	6.02
32	6.02
48	7.78
64	8.09

Source: G. D. Forney, Jr., et al., "Efficient Modulation for Bandlimited Channels," *IEEE J. Sel. Areas Commun.*, vol. SAC2, no. 5, September 1984, pp. 632–647.

TABLE 9.8 Combined Energy Savings from *N*-Sphere Mapping and Dense Lattices

Dimensions (N)	Combined Savings Gain (dB)
2	0.82
4	1.96
8	3.74
16	5.50
24	7.12
32	7.19
48	9.04
64	9.40

Source: G. D. Forney, Jr., et al., "Efficient Modulation for Bandlimited Channels," *IEEE J. Sel. Areas Commun.*, vol. SAC2, no. 5, September 1984, pp. 632–647.

9.10 TRELLIS-CODED MODULATION

The error-correction codes described in Chapters 6–8, when used in real-time communication systems, provide improvements in error performance at the cost of bandwidth expansion. For both block codes and convolutional codes, transforming each input data k-tuple into a larger output codeword n-tuple, requires additional transmission bandwidth. Therefore, in the past, coding generally was not popular for bandlimited channels such as telephone channels, where signal bandwidth expansion is not practical. Since about 1984, however, there has been active interest in combined modulation and coding schemes, called *trellis-coded modulation* (TCM), that achieve error-performance improvements without expansion of signal bandwidth. TCM schemes use redundant nonbinary modulation in combination with a *finite-state machine* (the encoder). What is a finite-state machine, and what is meant by its state? Finite-state machine is the general name given to a device that has a memory of past signals; the adjective *finite* refers to the fact that there are only a finite number of unique states that the machine can encounter. What is meant by the *state* of a finite-state machine? In the most general sense, the state consists of the smallest amount of information that, together with a current input to the machine, can predict the output of the machine. The state provides some knowledge of the past signaling events and the restricted set of possible outputs in the future. A future state is restricted by the past state.

For each symbol interval, a TCM finite-state encoder selects one of a set of waveforms, thereby generating a sequence of coded waveforms to be transmitted. The noisy received signals are detected and decoded by a soft-decision maximum-likelihood detector/decoder. In conventional systems involving modulation and coding, it is common to separately describe and implement the detector and the decoder. With TCM systems, however, these functions must be treated jointly. Coding gain can be achieved without sacrificing data rate or without increasing either bandwidth or power [6, 31]. At first, it may seem that this statement violates some basic principle of power-bandwidth, error-probability trade-off. However, there is still a trade-off involved, since TCM achieves coding gain at the expense of decoder complexity.

Trellis-coded modulation combines a multilevel/phase modulation signaling set with a *trellis-coding scheme.* The term "trellis-coding scheme" refers to any code system that has memory (a finite state machine), such as a convolutional code. Multilevel/phase signals have constellations involving multiple amplitudes, multiple phases, or combinations of multiple amplitudes and multiple phases. In other words, a TCM signal set is best represented by any signal set (greater than binary) whose vector representations can be depicted on a plane, such as that shown in Figure 9.16a for QAM signals. A trellis-coding scheme is one that can be characterized with a state-transition (trellis) diagram, similar to the trellis diagrams describing convolutional codes. The convolutional codes presented in Chapter 7 are linear, although trellis codes are not constrained to be linear. Coding gains can be realized with block codes or trellis codes, but only trellis codes will be considered because the availability of the *Viterbi decoding algorithm* makes trellis decoding simple and efficient. Ungerboeck showed that in the presence of AWGN, TCM schemes can yield net coding gains of about 3-dB relative to uncoded systems with relative ease, while gains of about 6-dB can be achieved with greater complexity.

9.10.1 The Idea Behind Trellis-Coded Modulation

In TCM, channel coding and modulation are performed together; where one begins and the other ends cannot be easily established. Can you speculate what ideas might have prompted the development of TCM? Perhaps it started with the notion that "not all signal subsets (in a constellation) have equal distance properties." That is, for a nonorthogonal signal set, such as MPSK, antipodal signals have the best distance properties for easily discriminating one signal from the other, while nearest neighbor signals have relatively poor distance properties. It may be that the initial idea of coded modulation came about by trying to exploit these differences.

