The parameter $h(Y|X)$ is the conditional differential entropy of Y, given X; it is defined in a manner similar to $h(X|Y)$.

9.10 *Information Capacity Theorem*

In this section, we use the idea of mutual information to formulate the information capacity theorem for *band-limited, power-limited Gaussian channels*. To be specific, consider a zero-mean stationary process $X(t)$ that is band-limited to B hertz. Let X_k, $k = 1, 2, \ldots, K$, denote the continuous random variables obtained by uniform sampling of the process $X(t)$ at the Nyquist rate of $2B$ samples per second. These samples are transmitted in T seconds over a noisy channel, also band-limited to B hertz. Hence, the number of samples, K, is given by

$$K = 2BT \tag{9.83}$$

We refer to X_k as a sample of the *transmitted signal*. The channel output is perturbed by *additive white Gaussian noise* (AWGN) of zero mean and power spectral density $N_0/2$. The noise is band-limited to B hertz. Let the continuous random variables Y_k, $k = 1, 2, \ldots, K$ denote samples of the received signal, as shown by

$$Y_k = X_k + N_k, \qquad k = 1, 2, \ldots, K \tag{9.84}$$

The noise sample N_k is Gaussian with zero mean and variance given by

$$\sigma^2 = N_0 B \tag{9.85}$$

We assume that the samples Y_k, $k = 1, 2, \ldots, K$ are statistically independent.

A channel for which the noise and the received signal are as described in Equations (9.84) and (9.85) is called a *discrete-time, memoryless Gaussian channel*. It is modeled as in Figure 9.13. To make meaningful statements about the channel, however, we have to assign a *cost* to each channel input. Typically, the transmitter is *power limited*; it is therefore reasonable to define the cost as

$$E[X_k^2] = P, \qquad k = 1, 2, \ldots, K \tag{9.86}$$

where P is the *average transmitted power*. The *power-limited Gaussian channel* described herein is of not only theoretical but also practical importance in that it models many communication channels, including line-of-sight radio and satellite links.

The *information capacity* of the channel is defined as the maximum of the mutual information between the channel input X_k and the channel output Y_k over all distributions on the input X_k that satisfy the power constraint of Equation (9.86). Let $I(X_k; Y_k)$ denote

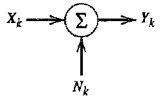

FIGURE 9.13 Model of discrete-time, memoryless Gaussian channel.

the mutual information between X_k and Y_k. We may then define the information capacity of the channel as

$$C = \max_{f_{X_k}(x)} \{I(X_k; Y_k) : E[X_k^2] = P\} \tag{9.87}$$

where the maximization is performed with respect to $f_{X_k}(x)$, the probability density function of X_k.

The mutual information $I(X_k; Y_k)$ can be expressed in one of the two equivalent forms shown in Equation (9.81). For the purpose at hand, we use the second line of this equation and so write

$$I(X_k; Y_k) = h(Y_k) - h(Y_k | X_k) \tag{9.88}$$

Since X_k and N_k are independent random variables, and their sum equals Y_k, as in Equation (9.84), we find that the conditional differential entropy of Y_k, given X_k, is equal to the differential entropy of N_k (see Problem 9.28):

$$h(Y_k \mid X_k) = h(N_k) \tag{9.89}$$

Hence, we may rewrite Equation (9.88) as

$$I(X_k; Y_k) = h(Y_k) - h(N_k) \tag{9.90}$$

Since $h(N_k)$ is independent of the distribution of X_k, maximizing $I(X_k; Y_k)$ in accordance with Equation (9.87) requires maximizing $h(Y_k)$, the differential entropy of sample Y_k of the received signal. For $h(Y_k)$ to be maximum, Y_k has to be a Gaussian random variable (see Example 9.8). That is, the samples of the received signal represent a noiselike process. Next, we observe that since N_k is Gaussian by assumption, the sample X_k of the transmitted signal must be Gaussian too. We may therefore state that the maximization specified in Equation (9.87) is attained by choosing the samples of the transmitted signal from a noiselike process of average power P. Correspondingly, we may reformulate Equation (9.87) as

$$C = I(X_k; Y_k) : X_k \text{ Gaussian}, \qquad E[X_k^2] = P \tag{9.91}$$

where the mutual information $I(X_k; Y_k)$ is defined in accordance with Equation (9.90).

For the evaluation of the information capacity C, we proceed in three stages:

1. The variance of sample Y_k of the received signal equals $P + \sigma^2$. Hence, the use of Equation (9.76) yields the differential entropy of Y_k as

$$h(Y_k) = \tfrac{1}{2} \log_2[2\pi e(P + \sigma^2)] \tag{9.92}$$

2. The variance of the noise sample N_k equals σ^2. Hence, the use of Equation (9.76) yields the differential entropy of N_k as

$$h(N_k) = \tfrac{1}{2} \log_2(2\pi e\sigma^2) \tag{9.93}$$

3. Substituting Equations (9.92) and (9.93) into Equation (9.90) and recognizing the definition of information capacity given in Equation (9.91), we get the desired result:

$$C = \frac{1}{2} \log_2\left(1 + \frac{P}{\sigma^2}\right) \text{ bits per transmission} \tag{9.94}$$

With the channel used K times for the transmission of K samples of the process $X(t)$ in T seconds, we find that the information capacity per unit time is (K/T) times the result

given in Equation (9.94). The number K equals $2BT$, as in Equation (9.83). Accordingly, we may express the information capacity in the equivalent form:

$$C = B \log_2\left(1 + \frac{P}{N_0 B}\right) \text{ bits per second} \tag{9.95}$$

where we have used Equation (9.85) for the noise variance σ^2.

Based on the formula of Equation (9.95), we may now state Shannon's third (and most famous) theorem, the *information capacity theorem*,[10] as follows:

The information capacity of a continuous channel of bandwidth B hertz, perturbed by additive white Gaussian noise of power spectral density $N_0/2$ and limited in bandwidth to B, is given by

$$C = B \log_2\left(1 + \frac{P}{N_0 B}\right) \text{ bits per second}$$

where P is the average transmitted power.

The information capacity theorem is one of the most remarkable results of information theory for, in a single formula, it highlights most vividly the interplay among three key system parameters: channel bandwidth, average transmitted power (or, equivalently, average received signal power), and noise power spectral density at the channel output. The dependence of information capacity C on channel bandwidth B is *linear*, whereas its dependence on signal-to-noise ratio P/N_0B is *logarithmic*. Accordingly, *it is easier to increase the information capacity of a communication channel by expanding its bandwidth than increasing the transmitted power for a prescribed noise variance.*

The theorem implies that, for given average transmitted power P and channel bandwidth B, we can transmit information at the rate of C bits per second, as defined in Equation (9.95), with arbitrarily small probability of error by employing sufficiently complex encoding systems. It is not possible to transmit at a rate higher than C bits per second by any encoding system without a definite probability of error. Hence, the channel capacity theorem defines the *fundamental limit* on the rate of error-free transmission for a power-limited, band-limited Gaussian channel. To approach this limit, however, the transmitted signal must have statistical properties approximating those of white Gaussian noise.

Sphere Packing[11]

To provide a plausible argument supporting the information capacity theorem, suppose that we use an encoding scheme that yields K code words, one for each sample of the transmitted signal. Let n denote the length (i.e., the number of bits) of each code word. It is presumed that the coding scheme is designed to produce an acceptably low probability of symbol error. Furthermore, the code words satisfy the power constraint; that is, the average power contained in the transmission of each code word with n bits is nP, where P is the average power per bit.

