two signal hypotheses as concatenations in time of consecutive PSK signals:

$$\begin{aligned} s_0(t) &\longleftrightarrow \left(\frac{2E_s}{T_s}\right)^{1/2}\cos(\omega_c t+\theta),\ \left(\frac{2E_s}{T_s}\right)^{1/2}\cos(\omega_c t+\theta), \\ s_1(t) &\longleftrightarrow \left(\frac{2E_s}{T_s}\right)^{1/2}\cos(\omega_c t+\theta),\ -\left(\frac{2E_s}{T_s}\right)^{1/2}\cos(\omega_c t+\theta). \end{aligned} \tag{3.5.12}$$

(We reemphasize that a new bit is transmitted every T_s seconds, and the previous signal becomes the new reference signal.)

We have just argued that the DPSK receiver is in effect a noncoherent detector over two intervals. Since in the binary case the signal pairs in (3.5.12) are orthogonal, the bit error probability for binary DPSK can be evaluated using binary noncoherent orthogonal performance, except with the effective symbol energy $E_s = 2E_b$. Thus, using (3.4.26) and this conversion, we find that

$$\boxed{P_b = \frac{1}{2}e^{-2E_b/2N_0} = \frac{1}{2}e^{-E_b/N_0}, \qquad \text{binary DPSK, AWGN}} \tag{3.5.13}$$

When $M = 2$, DPSK has only slight loss in energy efficiency relative to coherent PSK. At $P_b = 10^{-5}$, DPSK requires about 10.4 dB E_b/N_0, whereas coherent PSK requires about 9.6 dB. Differentially encoded, coherently detected PSK requires about 9.9 dB.

Some propensity exists for paired or back-to-back symbol errors with DPSK, but it is not so strong as in differentially encoded coherent PSK, where an isolated error on one bit produces a paired error upon differential decoding. To see why the tendency is less, consider again the case of $M = 2$. A common error event is of the following form: the first phasor experiences a phase error of, say, $\beta_{n-1} = -50°$, while the second phase error is $\beta_n = +45°$. The phase difference δ_n thus exceeds 90°, inducing a decision error. However, if the next phase error is less than 45°, a subsequent symbol error is not made. This pattern is far more likely than "error near 0°, error of 100°, error near 0°," which would induce paired symbol errors. Thus, back-to-back decision errors are not as predominant as might be expected. Salz and Salzberg [28] and Oberst and Schilling [29] give an analysis of this double-error effect. In any case, the marginal P_b is correctly expressed in (3.5.13).

Returning to the M-ary DPSK case, the calculation of symbol error probability would first calculate the p.d.f. for the modulo 2π phase difference of two phasors corrupted by two-dimensional independent Gaussian noise. This p.d.f. is formulated in Pawula et al. [30]:

$$f(\delta) = \frac{1}{2\pi}\int_0^{\pi/2}(\sin x)\left[1+\frac{E_s}{N_0}(1+\cos\delta\sin x)\right]\cdot\exp\left[\frac{-2E_s}{N_0}(1-\cos\delta\sin x)\right]dx. \tag{3.5.14}$$

In Figure 3.5.4, we show the p.d.f. for the phase difference measurement when $E_s/N_0 = 10$ dB, given that zero phase difference occurred at the transmitter.

This p.d.f. can be integrated numerically over the region $|\delta| \geq \pi/M$ to produce P_s, and presentation of this analysis is found in Lindsey and Simon [12]. Figure 3.5.5 presents the results graphically, showing M-ary coherent detection for comparison. As

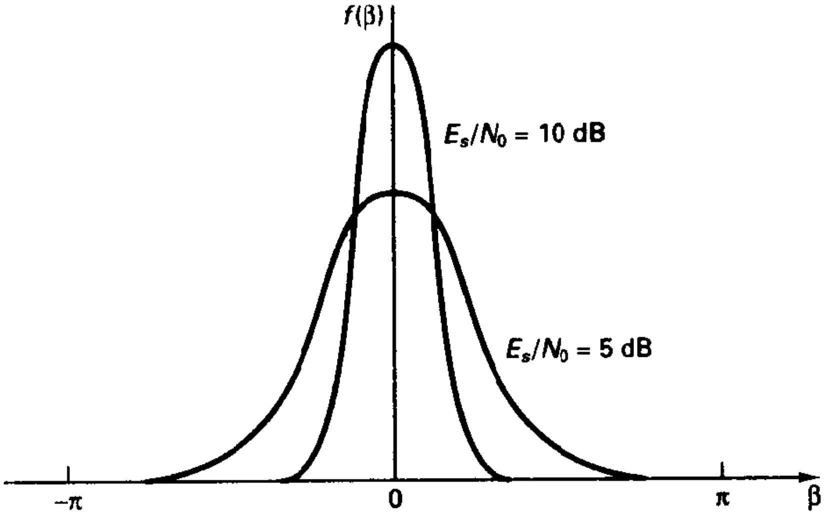

Figure 3.5.4 Phase difference p.d.f.

expected, the demodulator's lack of absolute phase knowledge always costs in energy efficiency; the difference is small for binary DPSK, but for larger M, say 16-DPSK, the penalty is nearly 3 dB. We might have anticipated this, because in DPSK detection two noise vectors influence the phase difference. At high signal-to-noise ratio, the phase error for each symbol is nearly Gaussian, and thus the phase difference δ is roughly Gaussian (see Figure 3.5.4 for example), but with twice the variance due to independence of the measurements. (This is pursued further in Exercise 3.5.2.)