A simple analogy might be helpful in understanding the overall goals in TCM. Imagine that there is an all-knowing wizard at the transmitter. As the message bits enter the system, the wizard recognizes that some of the bits are most vulnerable to the degradation effects of channel impairments; hence, they are assigned modulation waveforms associated with the best distance properties. Similarly, other bits are judged to be very robust, and hence, they are assigned waveforms with poorer distance properties. Modulation and coding take place together. The wizard is assigning waveforms to bits (modulation), but, the assignment is being performed according to the criterion of better or worse distance properties (channel coding).

9.10.1 The Idea Behind Trellis-Coded Modulation

In TCM, channel coding and modulation are performed together; where one begins and the other ends cannot be easily established. Can you speculate what ideas might have prompted the development of TCM? Perhaps it started with the notion that "not all signal subsets (in a constellation) have equal distance properties." That is, for a nonorthogonal signal set, such as MPSK, antipodal signals have the best distance properties for easily discriminating one signal from the other, while nearest neighbor signals have relatively poor distance properties. It may be that the initial idea of coded modulation came about by trying to exploit these differences.

A simple analogy might be helpful in understanding the overall goals in TCM. Imagine that there is an all-knowing wizard at the transmitter. As the message bits enter the system, the wizard recognizes that some of the bits are most vulnerable to the degradation effects of channel impairments; hence, they are assigned modulation waveforms associated with the best distance properties. Similarly, other bits are judged to be very robust, and hence, they are assigned waveforms with poorer distance properties. Modulation and coding take place together. The wizard is assigning waveforms to bits (modulation), but, the assignment is being performed according to the criterion of better or worse distance properties (channel coding).

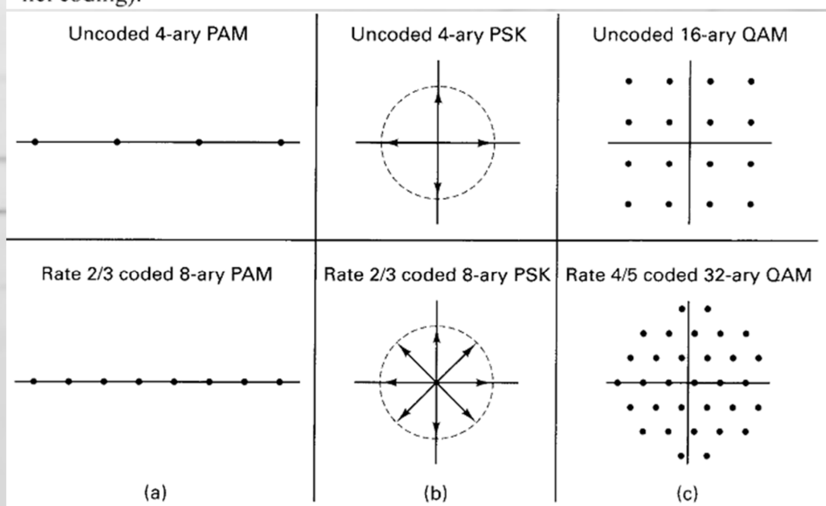


Figure 9.21 Increase of signal set size for trellis-coded modulation.