Suppose that any code word in the code is transmitted. The received vector of n bits is Gaussian distributed with mean equal to the transmitted code word and variance equal to $n\sigma^2$, where σ^2 is the noise variance. With high probability, the received vector lies inside a sphere of radius $\sqrt{n\sigma^2}$, centered on the transmitted code word. This sphere is itself contained in a larger sphere of radius $\sqrt{n(P + \sigma^2)}$, where $n(P + \sigma^2)$ is the average power of the received vector.

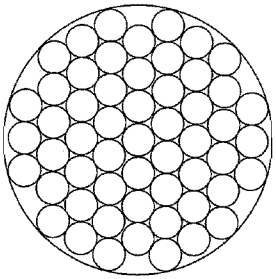

FIGURE 9.14 The sphere-packing problem.

We may thus visualize the picture portrayed in Figure 9.14. With everything inside a small sphere of radius $\sqrt{n\sigma^2}$ assigned to the code word on which it is centered, it is reasonable to say that when this particular code word is transmitted, the probability that the received vector will lie inside the correct "decoding" sphere is high. The key question is: How many decoding spheres can be packed inside the larger sphere of received vectors? In other words, how many code words can we in fact choose? To answer this question, we first recognize that the volume of an n-dimensional sphere of radius r may be written as $A_n r^n$, where A_n is a scaling factor. We may therefore make the following statements:

- The volume of the sphere of received vectors is $A_n[n(P + \sigma^2)]^{n/2}$.
- The volume of the decoding sphere is $A_n(n\sigma^2)^{n/2}$.

Accordingly, it follows that the maximum number of *nonintersecting* decoding spheres that can be packed inside the sphere of possible received vectors is

$$\frac{A_n[n(P + \sigma^2)]^{n/2}}{A_n(n\sigma^2)^{n/2}} = \left(1 + \frac{P}{\sigma^2}\right)^{n/2} = 2^{(n/2)\log_2(1+P/\sigma^2)} \tag{9.96}$$

Taking the logarithm of this result to base 2, we readily see that the maximum number of bits per transmission for a low probability of error is indeed as defined previously in Equation (9.94).

▶ EXAMPLE 9.9 Reconfiguration of Constellation for Reduced Power

To illustrate the idea of sphere packing, consider the 64-QAM square constellation of Figure 9.15*a*. The figure depicts two-dimensional nonintersecting decoding spheres centered on the message points in the constellation. In trying to pack the decoding spheres as tightly as possible while maintaining the same Euclidean distance between the message points as before, we obtain the alternative constellation shown in Figure 9.15*b*. With a common Euclidean distance between the message points, the two constellations of Figure 9.15 produce approximately the same bit error rate, assuming the use of a high enough signal-to-noise ratio over an AWGN channel; see, for example, Equation (5.95). However, comparing these two constellations, we find that the sum of squared Euclidean distances from the message points to the origin in Figure 9.15*b* is smaller than that in Figure 9.15*a*. It follows therefore that the tightly packed constellation of Figure 9.15*b* has an advantage over the square constellation

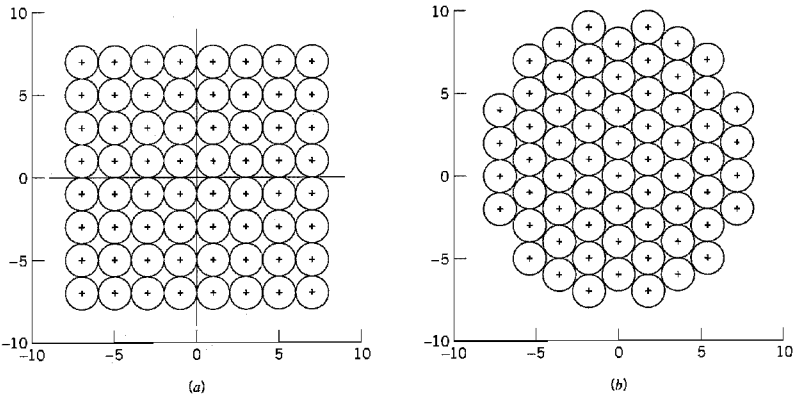

FIGURE 9.15 (*a*) Square 64-QAM constellation. (*b*) The most tightly coupled alternative to that of part *a*.

of Figure 9.15*a*: a smaller transmitted average signal energy per symbol for the same bit error rate on an AWGN channel. ◀

9.11 *Implications of the Information Capacity Theorem*

Now that we have an intuitive feel for the information capacity theorem, we may go on to discuss its implications in the context of a Gaussian channel that is limited in both power and bandwidth. For the discussion to be useful, however, we need an ideal framework against which the performance of a practical communication system can be assessed. To this end, we introduce the notion of an *ideal system* defined as one that transmits data at a bit rate R_b equal to the information capacity C. We may then express the average transmitted power as

$$P = E_b C \tag{9.97}$$

where E_b is the transmitted energy per bit. Accordingly, the ideal system is defined by the equation

$$\frac{C}{B} = \log_2\left(1 + \frac{E_b}{N_0}\frac{C}{B}\right) \tag{9.98}$$

Equivalently, we may define the *signal energy-per-bit to noise power spectral density ratio* E_b/N_0 in terms of the ratio C/B for the ideal system as

$$\frac{E_b}{N_0} = \frac{2^{C/B} - 1}{C/B} \tag{9.99}$$

A plot of bandwidth efficiency R_b/B versus E_b/N_0 is called the *bandwidth-efficiency diagram*. A generic form of this diagram is displayed in Figure 9.16, where the curve labeled

"capacity boundary" corresponds to the ideal system for which $R_b = C$. Based on Figure 9.16, we can make the following observations:

1. For *infinite bandwidth*, the ratio E_b/N_0 approaches the limiting value

$$\left(\frac{E_b}{N_0}\right)_\infty = \lim_{B\to\infty}\left(\frac{E_b}{N_0}\right) = \log 2 = 0.693 \tag{9.100}$$

This value is called the *Shannon limit* for an AWGN channel, assuming a code rate of zero. Expressed in decibels, it equals −1.6 dB. The corresponding limiting value of the channel capacity is obtained by letting the channel bandwidth B in Equation (9.95) approach infinity; we thus find that

$$C_\infty = \lim_{B\to\infty} C = \frac{P}{N_0}\log_2 e \tag{9.101}$$

where e is the base of the natural logarithm.

2. The *capacity boundary*, defined by the curve for the critical bit rate $R_b = C$, separates combinations of system parameters that have the potential for supporting error-free transmission ($R_b < C$) from those for which error-free transmission is not possible ($R_b > C$). The latter region is shown shaded in Figure 9.16.
3. The diagram highlights potential *trade-offs* among E_b/N_0, R_b/B, and probability of symbol error P_e. In particular, we may view movement of the operating point along

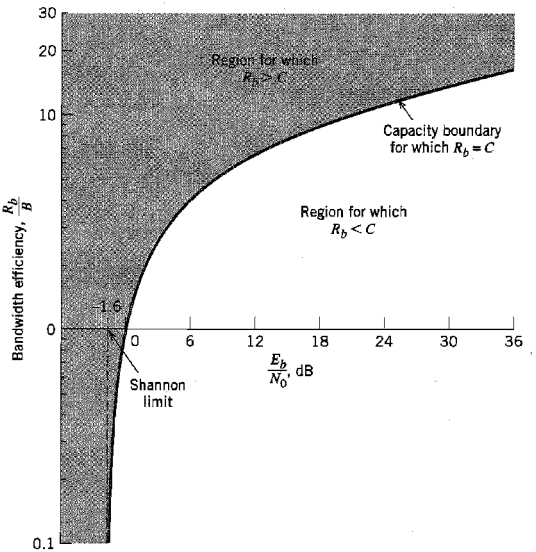

FIGURE 9.16 Bandwidth-efficiency diagram.

a horizontal line as trading P_e versus E_b/N_0 for a fixed R_b/B. On the other hand, we may view movement of the operating point along a vertical line as trading P_e versus R_b/B for a fixed E_b/N_0.