An alternative method to calculating error probability uses the p.d.f.'s for the random variables in (3.5.7), following the earlier analysis of the noncoherent detector. The signals are not orthogonal, however, over $2T_s$ in the nonbinary case, and noncentral chi-squared statistics are encountered [31].

Because the energy efficiency of M-DPSK is quite poor for $M \geq 8$, especially relative to the coherent counterpart, these designs are rarely found in modern practice when energy efficiency is a primary concern. Binary DPSK, however, represents an effective alternative to binary PSK, with or without additional coding, and 4-DPSK was selected as a modulation technique for one of the first high-speed modems, the Bell model 201 2400-bps telephone channel modem, implemented in 1962. There, receiver simplicity was of paramount concern, as well as bandwidth economy, and channel SNR was nominally rather high. 4-ary DPSK, combined with channel coding, has been selected as the modulation method for next-generation digital cellular telephony in the United States.

If bit error probability is to be minimized, the phase *changes* should be Gray-coded, since the most likely phase difference error is to an adjacent region, and such cases should produce minimal bit errors. Under such conditions,

$$P_b \approx \frac{P_s}{\log_2 M}, \tag{3.5.15}$$

as for coherent PSK.

Although the use of DPSK avoids needing to know absolute carrier phase, it is important that the receiver be well synchronized in frequency. If it is not, the measured

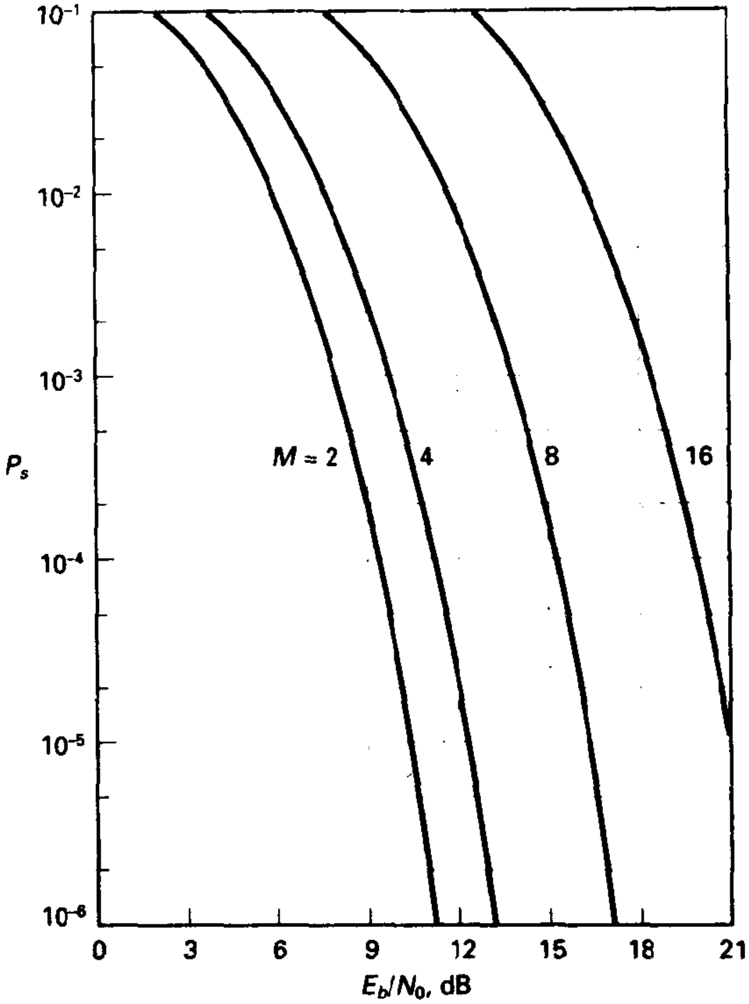

Figure 3.5.5 Symbol error probability for M-ary DPSK.

phase increments will be biased away from the middle of the decision zones, increasing the symbol error probability. A good rule of thumb for binary DPSK reception is to maintain $\Delta\omega T_b \leq 0.1$, where $\Delta\omega$ is the radian frequency offset [32]. This ensures that the carrier phase θ drifts by less than 0.1 radian during one bit interval. Generalizing this to the M-ary case, we might then require that $\Delta\omega T_s \leq 0.1/M$, since the decision regions shrink inversely with M. The frequency offset Δf must then be held to less than about $R_s/60M$, where R_s is the symbol rate.

Example 3.12 4-ary DPSK

Suppose in a 4-ary DPSK receiver that the following sequence of phase measurements (in degrees) is observed over five consecutive symbol intervals: 39, 110, 50, 239, 21. The phase difference sequence, modulo 2π, is 71°, 300°, 189°, 142°, and these are mapped to symbols 1, 3, 2, 2 according to (3.5.1). If $E_b/N_0 = 9$ dB on this channel, then the probability of a symbol error from Figure 3.5.5 is $3 \cdot 10^{-3}$.

For $R_s = 24$ ksps, approximately the rate for the IS-54 digital time-division cellular standard in North America, then by the above rule the required frequency accuracy must be less than about $\Delta f \leq 24{,}000/(60 \cdot 4) \approx 100$ Hz. This constitutes the allowable frequency offset for oscillator instability and Doppler shift combined.