9.10.1.1 Increasing Signal Redundancy

TCM may be implemented with a convolutional encoder, wherein k current bits and $K - 1$ prior bits are used to produce $n = k + p$ code bits, where K is the encoder constraint length (see Chapter 7) and p is the number of parity bits. Notice that encoding increases the signal set size from 2^k to 2^{k+p}. Ungerboeck [31] investigated the increase in channel capacity achievable by signal set expansion, and concluded that most of the achievable coding gain over conventional uncoded multilevel modulation could be realized by expanding the uncoded signaling set by the factor of two ($p = 1$). This can be accomplished by encoding with a rate $k/(k + 1)$ code and subsequently mapping groups of $(k + 1)$ bits into the set of 2^{k+1} waveforms. Figure 9.21a illustrates an uncoded 4-ary PAM signal set, before and after being rate 2/3 encoded into an 8-ary PAM signal set. Similarly, Figure 9.21b illustrates an uncoded 4-ary PSK (QPSK) signal set, before and after being rate 2/3 encoded into an 8-ary PSK signal set. Similarly, Figure 9.21c illustrates an uncoded 16-ary QAM signal set, before and after being rate 4/5 encoded into a 32-ary QAM signal set. In each of the cases shown in Figure 9.21, the system is configured to use the same average signal power before and after coding. Also, to provide the needed redundancy for coding, the signaling set is increased from $M = 2^k$ to $M' = 2^{k+1}$. Thus, $M' = 2M$; however, the increase in the alphabet size does not result in an increase in required bandwidth. Recall from Section 9.7.2 that the transmission bandwidth of nonorthogonal signaling does not depend on the density of signaling points in the constellation; it depends only on the rate of signaling. The expanded signal set *does* result in a reduced distance between adjacent symbol points (for signal sets with a constant average power), as seen in Figure 9.21. In an *uncoded* system, such a reduced distance degrades the error performance. However, because of the redundancy introduced by the code, this reduced distance no longer plays a critical role in determining the error performance. Instead, the *free distance,* which is the minimum distance between members of the set of *allowed* code sequences, determines the error performance. The free distance characterizes the "easiest way" for the decoder to make an error. (See Section 9.10.3.1.) Whenever there is a code at work, the signaling space is not the proper place for examining the error-performance advantage that coding can achieve. This is because the code is defined by rules and constraints that are not visible in the signaling space. When two signals are in close proximity in the signaling space of a coded system, their closeness may not have much performance significance because the rules of the code may not allow for the transitioning between two such vulnerable signal points. Where is the proper place for evaluating allowed code sequences and distance properties? It is the trellis diagram. Using this diagram, the objective of TCM is to assign waveforms to trellis transitions so as to increase the free distance between the waveforms that are the most likely to be confused.

Modulation and Coding Trade-Off

CONCLUSION

In this chapter we have integrated some of the ideas in earlier Chapters dealing with modulation and coding. We have reviewed the basic system design goals: to maximize data rate while simultaneously minimizing error probability, bandwidth, E_b/N_0, and complexity. We examined the trade-offs heuristically on two performance planes: the error probability plane and the bandwidth efficiency plane. The former explicitly illustrates the P_B versus E_b/N_0 trade-offs while only implicitly displaying the bandwidth expenditure. The latter explicitly illustrates the R/W versus E_b/N_0 trade-offs while only implicitly displaying the P_B performance. We outlined typical steps that need to be considered in meeting bandwidth power, and error-performance requirements of a digital communication system, and we discussed some of the basic constraints to improvement without limit. The Nyquist criterion establishes that we cannot continue to reduce system bandwidth indefinitely. There is a theoretical limitation; in order to transmit R_s symbols/second without intersymbol interference, we must utilize a minimum of $R_s/2$ hertz of bandwidth. The Shannon–Hartley theorem relates to the power–bandwidth trade-off and results in another important limitation, the Shannon limit. The Shannon limit of −1.6 dB is the theoretical minimum amount of E_b/N_0 that is necessary (in concert with channel coding) to achieve an arbitrarily low error probability over an AWGN channel. The more general limitation is the channel capacity, above which there cannot be error-free signaling. We have also examined some of the bandwidth-efficient modulation schemes, such as minimum shift keying (MSK), quadrature amplitude modulation (QAM), and trellis-coded modulation. The latter technique offers an attractive way to obtain coding gain without paying the price of additional bandwidth.

Problems and Questions

- Problems

9.2. Consider that a 100-kbits/s data stream is to be transmitted on a voice-grade telephone circuit (with a bandwidth of 3 kHz). Is it possible to approach error-free transmission with a SNR of 10 dB? Justify your answer. If it is not possible, suggest system modifications that might be made.