EXAMPLE 9.10 *M*-ary PCM

In this example, we look at an M-ary PCM system in light of the channel capacity theorem under the assumption that the system operates above the error threshold. That is, the average probability of error due to channel noise is negligible.

We assume that the M-ary PCM system uses a code word consisting of n code elements, each having one of M possible discrete amplitude levels; hence the name "M-ary." From Chapter 3 we recall that for a PCM system to operate above the error threshold, there must be provision for a noise margin that is sufficiently large to maintain a negligible error rate due to channel noise. This, in turn, means there must be a certain separation between these M discrete amplitude levels. Call this separation $k\sigma$, where k is a constant and $\sigma^2 = N_0B$ is the noise variance measured in a channel bandwidth B. The number of amplitude levels M is usually an integer power of 2. The average transmitted power will be least if the amplitude range is symmetrical about zero. Then the discrete amplitude levels, normalized with respect to the separation $k\sigma$, will have the values $\pm 1/2, +3/2, \ldots, \pm (M-1)/2$. We assume that these M different amplitude levels are equally likely. Accordingly, we find that the average transmitted power is given by

$$\begin{aligned} P &= \frac{2}{M}\left[\left(\frac{1}{2}\right)^2 + \left(\frac{3}{2}\right)^2 + \cdots + \left(\frac{M-1}{2}\right)^2\right](k\sigma)^2 \\ &= k^2\sigma^2\left(\frac{M^2-1}{12}\right) \end{aligned} \tag{9.102}$$

Suppose that the M-ary PCM system described herein is used to transmit a message signal with its highest frequency component equal to W hertz. The signal is sampled at the Nyquist rate of $2W$ samples per second. We assume that the system uses a quantizer of the midrise type, with L equally likely representation levels. Hence, the probability of occurrence of any one of the L representation levels is $1/L$. Correspondingly, the amount of information carried by a single sample of the signal is $\log_2 L$ bits. With a maximum sampling rate of $2W$ samples per second, the maximum rate of information transmission of the PCM system, measured in bits per second, is given by

$$R_b = 2W \log_2 L \text{ bits per second} \tag{9.103}$$

Since the PCM system uses a code word consisting of n code elements, each having one of M possible discrete amplitude values, we have M^n different possible code words. For a unique encoding process, we require

$$L = M^n \tag{9.104}$$

Clearly, the rate of information transmission in the system is unaffected by the use of an encoding process. We may therefore eliminate L between Equations (9.103) and (9.104) to obtain

$$R_b = 2Wn \log_2 M \text{ bits per second} \tag{9.105}$$

Equation (9.102) defines the average transmitted power required to maintain an M-ary PCM system operating above the error threshold. Hence, solving this equation for the number of discrete amplitude levels, M, we get

$$M = \left(1 + \frac{12P}{k^2N_0B}\right)^{1/2} \tag{9.106}$$

where $\sigma^2 = N_0B$ is the variance of the channel noise measured in a bandwidth B. Therefore, substituting Equation (9.106) into Equation (9.105), we obtain

$$R_b = Wn \log_2\left(1 + \frac{12P}{k^2N_0B}\right) \tag{9.107}$$

The channel bandwidth B required to transmit a rectangular pulse of duration $1/2nW$ (representing a code element in the code word) is given by (see Chapter 3)

$$B = \kappa nW$$

where κ is a constant with a value lying between 1 and 2. Using the minimum possible value $\kappa = 1$, we find that the channel bandwidth $B = nW$. We may thus rewrite Equation (9.107) as

$$R_b = B \log_2\left(1 + \frac{12P}{k^2N_0B}\right) \tag{9.108}$$

The *ideal system* is described by Shannon's channel capacity theorem, given in Equation (9.95). Hence, comparing Equation (9.108) with Equation (9.95), we see that they are identical if the average transmitted power in the PCM system is increased by the factor $k^2/12$, compared with the ideal system. Perhaps the most interesting point to note about Equation (9.108) is that the form of the equation is right: *Power and bandwidth in a PCM system are exchanged on a logarithmic basis, and the information capacity C is proportional to the channel bandwidth B.* ◀

▶ EXAMPLE 9.11 *M*-ary PSK and *M*-ary FSK

In this example, we compare the bandwidth-power exchange capabilities of M-ary PSK and M-ary FSK signals in light of Shannon's information capacity theorem. Consider first a coherent M-ary PSK system that employs a *nonorthogonal* set of M phase-shifted signals for the transmission of binary data. Each signal in the set represents a symbol with $\log_2 M$ bits. Using the definition of null-to-null bandwidth, we may express the bandwidth efficiency of M-ary PSK as follows [see Equation (6.51)]:

$$\frac{R_b}{B} = \frac{\log_2 M}{2}$$

In Figure 9.17*a*, we show the operating points for different numbers of phase levels $M = 2$, 4, 8, 16, 32, 64. Each point corresponds to an average probability of symbol error $P_e = 10^{-5}$. In the figure we have also included the capacity boundary for the ideal system. We observe from Figure 9.17*a* that as M is increased, the bandwidth efficiency is improved, but the value of E_b/N_0 required for error-free transmission moves away from the Shannon limit.

Consider next a coherent M-ary FSK system that uses an *orthogonal* set of M frequency-shifted signals for the transmission of binary data, with the separation between adjacent signal frequencies set at $1/2T$, where T is the symbol period. As with the M-ary PSK, each signal in the set represents a symbol with $\log_2 M$ bits. The bandwidth efficiency of M-ary FSK is as follows [see Equation (6.143)]:

$$\frac{R_b}{B} = \frac{2 \log_2 M}{M}$$

In Figure 9.17*b*, we show the operating points for different numbers of frequency levels $M = 2$, 4, 8, 16, 32, 64 for an average probability of symbol error $P_e = 10^{-5}$. In the figure, we have also included the capacity boundary for the ideal system. We see that increasing M in (orthogonal) M-ary FSK has the opposite effect to that in (nonorthogonal) M-ary PSK. In particular, as M is increased, which is equivalent to increased bandwidth requirement, the operating point moves closer to the Shannon limit. ◀

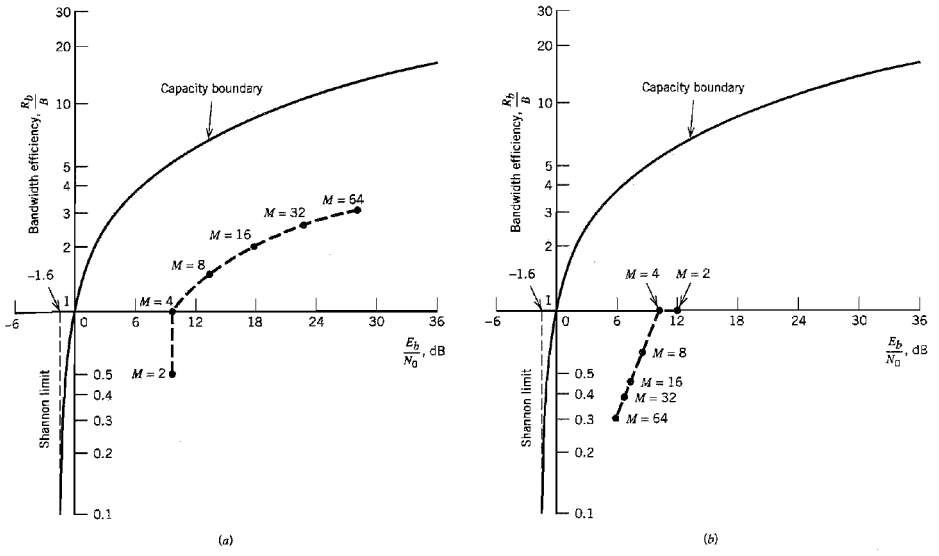

FIGURE 9.17 (*a*) Comparison of *M*-ary PSK against the ideal system for $P_e = 10^{-5}$ and increasing *M*. (*b*) Comparison of *M*-ary FSK against the ideal system for $P_e = 10^{-5}$ and increasing *M*.