It is possible to improve the performance of M-DPSK by forming decisions based on more than two consecutive symbols, which is called multisymbol detection of M-DPSK. Specifically, we can employ a sliding (or block) window of length $(N + 1)T_s$ to decide N consecutive data symbols, or perhaps just the oldest symbol of a sliding block. The qualitative notion is that a longer observation window allows effectively the establishment of a higher-quality phase reference for detection than that obtained from just the previous symbol. In some sense such detectors are acting as short-memory phase estimating schemes, and in the limit of large observation interval, the performance approaches that of coherent detection with differential detection. For $M > 2$, the potential gains in energy efficiency are significant, and it has been shown that use of $N = 3$ provides at least half the available gain. However, the receiver processing, if optimal noncoherent detection is pursued, is considerably larger, for M^N hypotheses need to be examined for a window of length $(N+1)T_s$ seconds. Furthermore, the constant phase assumption about the channel becomes more questionable. The interested reader is referred to [33] and [34] for a discussion of these possibilities.

3.6 PERFORMANCE ON THE SLOW, NONSELECTIVE RAYLEIGH FADING CHANNEL

We now study the performance of the previous modulation and detection strategies on the slow, flat-fading Rayleigh channel and will observe a fundamentally different dependence on signal-to-noise ratio than seen thus far for the nonfading, Gaussian noise channel. Specifically, instead of a negative exponential dependence on E_b/N_0 common to all cases in Sections 3.3, 3.4, and 3.5, we shall find that the infrequent, very deep amplitude fading events induce a much weaker (inverse) dependence of P_s on average E_b/N_0. This will be true for all uncoded transmission strategies, and the potential performance penalties due to fading are enormous for high-reliability systems. However, various channel coding techniques studied in later chapters will be able to substantially mitigate the effect of fading.

To recall the model assumptions made at the beginning of the chapter, we assume the channel gain $A(t)$ is a Rayleigh random process, but essentially fixed over the duration of one symbol's decision interval. In actuality, the amplitude is a slowly varying random process, and our primary interest is in the *average* error probability computed over the fading distribution. Assuming ergodicity holds for the process, the ensemble average performance we will compute would correspond to the time-averaged performance on an actual link. We should be aware though, that for any given channel the "instantaneous" error probability will fluctuate.

A practical difficulty associated with fading channels is that the demodulator must know the channel's scale factor A for optimal detection in those cases where the signals are not equal energy, for example, with on–off keying or 16-QAM. Because this is sometimes difficult to establish and because performance is sensitive to errors in this estimate, equal-energy schemes, notably M-PSK and M-FSK, are commonly utilized on fading channels. Consequently, we shall focus on these cases, although the analysis we follow is easily extendable to other situations. Also, normally coincident with time-varying amplitude is a time-varying channel phase, whose rate of change is on the same order as that of the amplitude. If the demodulation is to be coherent, this time-varying channel phase must also be estimated, a procedure made more difficult by the occasional deep fades. Thus, in practice we typically find noncoherent detection utilized on strongly fading channels.

Analysis of various cases is procedurally straightforward and identical for all techniques we consider. The error probability conditioned on a fixed channel gain $A = a$ is determined, as performed earlier in this chapter, then we average this conditional error probability with respect to the random variable A. That is,

$$P_s = \int_0^\infty P(\epsilon|a) f_A(a)\, da. \tag{3.6.1}$$

This procedure is quite general, making only a slow-fading assumption, and is applicable to other slow fading models such as Rician and log-normal (see Exercise 3.6.2).

Let's consider the Rayleigh channel in particular. Recall that the p.d.f. for A, again assuming a mean-square value of 1 for the random gain parameter, is

$$f_A(a) = 2ae^{-a^2}, \qquad a \geq 0. \tag{3.6.2}$$

In keeping with the model of Figure 3.1.1, the average, or expected, signal energy received per symbol will then be E_s. We shall use E_s and E_b to denote, respectively, the *average* symbol energy and *average* energy per bit communicated.

To establish the procedure and the general nature of the results, we first analyze the coherent detection of binary antipodal signals and binary orthogonal signals, as well as DPSK and noncoherent detection of binary orthogonal signals. Extension to other M-ary signaling cases is then made for both coherent and noncoherent detection.

3.6.1 Binary Signaling with Rayleigh Fading

The principal binary signal designs of interest are antipodal and orthogonal, represented by PSK and FSK, respectively. Both can be detected coherently or noncoherently (PSK in the form of DPSK), and receiver gain control is not crucial.

The error probability for ***binary PSK (antipodal signaling)***, given an available energy per bit of E_b joules, is $P_b = Q\left[(2E_b/N_0)^{1/2}\right]$. To generalize this for the case at hand, assume the selection of a specific channel amplitude $A = a$. Then, from our discussion of Section 3.3,

$$P(\epsilon|a) = Q\left[\left(\frac{a^2 2E_b}{N_0}\right)^{1/2}\right], \tag{3.6.3}$$

where $a^2 E_b/N_0$ is the instantaneous energy per bit-to-noise density ratio. The unconditional bit error probability is then obtained by averaging as in (3.6.1):

$$P_b = \int_0^\infty (2ae^{-a^2}) Q[(a^2 2E_b/N_0)^{1/2}]\, da. \tag{3.6.4}$$

It is now convenient to introduce the random variable $Y = A^2 E_b/N_0$. Whereas A is Rayleigh distributed, Y has a one-sided exponential density given by (see Example 2.11)

$$f_Y(y) = \frac{1}{E_b/N_0} e^{-y/(E_b/N_0)} \qquad y \geq 0, \tag{3.6.5}$$

where again E_b/N_0 is interpreted as the expected energy per bit-to-noise power density ratio. With this definition, (3.6.4) may be rewritten as

$$P_b = \int_0^\infty Q\left[(2y)^{1/2}\right] \frac{1}{E_b/N_0} e^{-y/(E_b/N_0)}\, dy. \tag{3.6.6}$$