9.3. Consider a source that produces six messages with probabilities $\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{32}$, and $\frac{1}{32}$. Determine the average information content in bits, of a message.

9.8. Suppose a binary noncoherent FSK link has a maximum data rate of 2.4 kbits/s without ISI over a channel whose nominal bandwidth is 2.4 kHz. Suggest ways of increasing the data rate under the following system constraints.

(a) The system is power limited.
(b) The system is bandwidth limited.
(c) The system is both power and bandwidth limited.

9.17. Consider a telephone modem operating at 28.8 kbits/s that uses trellis-coded QAM modulation.

(a) Calculate the bandwidth efficiency of such a modem, assuming that the usable channel bandwidth is 3429 Hz.
(b) Assuming AWGN and an available E_b/N_0 = 10 dB, calculate the theoretically available capacity in the 3429-Hz bandwidth.
(c) What is the required E_b/N_0 that will enable a 3429-Hz bandwidth to have a capacity of 28.8 kbits/s?

9.14. Consider a real-time communication satellite system, operating over an AWGN channel (disturbed by periodic fades). The overall link is described by the following specifications from a mobile transmitter to a low-earth-orbit satellite receiver:

Data rate $R = 9600$ bits/s
Available bandwidth $W = 3000$ Hz
Link margin $M = 0$ dB (see Section 5.6)
Carrier frequency $f_c = 1.5$ GHz
EIRP = 6 dBW
Distance between transmitter and receiver $d = 1000$ km
Satellite receiver figure of merit $G/T = 30$ dBI
Receiver antenna temperature $T_A^\circ = 290\ K$
Line loss from the receiver antenna to the receiver, $L = 3$ dB
Receiver noise figure $F = 10$ dB
Losses due to fading $L_f = 20$ dB
Other losses $L_o = 6$ dB

You are allowed to choose one of two modulation schemes—MPSK with Gray coding, or noncoherent orthogonal MFSK—such that the available bandwidth is not exceeded and power is conserved. For error-correction coding, you are to choose one of the (127, k) BCH codes from Table 9.2 that provides the most redundancy, but still meets the bandwidth constraints. Calculate the output decoded bit-error probability. How much coding gain, if any, characterizes your choices. Hint: Proceed by calculating parameters in the following order, E_b/N_0, E_s/N_0, $P_E(M)$, p_c, P_B. When using Equation (9.41) for computing decoded bit-error probability, a small E_b/N_0 necessitates using many terms in the summation. Hence, computer assistance is helpful here.

Problems and Questions

- Questions

9.1. Why do binary and 4-ary orthogonal frequency shift keying (4-FSK) manifest the same *bandwidth-efficiency* relationship? (See Section 9.5.1.)

9.2. For MPSK modulation, *bandwidth efficiency* increases with higher-dimensional signaling, but for MFSK, it decreases. Explain why this is the case. (See Sections 9.7.2 and 9.7.3.)

9.3. Consider the assortment of signaling elements that flow through a typical system, and describe the subtle energy and rate transformations among them: from data-bits to channel-bits to symbols to chips. (See Section 9.7.7.)

9.4. The steep decrease in MSK *spectral sidelopes* in Figure 9.15 illustrates why MSK is considered to be more spectrally efficient then QPSK. How do you explain then that the QPSK spectrum has a *narrower mainlobe* than the MSK spectrum? (See Section 9.8.2.)

9.5. In Chapter 4, we presented binary phase shift keying (BPSK) and quaternary phase shift keying (QPSK) as manifesting the same bit-error-probability relationship. (See Section 4.8.4.) Does the same hold true for M-ary pulse amplitude modulation (M-PAM) and M^2-ary quadrature amplitude modulation (M^2-QAM); that is, do these schemes also manifest the same bit-error probability? (See Section 9.8.3.1.)

9.6 Although trellis-coded modulation schemes do not require additional bandwidth or power, their use still involves a *trade-off.* What is the cost of achieving coding gain with TCM? (See Section 9.10.)

9.7 What is meant by the *state* of a finite-state machine? (See Section 9.10.)