EXAMPLE 9.12 Capacity of Binary-Input AWGN Channel

In this example, we investigate the capacity of an AWGN channel using *encoded* binary antipodal signaling (i.e., levels -1 and $+1$ for binary symbols 0 and 1, respectively). In particular, we address the issue of determining the minimum achievable bit error rate as a function of E_b/N_0 for varying code rate r. It is assumed that the binary symbols 0 and 1 are equiprobable.

Let the random variables X and Y denote the channel input and channel output, respectively; X is a discrete variable, whereas Y is a continuous variable. In light of the second line of Equation (9.81), we may express the mutual information between the channel input and channel output as

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

The second term, $h(Y|X)$, is the conditional differential entropy of the channel output Y, given the channel input X. By virtue of Equations (9.89) and (9.93), this term is just the entropy of a Gaussian distribution. Hence, using σ^2 to denote the variance of the channel noise, we may write

$$h(Y|X) = \frac{1}{2}\log_2(2\pi e\sigma^2)$$

Next, the first term, $h(Y)$, is the differential entropy of the channel output Y. With the use of binary antipodal signaling, the probability density function of Y, given $X = x$, is a mixture of two Gaussian distributions with common variance σ^2 and mean values -1 and $+1$, as shown by

$$f_Y(y_i|x) = \frac{1}{2}\left[\frac{\exp(-(y_i + 1)^2/2\sigma^2)}{\sqrt{2\pi}\sigma} + \frac{\exp(-(y_i - 1)^2/2\sigma^2)}{\sqrt{2\pi}\sigma}\right] \tag{9.109}$$

Hence, we may determine the differential entropy of Y using the formula

$$h(Y) = -\int_{-\infty}^{\infty} f_Y(y_i|x) \log_2[f_Y(y_i|x)]\, dy_i$$

where $f_Y(y_i|x)$ is defined by Equation (9.109). From the formulas of $h(Y|X)$ and $h(Y)$, it is clear that the mutual information is solely a function of the noise variance σ^2. Using $M(\sigma^2)$ to denote this functional dependence, we may thus write

$$I(X; Y) = M(\sigma^2)$$

Unfortunately, there is no closed formula that we can derive for $M(\sigma^2)$ because of the difficulty of determining $h(Y)$. Nevertheless, the differential entropy $h(Y)$ can be well approximated using *Monte Carlo integration*, which is straightforward to program on a digital computer; see Problem 9.36.

Because symbols 0 and 1 are equiprobable, it follows that the channel capacity C is equal to the mutual information between X and Y. Hence, for error-free data transmission over the AWGN channel, the code rate r must satisfy the condition

$$r < M(\sigma^2) \tag{9.110}$$

A robust measure of the ratio E_b/N_0 is

$$\frac{E_b}{N_0} = \frac{P}{N_0 r} = \frac{P}{2\sigma^2 r}$$

where P is the average transmitted power, and $N_0/2$ is the two-sided power spectral density of the channel noise. Without loss of generality, we may set $P = 1$. We may then express the noise variance as

$$\sigma^2 = \frac{N_0}{2E_b r} \tag{9.111}$$

Substituting Equation (9.111) into (9.110) and rearranging terms, we get the desired relation:

$$\frac{E_b}{N_0} = \frac{1}{2rM^{-1}(r)} \tag{9.112}$$

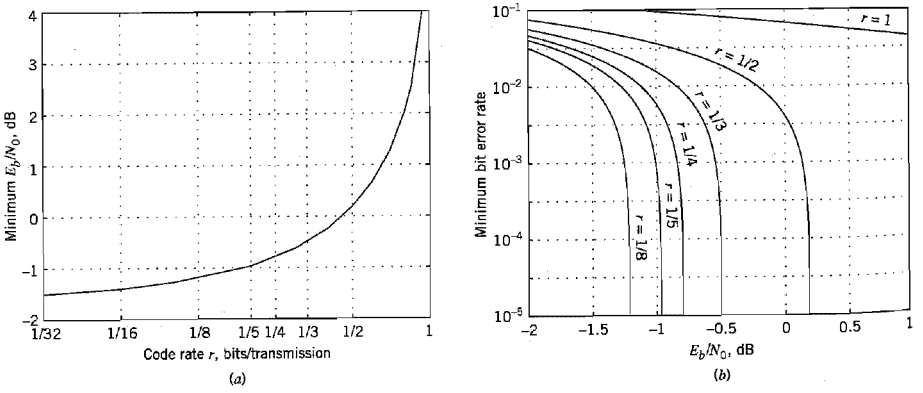

FIGURE 9.18 Binary antipodal signaling over an AWGN channel. (*a*) Minimum E_b/N_0 versus the code rate r. (*b*) Minimum bit error rate (BER) versus E_b/N_0 for varying code rate r.

where $M^{-1}(r)$ is the *inverse* of the mutual information between the channel input and output, expressed as a function of the code rate r.

Using the Monte Carlo method to estimate the differential entropy $h(Y)$ and therefore $M^{-1}(r)$, the plots of Figure 9.18 are computed.[12] Figure 9.18*a* plots the minimum E_b/N_0 versus the code rate r for error-free communication. Figure 9.18*b* plots the minimum achievable bit error rate versus E_b/N_0 with the code rate r as a running parameter. From Figure 9.18 we may draw the following conclusions:

- For uncoded binary signaling (i.e., $r = 1$), an infinite E_b/N_0 is required for error-free communication, which agrees with what we know about uncoded data transmission over an AWGN channel.
- The minimum E_b/N_0 decreases with decreasing code rate r, which is intuitively satisfying. For example, for $r = 1/2$, the minimum value of E_b/N_0 is slightly less than 0.2 dB.
- As r approaches zero, the minimum E_b/N_0 approaches the limiting value of -1.6 dB, which agrees with the Shannon limit derived earlier; see Equation (9.100). ◀

9.12 *Information Capacity of Colored Noise Channel*[13]

The information capacity theorem as formulated in Equation (9.95) applies to a band-limited white noise channel. In this section, we extend Shannon's information capacity theorem to the more general case of a *nonwhite*, or *colored*, *noise channel*. To be specific, consider the channel model shown in Figure 9.19*a* where the transfer function of the channel is denoted by $H(f)$. The channel noise $n(t)$, which appears additively at the channel output, is modeled as the sample function of a stationary Gaussian process of zero mean and power spectral density $S_N(f)$. The requirement is twofold:

1. Find the input ensemble, described by the power spectral density $S_X(f)$, that maximizes the mutual information between the channel output $y(t)$ and the channel input $x(t)$, subject to the constraint that the average power of $x(t)$ is fixed at a constant value P.
2. Hence, determine the optimum information capacity of the channel.