Integration by parts gives

$$\begin{aligned} P_b &= -Q\left[(2y)^{1/2}\right] e^{-y/(E_b/N_0)} \Big|_0^\infty - \int_0^\infty e^{-y/(E_b/N_0)} \frac{1}{(4\pi y)^{1/2}} e^{-y}\, dy \\ &= \frac{1}{2} - \int_0^\infty \frac{1}{(4\pi y)^{1/2}} \exp\left(-y\left[1 + \frac{1}{E_b/N_0}\right]\right) dy. \end{aligned} \tag{3.6.7}$$

This last integral may be found in a table of definite integrals, and, upon simplifying, we obtain the (exact) result that

$$\boxed{P_b = \frac{1}{2}\left(1 - \left[\frac{E_b/N_0}{1 + (E_b/N_0)}\right]^{1/2}\right) \qquad \text{(coherent antipodal, Rayleigh fading)}} \tag{3.6.8}$$

The approximation $[x/(1+x)]^{1/2} \approx 1 - \frac{1}{2x}$ for x large allows us to estimate the error probability for large E_b/N_0 as

$$P_b \approx \frac{1}{4E_b/N_0} \tag{3.6.9}$$

This approximation is accurate provided $E_b/N_0 \geq 20$, or 13 dB.

The difference between this functional dependence and that found for the nonfading channel is quite profound. Specifically, to achieve a $P_b = 10^{-5}$ on the nonfading channel necessitates $E_b/N_0 = 9.6$ dB, while to do so on a Rayleigh fading channel requires $E_b/N_0 = 44$ dB, a roughly 2500-fold increase in signal-to-noise ratio! Furthermore, another decrease by a factor of 10 in error probability comes only at the expense of 10 dB increase in SNR.

Before proceeding to other cases, we should try to understand the basic difficulty with this channel and why brute-force methods, such as merely increasing the signal power, are an inefficient attack on the problem. The problem is simply that the error probability is heavily dominated by the infrequent, but deep, fading events. Under a slow-fading assumption, we can visualize the receiver operating point moving up and down the AWGN curve, such as in Figure 3.3.21 for PSK, as the channel amplitude changes. This averaging is illustrated in Figure 3.6.1. To simply model this, we might imagine a two-level approximation to the Q-function: we assume the error probability is $Q(2^{1/2}) = 0.079$ when the received energy-to-noise density ratio drops below 1, or 0 dB, and is zero if signal energy-to-noise density ratio exceeds 0 dB. Clearly, the resulting error

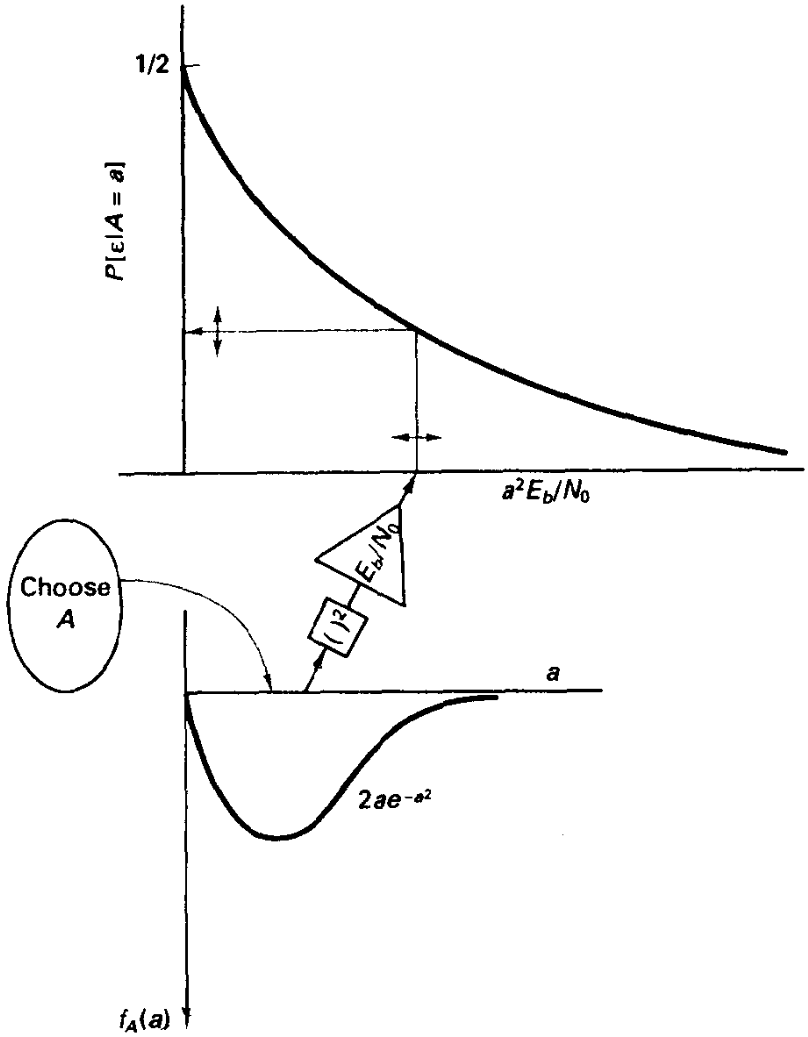

Figure 3.6.1 Nomograph for fading channel performance as an averaged performance obtained from nonfading analysis.