This problem is a constrained optimization problem. To solve it, we proceed as follows:

- Because the channel is linear, we may replace the model of Figure 9.19*a* with the equivalent model shown in Figure 9.19*b*. From the viewpoint of the spectral characteristics of the signal plus noise measured at the channel output, the two models of Figure 9.19 are equivalent, provided that the power spectral density of the noise

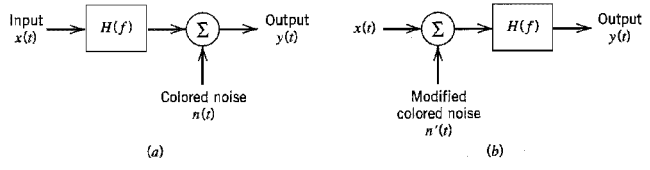

FIGURE 9.19 (*a*) Model of band-limited, power-limited noisy channel. (*b*) Equivalent model of the channel.

$n'(t)$ in Figure 9.19*b* is defined in terms of the power spectral density of the noise $n(t)$ in Figure 9.19*a* as

$$S_{N'}(f) = \frac{S_N(f)}{|H(f)|^2} \tag{9.113}$$

where $|H(f)|$ is the magnitude response of the channel.

- To simplify the analysis, we use the "principle of divide and conquer" in a manner similar to that described in Section 6.12. Specifically, the channel is divided into a large number of adjoining frequency slots, as illustrated in Figure 9.20. The smaller we make the incremental frequency interval Δf of each subchannel, the better is this approximation.

The net result of these two points is that the original model of Figure 9.19*a* is replaced by the parallel combination of a finite number of subchannels, N, each of which is corrupted essentially by "band-limited white Gaussian noise."

The kth subchannel in the approximation to the model of Figure 9.19*b* is described by

$$y_k(t) = x_k(t) + n_k(t), \qquad k = 1, 2, \ldots, N \tag{9.14}$$

The average power of the signal component $x_k(t)$ is

$$P_k = S_X(f_k)\,\Delta f, \qquad k = 1, 2, \ldots, N \tag{9.115}$$

where $S_X(f_k)$ is the power spectral density of the input signal evaluated at the frequency $f = f_k$. The variance of the noise component $n_k(t)$ is

$$\sigma_k^2 = \frac{S_N(f_k)}{|H(f_k)|^2}\,\Delta f, \qquad k = 1, 2, \ldots, N \tag{9.116}$$

where $S_N(f_k)$ and $|H(f_k)|$ are the noise spectral density and the channel's magnitude response evaluated at the frequency f_k, respectively. The information capacity of the kth subchannel is

$$C_k = \frac{1}{2}\,\Delta f \log_2\left(1 + \frac{P_k}{\sigma_k^2}\right), \qquad k = 1, 2, \ldots, N \tag{9.117}$$

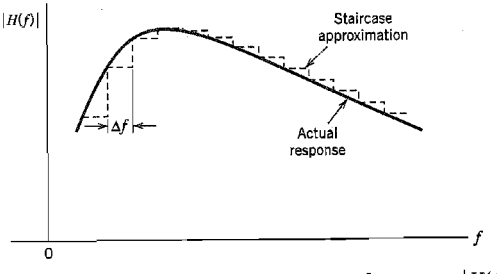

FIGURE 9.20 Staircase approximation of an arbitrary magnitude response $|H(f)|$; only positive-frequency portion of the response is shown.

where the factor 1/2 accounts for the fact that Δf applies to both positive and negative frequencies. All the N subchannels are independent of one another. Hence the total capacity of the overall channel is approximately given by the summation

$$\begin{aligned} C &\simeq \sum_{k=1}^{N} C_k \\ &= \frac{1}{2} \sum_{k=1}^{N} \Delta f \log_2\left(1 + \frac{P_k}{\sigma_k^2}\right) \end{aligned} \tag{9.118}$$

The problem we have to address is to maximize the overall information capacity C subject to the constraint:

$$\sum_{k=1}^{N} P_k = P = \text{constant} \tag{9.119}$$

The usual procedure to solve a constrained optimization problem is to use the *method of Lagrange multipliers*; see Note 19 in Chapter 6. To proceed with this optimization, we first define an objective function that incorporates both the information capacity C and the constraint [i.e., Equations (9.118) and (9.119)], as shown by

$$J = \frac{1}{2} \sum_{k=1}^{N} \Delta f \log_2\left(1 + \frac{P_k}{\sigma_k^2}\right) + \lambda\left(P - \sum_{k=1}^{N} P_k\right) \tag{9.120}$$

where λ is the Lagrange multiplier. Next, differentiating the objective function J with respect to P_k and setting the result equal to zero, we obtain

$$\frac{\Delta f \log_2 e}{P_k + \sigma_k^2} - \lambda = 0$$

To satisfy this optimizing solution, we impose the following requirement:

$$P_k + \sigma_k^2 = K \Delta f \qquad \text{for } k = 1, 2, \ldots, N \tag{9.121}$$

where K is a constant that is the same for all k. The constant K is chosen to satisfy the average power constraint.

Inserting the defining values of Equations (9.115) and (9.116) in the optimizing condition of Equation (9.121), simplifying, and rearranging terms, we get

$$S_X(f_k) = K - \frac{S_N(f_k)}{|H(f_k)|^2}, \qquad k = 1, 2, \ldots, N \tag{9.122}$$

Let $\mathcal{F}_A$ denote the frequency range for which the constant K satisfies the condition

$$K \geq \frac{S_N(f)}{|H(f)|^2}$$

Then, as the incremental frequency interval Δf is allowed to approach zero and the number of subchannels N goes to infinity, we may use Equation (9.122) to formally state that the power spectral density of the input ensemble that achieves the optimum information capacity is a nonnegative quantity defined by

$$S_X(f) = \begin{cases} K - \dfrac{S_N(f)}{|H(f)|^2} & \text{for } f \in \mathcal{F}_A \\ 0 & \text{otherwise} \end{cases} \tag{9.123}$$

Since the average power of a random process is the total area under the curve of the power spectral density of the process, we may express the average power of the channel input $x(t)$ as

$$P = \int_{f \in \mathcal{F}_A} \left(K - \frac{S_N(f)}{|H(f)|^2} \right) df \tag{9.124}$$

For a prescribed P and specified $S_N(f)$ and $H(f)$, the constant K is the solution to Equation (9.124).

The only thing that remains for us to do is to find the optimum information capacity. Substituting the optimizing solution of Equation (9.121) into Equation (9.118) and then using the defining values of Equations (9.115) and (9.116), we obtain

$$C \simeq \frac{1}{2} \sum_{k=1}^{N} \Delta f \log_2 \left(K \frac{|H(f_k)|^2}{S_N(f_k)} \right)$$

When the incremental frequency interval Δf is allowed to approach zero, this equation takes the limiting form:

$$C = \frac{1}{2} \int_{-\infty}^{\infty} \log_2 \left(K \frac{|H(f)|^2}{S_N(f)} \right) df \tag{9.125}$$

where the constant K is chosen as the solution to Equation (9.124) for a prescribed input signal power P.

■ WATER-FILLING INTERPRETATION OF THE INFORMATION CAPACITY THEOREM

Equations (9.123) and (9.124) suggest the picture portrayed in Figure 9.21. Specifically, we make the following observations:

- The appropriate input power spectral density $S_X(f)$ is described as the bottom regions of the function $S_N(f)/|H(f)|^2$ that lie below the constant level K, which are shown shaded.
- The input power P is defined by the total area of these shaded regions.