probability estimate, $0.079P(Y \leq 1)$, lower-bounds the true result. According to probability distribution for the random variable Y,

$$P(Y \leq 1) = \int_0^1 f_Y(y)\,dy = 1 - e^{-[1/(E_b/N_0)]}. \qquad (3.6.10)$$

We can truncate a Taylor series expansion for the exponential and obtain the lower bound

$$P(Y \leq 1) \geq \frac{1}{E_b/N_0} - \frac{1}{2(E_b/N_0)^2}. \qquad (3.6.11a)$$

This in turn implies that the probability of error, averaged over the fading random variable, is lower-bounded by

$$P_b \geq \frac{0.079}{E_b/N_0}\left[1 - \frac{1}{2E_b/N_0}\right]. \qquad (3.6.11b)$$

Although a crude argument, this demonstrates that increasing the average SNR on the channel only slowly diminishes the probability that the channel will be found in the below-threshold region, and hence only slowly reduces P_b.

We can also say something about the distribution of the (random) error probability whose expected value was given previously. The Markov inequality, for example, will hold that, with 0.9 probability, the instantaneous error probability is no worse than $10P_b$.

Proceeding as before for the case of ***coherent orthogonal signaling***, say with FSK, we obtain

$$\boxed{P_b = \frac{1}{2}\left[1 - \left(\frac{E_b/N_0}{2 + E_b/N_0}\right)^{1/2}\right] \qquad \text{(coherent orthogonal, Rayleigh fading),}} \tag{3.6.12}$$

which for large SNR behaves as

$$P_b \approx \frac{1}{2E_b/N_0}. \tag{3.6.13}$$

This points to a 3-dB loss for orthogonal signaling relative to antipodal signaling, which should not be surprising given our study of performance on nonfading channels and the graphical interpretation of Figure 3.6.1. In fact the 3-dB difference is exact at all values of E_b/N_0.

Next we analyze ***noncoherent detection*** of binary orthogonal signaling (say FSK), as well as DPSK. For noncoherent detection of orthogonal signals, the conditional error probability is

$$P(\epsilon|a) = \frac{1}{2}e^{-a^2E_b/2N_0}. \tag{3.6.14}$$

Averaging this as in (3.6.1) with respect to the Rayleigh density function for the fading amplitude, we have

$$\begin{aligned} P_b &= \int_0^\infty (2ae^{-a^2})\frac{1}{2}e^{-a^2E_b/2N_0}\,da, \\ &= \int_0^\infty ae^{-a^2(1+E_b/2N_0)}\,da, \end{aligned} \tag{3.6.15}$$

which integrates to

$$\boxed{P_b = \frac{1}{2 + (E_b/N_0)}, \qquad \text{(orthogonal, noncoherent, Rayleigh fading).}} \tag{3.6.16}$$

Notice again the inverse dependence on mean energy-to-noise density ratio for large SNR and that the efficiency is a factor of 4, or 6 dB, poorer than that of coherent PSK for high signal-to-noise ratios. This is a somewhat larger gap than experienced on the nonfading channel.

DPSK, at any fixed signal level, is exactly 3 dB more efficient than FSK with noncoherent detection. Carrying out the same averaging as before for DPSK would thus give

$$\boxed{P_b = \frac{1}{2+(2E_b/N_0)}, \qquad \text{(DPSK, Rayleigh fading),}} \tag{3.6.17}$$

which remains 3 dB superior to noncoherent orthogonal transmission in energy efficiency under fading conditions. This relative superiority is not restricted to the Rayleigh case, but would pertain to any slow-fading channel.

Figure 3.6.2 shows these four binary detection performances versus E_b/N_0. Notice that all have the same slope, -1, on a logarithmic plot, for high SNR, equivalent to the statement that P_b depends inversely on E_b/N_0. The comparison of these results with the

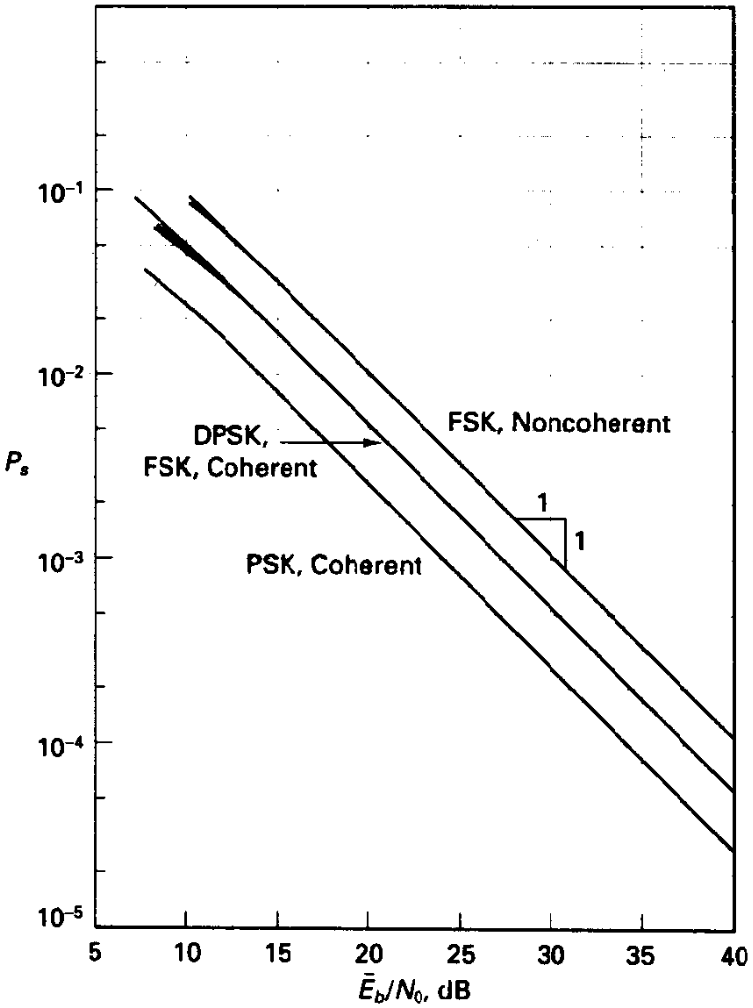

Figure 3.6.2 Bit error probability for binary signaling, Rayleigh channel.

corresponding results on the fixed-amplitude AWGN channel is striking. We must supply much greater average E_b/N_0 on the fading channel to achieve a given error probability, some 20 to 40 dB more, depending on the desired performance. This qualitative statement pertains to all the binary transmission strategies. Also, for the AWGN channel we found that PSK and DPSK were asymptotically equivalent in energy efficiency at high SNR, and likewise coherent and noncoherent detection of orthogonal signals is asymptotically equivalent, and at a typical error probability target of $P_b = 10^{-5}$, the difference in AWGN channel efficiencies is about 0.8 dB. For the Rayleigh channel, however, we have established that the noncoherence penalty is 3 dB in each case, at least in the high SNR regime. The reason has to do again with the dominance of the average error probability by the deep fading events, that is, when A is small. Careful comparison of error probability plots for the fixed-gain channel in the low SNR region, say for $E_b/N_0 < 3$ dB, will reveal this inferiority of the noncoherent techniques.

3.6.2 *M*-ary Orthogonal Signaling with Noncoherent Detection

Next we turn to M-ary orthogonal signaling on the Rayleigh fading channel. We shall emphasize the noncoherent detection case for two reasons. First, it is difficult to maintain phase coherence in the receiver in fading events, and, second, for large M, noncoherent detection performs comparably with coherent detection, as we have seen in Section 3.4. The expression derived earlier, (3.4.29), for the symbol error probability of noncoherent detection of M-ary orthogonal signal sets gives the conditional error probability

$$P(\epsilon|a) = \sum_{j=1}^{M-1} \frac{(-1)^{j+1}}{j+1} C_j^{M-1} \exp\left[-\frac{ja^2E_s}{(j+1)N_0}\right]. \tag{3.6.18}$$

To determine the unconditional error probability, we simply average (3.6.18) term by term with respect to the random variable A. The integrand involved in each term is a simple exponential form, and the resulting expression for P_s is

$$\boxed{P_s = \sum_{j=1}^{M-1} \frac{(-1)^{j+1}C_j^{M-1}}{1+j+(jE_s/N_0)} \qquad (M\text{-ary orthogonal, noncoherent, Rayleigh fading}).} \tag{3.6.19}$$

where again $E_s = (\log_2 M)E_b$.

The symbol error probability is shown in Figure 3.6.3 versus E_b/N_0 for $M = 2$, 8, and 32. All cases exhibit the same inverse dependence on E_b/N_0 for large average SNR, again showing that a simple energy-increasing attack to improve link performance is quite expensive. Also, we see that increasing M is only marginally helpful on the Rayleigh channel, in some contrast to the result for the AWGN channel. A qualitative rationale for this is that performance is dominated by low-amplitude events, and large M signaling is little better than $M = 2$ signaling for small instantaneous SNR, as may be seen in Figure 3.4.5 for example.