The spectral domain picture portrayed here is called the *water-filling (pouring) interpretation* in the sense that the process by which the input power is distributed across

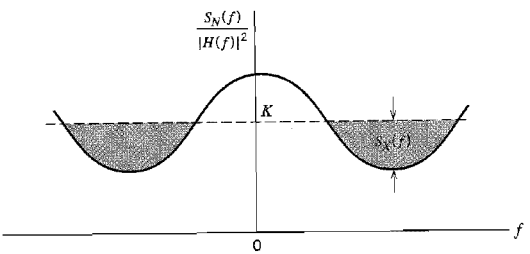

FIGURE 9.21 Water-filling interpretation of information-capacity theorem for a colored noisy channel.

the function $S_N(f)/|H(f)|^2$ is identical to the way in which water distributes itself in a vessel.

Consider now the idealized case of a band-limited signal in additive white Gaussian noise of power spectral density $N(f) = N_0/2$. The transfer function $H(f)$ is that of an ideal band-pass filter defined by

$$H(f) = \begin{cases} 1, & 0 \le f_c - \dfrac{B}{2} \le |f| \le f_c + \dfrac{B}{2} \\ 0, & \text{otherwise} \end{cases}$$

where f_c is the midband frequency and B is the channel bandwidth. For this special case, Equations (9.124) and (9.125) reduce to, respectively,

$$P = 2B\left(K - \frac{N_0}{2}\right)$$

and

$$C = B \log_2\left(\frac{2K}{N_0}\right)$$

Hence, eliminating K between these two equations, we get the standard form of Shannon's capacity theorem, defined by Equation (9.95).

▶ EXAMPLE 9.13 Capacity of NEXT-Dominated Channel

From the discussion presented in Section 4.8, we recall that a major channel impairment in digital subscriber lines is near-end crosstalk (NEXT). The power spectral density of this crosstalk may be taken as

$$S_N(f) = |H_{\text{NEXT}}(f)|^2 S_X(f) \tag{9.126}$$

where $S_X(f)$ is the power spectral density of the transmitted signal and $H_{\text{NEXT}}(f)$ is the transfer function that couples adjacent twisted pairs. The only constraint we have to satisfy in this example is that the power spectral density function $S_X(f)$ be *nonnegative for all* f. Substituting Equation (9.126) into (9.123), we readily find that this condition is satisfied by solving for K as

$$K = \left(1 + \frac{|H_{\text{NEXT}}(f)|^2}{|H(f)|^2}\right) S_X(f)$$

Finally, using this result in Equation (9.125), we find that the capacity of the NEXT-dominated digital subscriber channel is given by

$$C = \frac{1}{2} \int_{\mathcal{F}_A} \log_2\left(1 + \frac{|H(f)|^2}{|H_{\text{NEXT}}(f)|^2}\right) df$$

where $\mathcal{F}_A$ is the set of positive and negative frequencies for which $S_X(f) > 0$.

9.13 *Rate Distortion Theory*

In Section 9.3 we introduced the source coding theorem for a discrete memoryless source, according to which the average code-word length must be at least as large as the source entropy for perfect coding (i.e., perfect representation of the source). However, in many practical situations there are constraints that force the coding to be imperfect, thereby

resulting in unavoidable *distortion*. For example, constraints imposed by a communication channel may place an upper limit on the permissible code rate and therefore average code-word length assigned to the information source. As another example, the information source may have a continuous amplitude as in the case of speech, and the requirement is to quantize the amplitude of each sample generated by the source to permit its representation by a code word of finite length as in pulse-code modulation. In such cases, the problem is referred to as *source coding with a fidelity criterion*, and the branch of information theory that deals with it is called *rate distortion theory*.[14] Rate distortion theory finds applications in two types of situations:

- Source coding where the permitted coding alphabet cannot exactly represent the information source, in which case we are forced to do lossy *data compression*.
- Information transmission at a rate greater than channel capacity.

Accordingly, rate distortion theory may be viewed as a natural extension of Shannon's coding theorems.

■ RATE DISTORTION FUNCTION

Consider a discrete memoryless source defined by an M-ary alphabet X: $\{x_i | i = 1, 2, \ldots, M\}$, which consists of a set of statistically independent symbols together with the associated symbol probabilities $\{p_i | i = 1, 2, \ldots, M\}$. Let R be the average code rate in bits per code word. The representation code words are taken from another alphabet Y: $\{y_j | j = 1, 2, \ldots, N\}$. The source coding theorem states that this second alphabet provides a perfect representation of the source provided that $R > H$, where H is the source entropy. But if we are forced to have $R < H$, then there is unavoidable distortion and therefore loss of information.

Let $p(x_i, y_j)$ denote the joint probability of occurrence of source symbol x_i and representation symbol y_j. From probability theory, we have

$$p(x_i, y_j) = p(y_j | x_i) p(x_i) \tag{9.127}$$

where $p(y_j | x_i)$ is a transition probability. Let $d(x_i, y_j)$ denote a measure of the cost incurred in representing the source symbol x_i by the symbol y_j; the quantity $d(x_i, y_j)$ is referred to as a *single-letter distortion measure*. The statistical average of $d(x_i, y_j)$ over all possible source symbols and representation symbols is given by

$$\bar{d} = \sum_{i=1}^{M} \sum_{j=1}^{N} p(x_i) p(y_j | x_i)\, d(x_i, y_j) \tag{9.128}$$

Note that the average distortion $\bar{d}$ is a nonnegative continuous function of the transition probabilities $p(y_j | x_i)$ that are determined by the source encoder-decoder pair.

A conditional probability assignment $p(y_j | x_i)$ is said to be *D-admissible* if and only if the average distortion $\bar{d}$ is less than or equal to some acceptable value D. The set of all D-admissible conditional probability assignments is denoted by

$$P_D = \{p(y_j | x_i) : \bar{d} \leq D\} \tag{9.129}$$

For each set of transition probabilities, we have a mutual information

$$I(X; Y) = \sum_{i=1}^{M} \sum_{j=1}^{N} p(x_i) p(y_j | x_i) \log\left(\frac{p(y_j | x_i)}{p(y_j)}\right) \tag{9.130}$$

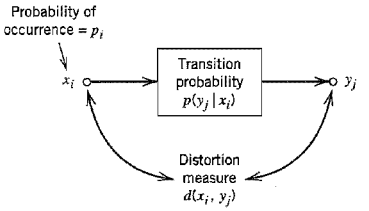

FIGURE 9.22 Summary of rate distortion theory.

A *rate distortion function* $R(D)$ is defined as *the smallest coding rate possible for which the average distortion is guaranteed not to exceed* D. Let P_D denote the set to which the conditional probability $p(y_j|x_i)$ belongs for a prescribed D. Then, for a fixed D we write[15]

$$R(D) = \min_{p(y_j|x_i) \in P_D} I(X; Y) \tag{9.131}$$

subject to the constraint

$$\sum_{j=1}^{N} p(y_j|x_i) = 1 \qquad \text{for } i = 1, 2, \ldots, M \tag{9.132}$$

The rate distortion function $R(D)$ is measured in units of bits if the base-2 logarithm is used in Equation (9.130). Intuitively, we expect the distortion D to decrease as the rate distortion function $R(D)$ is increased. We may say conversely that tolerating a large distortion D permits the use of a smaller rate for coding and/or transmission of information.

Figure 9.22 summarizes the main parameters of rate distortion theory. In particular, given the source symbols $\{x_i\}$ and their probabilities $\{p_i\}$ and given a definition of the single-letter distortion measure $d(x_i, y_j)$, the calculation of the rate distortion function $R(D)$ involves finding the conditional probability assignment $p(y_j|x_i)$ subject to certain constraints imposed on $p(y_j|x_i)$. This is a variational problem, the solution of which is unfortunately not straightforward in general.