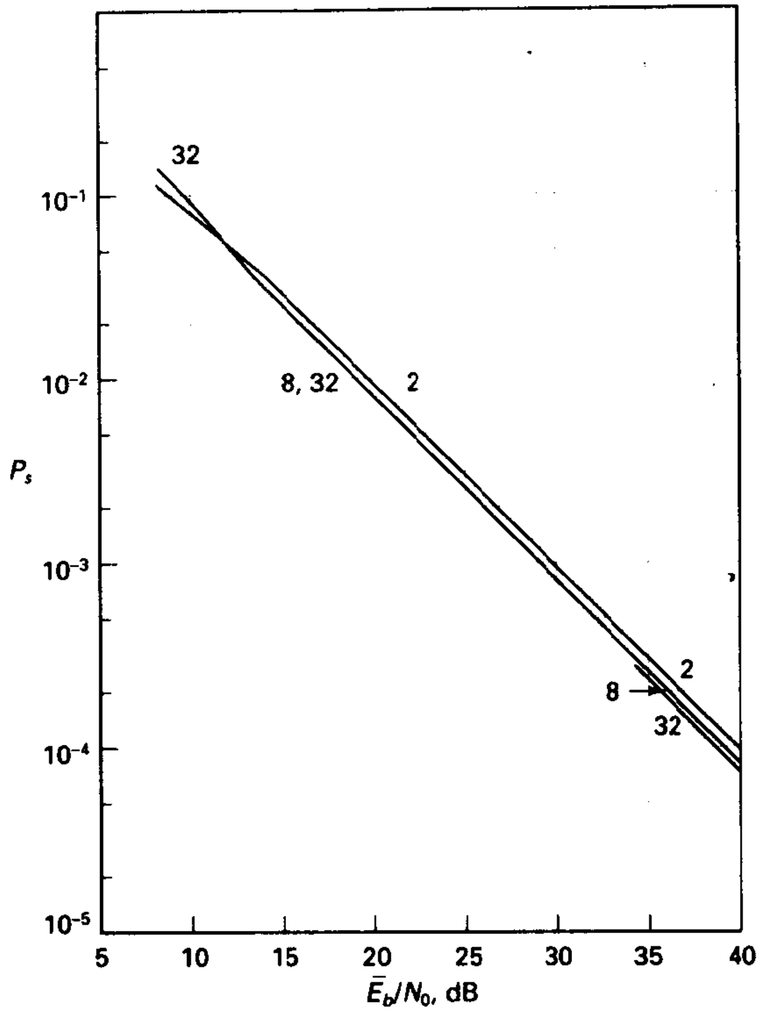

Figure 3.6.3 Symbol error probability for M-ary orthogonal signals, Rayleigh channel, noncoherent detection.

3.6.3 M-ary PSK and DPSK

Analysis of M-PSK and M-DPSK is in principle straightforward, but analytically tedious, and we shall omit the details. Proakis [31] devotes Appendix 7A to the exact treatment and obtains remarkably similar expressions for the two cases. For large SNR, which is the usual case of interest, the expressions simplify to

$$\boxed{P_s \approx \frac{M-1}{2M(\log_2 M)\sin^2(\pi/M)E_b/N_0} \quad (M\text{-PSK, Rayleigh fading}) \qquad (3.6.20)}$$

and

$$\boxed{P_s \approx \frac{M-1}{M(\log_2 M)\sin^2(\pi/M)E_b/N_0} \quad (M\text{-DPSK, Rayleigh fading}) \qquad (3.6.21)}$$

The first result can quite easily be argued to be approximately correct. We use the upper bound for M-PSK detection on a fixed-gain channel

$$P(\epsilon|a) < 2Q\left[\left(\frac{a^2 2E_b}{N_0}\log M \sin^2\frac{\pi}{M}\right)^{1/2}\right], \tag{3.6.22}$$

and then average this conditional probability of error over the distribution for channel amplitude, using integration by parts as earlier demonstrated for binary modulation.

The asymptotic results of (3.6.20) and (3.6.21) display a 3-dB difference in performance[25] on the Rayleigh channel, which is anticipated, given that the DPSK receiver uses two noisy phase measurements to form its decision, rather than one. Figure 3.6.4

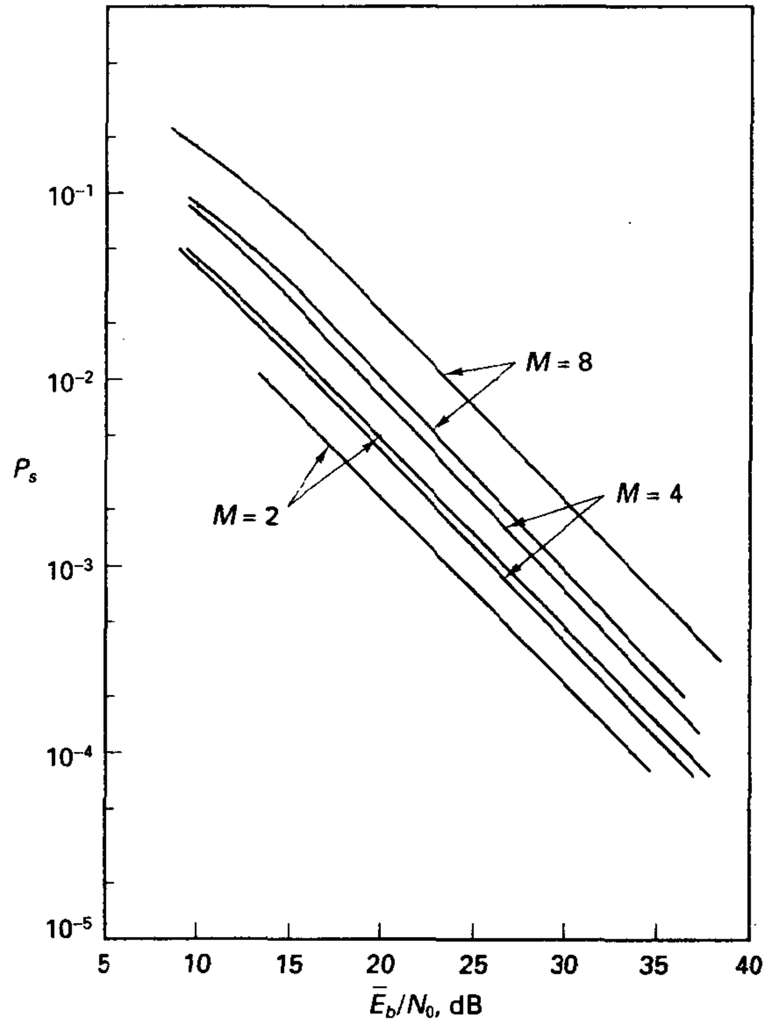

Figure 3.6.4 PSK and DPSK symbol error probability, Rayleigh channel. Lower of each pair is PSK.

[25] Notice that here a factor of 2 difference in error probability translates to 3 dB in energy efficiency, whereas on the AWGN nonfading channel, the energy penalty attached to a factor of 2 shift in probability is small, due to exponential dependence on E_b/N_0.

shows the performance for $M = 2$, 4, and 8.

The overriding message of the section should be clear: on the slow, flat-fading Rayleigh channel, the error performance for all modulation formats is markedly changed to a simple inverse dependence on SNR. Attempts to improve link quality by obvious engineering methods are expensive, but various forms of coding (including various "diversity" methods) will be seen to yield enormous improvements in this rather dismal situation. We shall resume this study in the next several chapters.