EXAMPLE 9.14 Gaussian Source

Consider a discrete-time, memoryless Gaussian source with zero mean and variance σ^2. Let x denote the value of a sample generated by such a source. Let y denote a quantized version of x that permits a finite representation of it. The *squared error distortion*

$$d(x, y) = (x - y)^2$$

provides a distortion measure that is widely used for continuous alphabets. The rate distortion function for the Gaussian source with squared error distortion, as described herein, is given by

$$R(D) = \begin{cases} \frac{1}{2} \log\left(\frac{\sigma^2}{D}\right), & 0 \le D \le \sigma^2 \\ 0, & D > \sigma^2 \end{cases} \tag{9.133}$$

In this case, we see that $R(D) \to \infty$ as $D \to 0$, and $R(D) = 0$ for $D = \sigma^2$.

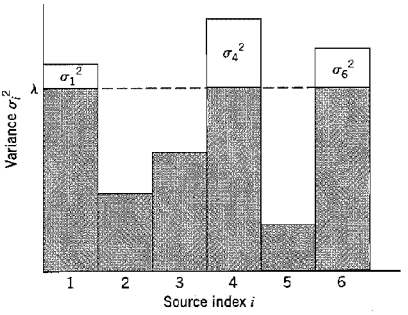

FIGURE 9.23 Reverse water-filling picture for a set of parallel Gaussian processes.

▶ EXAMPLE 9.15 Set of Parallel Gaussian Sources

Consider next a set of N independent Gaussian random variables $\{X_i\}_{i=1}^N$, where X_i has zero mean and variance σ_i^2. Using the distortion measure

$$d = \sum_{i=1}^{N} (x_i - \hat{x}_i)^2$$

and building on the result of Example 9.14, we may express the rate distortion function for the set of parallel Gaussian sources described here as

$$R(D) = \sum_{i=1}^{N} \frac{1}{2} \log\left(\frac{\sigma_i^2}{D_i}\right) \tag{9.134}$$

where D_i is itself defined by

$$D_i = \begin{cases} \lambda & \text{if } \lambda < \sigma_i \\ \sigma_i^2 & \text{if } \lambda \geq \sigma_i^2 \end{cases} \tag{9.135}$$

and the constant λ is chosen so as to satisfy the condition

$$\sum_{i=1}^{N} D_i = D \tag{9.136}$$

Equations (9.135) and (9.136) may be interpreted as a kind of "water-filling in reverse," as illustrated in Figure 9.23. First, we choose a constant λ and only the subset of random variables whose variances exceed the constant λ. No bits are used to describe the remaining subset of random variables whose variances are less than the constant λ. ◀

9.14 *Data Compression*

Rate distortion theory naturally leads us to consider the idea of *data compression* that involves a purposeful or unavoidable reduction in the information content of data from a continuous or discrete source. Specifically, we may think of a *data compressor*, or *signal compressor*, as a device that supplies a code with the least number of symbols for the representation of the source output, subject to a permissible or acceptable *distortion*. The data compressor thus retains the essential information content of the source output by blurring fine details in a deliberate but controlled manner. Accordingly, data compression

is a *lossy* operation in the sense that the source entropy is reduced (i.e., information is lost), irrespective of the type of source being considered.

In the case of a discrete source, the reason for using data compression is to encode the source output at a rate smaller than the source entropy. By so doing, the source coding theorem is violated, which means that exact reproduction of the original data is *no longer* possible.

In the case of a continuous source, the entropy is infinite, and therefore a signal compression code must always be used to encode the source output at a finite rate. Consequently, it is impossible to digitally encode an analog signal with a finite number of bits without producing some distortion. This statement is in perfect accord with the idea of pulse-code modulation, which was studied in Chapter 3. There it was shown that quantization, which is basic to the analog-to-digital conversion process in pulse-code modulation, always introduces distortion (known as quantization noise) into the transmitted signal. A quantizer may therefore be viewed as a signal compressor.

The uniform and nonuniform quantizers considered in Chapter 3 are said to be *scalar quantizers* in the sense that they deal with samples of the analog signal (i.e., continuous source output) one at a time. Each sample is converted into a quantized value, with the conversion being independent from sample to sample. A scalar quantizer is a rather simple signal compressor, which makes it attractive for practical use. Yet it can provide a surprisingly good performance; this is especially so if nonuniform quantization is used.

There is another class of quantizers known as *vector quantizers* that use blocks of consecutive samples of the source output to form vectors, each of which is treated as a single entity. The essential operation in a vector quantizer is the quantization of a random vector[16] by encoding it as a binary code word. The vector is encoded by comparing it with a *codebook* consisting of a set of stored reference vectors known as *code vectors* or *patterns*. Each pattern in the codebook is used to represent input vectors that are identified by the encoder to be similar to the particular pattern, subject to the maximization of an appropriate fidelity criterion. The encoding process in a vector quantizer may thus be viewed as a *pattern matching operation*.

Let N be the number of code vectors in the codebook, k be the dimension of each vector (i.e., the number of samples in each pattern), and r be the coded transmission rate in bits per sample. These three parameters are related as follows:

$$r = \frac{\log_2 N}{k} \tag{9.137}$$

Then, assuming that the size of the code book is sufficiently large, the signal-to-quantization noise ratio (SNR) for the vector quantizer is given by

$$10 \log_{10}(\text{SNR}) = 6\left(\frac{\log_2 N}{k}\right) + C_k \text{ dB} \tag{9.138}$$

where C_k is a constant (expressed in dB) that depends on the dimensions k. According to Equation (9.138), the SNR for a vector quantizer increases approximately at the rate of $6/k$ dB for each doubling of the codebook size. Equivalently, we may state that the SNR increases by 6 dB per unit increase in rate (bits per sample) as in the standard PCM using a uniform scalar quantizer. The advantage of the vector quantizer over the scalar quantizer is that its constant term C_k has a higher value, because the vector quantizer optimally exploits the correlations among the samples constituting a vector. Specifically, the constant C_k increases with the dimension k, approaching the ultimate rate-distortion limit for a

given source of information. However, the improvement in SNR is attained at the cost of increased encoding complexity, which grows exponentially with the dimension k for a specified rate r. Unfortunately, this is the main obstacle to the wide use of vector quantization in practice. Nevertheless, in certain applications, the issue of computational complexity is mitigated by exploiting the capability of VLSI technology to concentrate a highly complex signal processor on a silicon chip. For example, that is precisely what is done in the use of code-excited linear predictive (CELP) modeling of speech in wireless communication systems of the CDMA type, namely, the IS-95 system. From the description of CELP presented in Section 8.9, it is clear that the CELP modeling of speech is an example of vector quantization.

9.15 *Summary and Discussion*

In this chapter we established four fundamental limits on different aspects of a communication system. The limits are embodied in the source coding theorem, the channel coding theorem, the information capacity theorem, and the rate distortion function.

The *source coding theorem*, Shannon's first theorem, provides the mathematical tool for assessing *data compaction*, that is, *lossless compression* of data generated by a discrete memoryless source. The theorem tells us that we can make the average number of binary code elements (bits) per source symbol as small as, but no smaller than, the entropy of the source measured in bits. The *entropy* of a source is a function of the probabilities of the source symbols that constitute the alphabet of the source. Since entropy is a measure of uncertainty, the entropy is maximum when the associated probability distribution generates maximum uncertainty.

The *channel coding theorem*, Shannon's second theorem, is both the most surprising and the single most important result of information theory. For a *binary symmetric channel*, the channel coding theorem tells us that for any *code rate* r less than or equal to the *channel capacity* C, codes do exist such that the average probability of error is as small as we want it. A binary symmetric channel is the simplest form of a discrete memoryless channel. It is symmetric because the probability of receiving a 1 if a 0 is sent is the same as the probability of receiving a 0 if a 1 is sent. This probability, the probability that an error will occur, is termed a *transition probability*. The transition probability p is determined not only by the additive noise at the channel output but also by the kind of receiver used. The value of p uniquely defines the channel capacity C.

Shannon's third remarkable theorem, the *information capacity theorem*, tells us that there is a maximum to the rate at which any communication system can operate reliably (i.e., free of errors) when the system is constrained in power. This maximum rate is called the *information capacity*, measured in bits per second. When the system operates at a rate greater than the information capacity, it is condemned to a high probability of error, regardless of the choice of signal set used for transmission or the receiver used for processing the received signal.

Finally, the rate distortion function provides the mathematical tool for signal compression (i.e., solving the problem of source coding with a fidelity criterion): The rate distortion function can be applied to a discrete as well as continuous memoryless source.

When the output of a source of information is compressed in a lossless manner, the resulting data stream usually contains redundant bits. These redundant bits can be removed by using a lossless algorithm such as Huffman coding or the Lempel–Ziv algorithm for data compaction. We may thus speak of data compression followed by data compaction as two constituents of the *dissection of source coding*, which is so called because it refers

exclusively to the sources of information. In some source coding applications, we have a third constituent, namely, *data encryption*, which follows data compaction. The purpose of data encryption is to disguise the data (bit) stream in such a way that it has no meaning to an unauthorized receiver. Some basic aspects of *cryptography*, which encompasses both encryption and decryption, follow quite naturally from information theory, as discussed in Appendix 5. Other issues relating to cryptography are also discussed in that appendix.

One last comment is in order. Shannon's information theory, as presented in this chapter, has been entirely in the context of memoryless sources and channels. The theory can be extended to deal with sources and channels with *memory*, in which case a symbol of interest depends on preceding symbols; however, the level of exposition needed to do this is beyond the scope of this book.[17]

NOTES AND REFERENCES

1. According to Lucky (1989), the first mention of the term *information theory* by Shannon occurs in a 1945 memorandum entitled "A Mathematical Theory of Cryptography." It is rather curious that the term was never used in the classic 1948 paper by Shannon, which laid down the foundations of information theory. For an introductory treatment of information theory, see Chapter 2 of Lucky (1989) and the paper by Wyner (1981); see also the books of Adámek (1991), Hamming (1980), and Abramson (1963). For more advanced treatments of the subject, see the books of Cover and Thomas (1991), Blahut (1987), McEliece (1977), and Gallager (1968). For a collection of papers on the development of information theory (including the 1948 classic paper by Shannon), see Slepian (1974). For a collection of the papers published by Shannon, see Sloane and Wyner (1993).

2. The use of a logarithmic measure of information was first suggested by Hartley (1928); however, Hartley used logarithms to base 10.

3. In statistical physics, the entropy of a physical system is defined by (Reif, 1967, p. 147)

$$\mathcal{S} = k \log \Omega$$

where k is Boltzmann's constant, Ω is the number of states accessible to the system, and log denotes the natural logarithm. This entropy has the dimensions of energy because its definition involves the constant k. In particular, it provides a *quantitative measure of the degree of randomness of the system*. Comparing the entropy of statistical physics with that of information theory, we see that they have a similar form. For a detailed discussion of the relation between them, see Pierce (1961, pp. 184–207) and Brillouin (1962).

4. For the original proof of the source coding theorem, see Shannon (1948). A general proof of the source coding theorem is also given in the following books: Viterbi and Omura (1979, pp. 13–19), McEliece (1977, Chapter 3), and Gallager (1968, pp. 38–55). The source coding theorem is also referred to in the literature as the *noiseless coding theorem*, noiseless in the sense that it establishes the condition for error-free encoding to be possible.

5. For proof of the Kraft–McMillan inequality, see Cover and Thomas (1991, pp. 82–84), Blahut (1990, pp. 298–299), and McEliece (1977, pp. 239–240). For a proof of Equation (9.23), see Cover and Thomas (1991), pp. 87–88), Blahut (1990, pp. 300–301), and McEliece (1977, pp. 241–242).

6. The Huffman code is named after its inventor: D. A. Huffman (1952). For a readable account of Huffman coding and its use in data compaction, see Adámek (1991).

7. The original papers on the Lempel–Ziv algorithm are Ziv and Lempel (1977, 1978). For readable descriptions of the Lempel–Ziv algorithm, see Lucky (1989, pp. 118–122), Blahut

(1990, pp. 314–319), and Gitlin, Hayes, and Weinstein (1992, pp. 120–122). For the application of the Lempel–Ziv algorithm to the compaction of English text, see Lucky (1989, pp. 122–128) and the paper by Welch (1984); see also the review paper by Weiss and Shremp (1993).

8. The channel coding theorem is also known as the *noisy coding theorem*. The original proof of the theorem is given in Shannon (1948). A proof of the theorem is also presented in Hamming (1980, Chapters 9 and 10) in sufficient detail so that a general appreciation of relevant results is developed. The second part of the theorem is referred to in the literature as *the converse to the coding theorem*. A proof of this theorem is presented in the following references: Viterbi and Omura (1979, pp. 28–34) and Gallager (1968, pp. 76–82).

9. The quantity

$$\int_{-\infty}^{\infty} f_Y(x) \log_2\left(\frac{f_X(x)}{f_Y(x)}\right) dx$$

on the left-hand side of Equation (9.70) is called *relative entropy* or the *Kullback–Leibler divergence* between the probability density functions $f_X(x)$ and $f_Y(x)$; see Kullback (1968).

10. Shannon's information capacity theorem is also referred to in the literature as the *Shannon–Hartley law* in recognition of early work by Hartley on information transmission (Hartley, 1928). In particular, Hartley showed that the amount of information that can be transmitted over a given channel is proportional to the product of the channel bandwidth and the time of operation.

11. A lucid exposition of sphere packing is presented in Cover and Thomas (1991, pp. 242–243); see also Wozencraft and Jacobs (1965, pp. 323–341).

12. Parts *a* and *b* of Figure 9.18 follow the corresponding parts of Figure 6.2 in the book by Frey (1998).

13. For a rigorous treatment of the information capacity of a colored noisy channel, see Gallager (1968). The idea of replacing the channel model of Figure 9.19*a* with that of Figure 9.19*b* is discussed in Gitlin, Hayes, and Weinstein (1992).

14. For a complete treatment of rate distortion theory, see the book by Berger (1971); this subject is also treated in somewhat less detail in Cover and Thomas (1991), McEliece (1977), and Gallager (1968).

15. For the derivation of Equation (9.131), see Cover and Thomas (1991, p. 345). An algorithm for computation of the rate distortion function $R(D)$ defined in Equation (9.131) is described in Blahut (1987, pp. 220–221) and Cover and Thomas (1991, pp. 364–367).

16. For the early papers on vector quantization, see Gersho (1979) and Linde, Buzo, and Gray (1980). For a tutorial review of vector quantization, see Gray (1984). Equation (9.138), defining the SNR for a vector quantizer, is discussed in Gersho and Cuperman (1983). For a complete treatment of vector quantization, see the book by Gersho and Gray (1992).

17. For detailed discussion of discrete channels with memory, see Gallager (1968, pp. 97–112) and Ash (1965, pp. 211–229).

PROBLEMS

Entropy

9.1 Let p denote the probability of some event. Plot the amount of information gained by the occurrence of this event for $0 \leq p \leq 1$.