3.7 POWER SPECTRA OF DIGITALLY MODULATED SIGNALS

3.7.1 Overview on Power Spectrum and Some Cautions

In the previous sections of this chapter, our attention has been on the description of modulator signal constellations and on the error probability of optimal demodulators under various channel and receiver assumptions. We now shift the focus to a more transmission-oriented concern—the nature of the digital signal's ***power spectral density***, or power spectrum for short. In many applications the power spectrum of the transmitted signal is of just as much interest as the energy efficiency, and in some situations, for example, high-density magnetic recording and close-packed frequency-division channelization schemes, power spectrum issues may be foremost in selection of the signaling format.

There are several reasons why a detailed understanding of the power spectrum is important. First, if we are designing a system to communicate through a certain channel with special frequency response, either induced by electronic equipment or by the physical medium, we must have some notion of the power spectrum to be able to determine the channel's effect on the signals that are used in transmission. Knowledge of the channel response may strongly dictate our choices as to modulation; if a channel has poor low-frequency response, then signals with spectra concentrated at low frequency, for example, the baseband NRZ format introduced in Section 2.5, are poor candidates. (We should caution that the power spectrum, which is an averaged, second-order property of a signal, may obscure certain rare but important signal patterns that actually limit performance, and judgment of a signal's suitability should not be based on the spectrum alone. Furthermore, we may pass a digital signal through an all-pass linear filter, which does not alter the power spectrum but whose attendant phase distortion may produce disastrous effects on performance—once again, power spectrum is not by any means a total description.)

. Often, regulatory constraints imposed by bodies such as the Federal Communications Commission in the United States and similar telecommunication authorities in other countries force the power spectrum to meet certain constraints, and doing so requires either theoretical or empirical knowledge of the transmitted signal's power spectrum. A practical example might require that a microwave digital radio transmitter produce a power spectral density, measured in a 1-kHz bandwidth at all frequencies more than 10 MHz from the center frequency, at least 60 dB below the total signal power. Such restrictions are often expressed in the form of a spectral *mask* that the power spectrum must satisfy.

A final reason for our interest is the question of interference between different transmissions in a channelized multiuser system. An example could be the use of frequency-division access in mobile radio systems, where adjacent channel crosstalk due to the spectral overlap is a primary concern, especially given the varying proximity of users and fading possibilities. Analysis of the power spectra can help assess the amount of interference to be anticipated.

In communications parlance, we frequently encounter reference to ***bandwidth***, pertaining to the spectral extent of signals. This has potential for misinterpretation; indeed the very definition of bandwidth is elusive.[26] In a formal sense, most of the signals we encounter have infinite spectral extent, either as baseband or bandpass signals. Any signal produced as a time superposition of time-limited signal shapes must possess a Fourier transform that has infinite extent in frequency, a basic result of signal theory. Nonetheless, typical signals can be characterized as having a range of frequencies in which most of the power is located. More precisely, it is common to specify the frequency range, or bandwidth, that includes 90%, 99%, 99.9%, and so on, of a signal's power. Equivalently, we can specify the fraction of power outside a given range, leading to the power-out-of-band specification.

It is also possible for a certain modulation format to be *bandwidth efficient* for some applications, yet not so in other senses of the word. A signal may possess very low spectral sidelobes at large frequency separation from the center frequency and thus be a low source of interference to other channels, yet this may have been achieved at the expense of widening the main lobe of the spectrum, making the signal more sensitive to effects of channel filtering.

The power spectrum is a signal property derived from a probability model that we would hope reflects the power distribution versus frequency for the modulator output induced by any sequence of inputs, at least in the long-term sense. An ergodic property, which we shall assume, would hold that the power spectrum computed by time-averaging on a single sample function of the process converges to that obtained by probabilistic methods. Of course, if our probabilistic model is not representative of the modulation system, these two assessments of power spectrum may be quite different. For example, it may be typical that from certain digital sources or source encoders a sequence of one type of symbol persists for long periods, or certain pairs of symbols are highly likely, whereas the statistical model assumes independence. This is merely the usual difficulty with models—they are just that.

A related issue is that the power spectrum, by definition a long-term description, describes observations over a long interval. Many measurements, whether obtained by analog spectrum analysis or digital signal processing, are short-term statistics in some sense, and we must be careful when interpreting such results. These are particularly sensitive to the actual versus modeled behavior of the modulator input sequence; scramblers are commonly inserted in the transmission path, in fact, to counter the possibility of long runs of one type of symbol, for example.

With these caveats, we now proceed to develop important power spectrum relationships.

[26] Amoroso [35] discusses several common notions of bandwidth.

3.7.2 Power Spectrum for General Memoryless Modulation

Our basic premise about modulation stated at the beginning of the chapter is that modulation is a *memoryless process*; that is, every T_s seconds the modulator produces one of M signal waveforms, according to the symbol x_n presented to it, and time-superposes this with the other modulator responses as in (3.1.1):

$$s(t) = \sum_n s_{x_n}(t - nT_s). \qquad (3.7.1)$$

Normally, the probabilities of the M signals are equal, and a common model is one for which the input symbols are chosen in independent fashion.[27] On the other hand, in many situations the signals are not selected independently, specifically when the modulator input sequence $\{x_n\}$ is coded. Such coding may be either for purposes of improving the communication reliability in the presence of noise, or for shaping the spectrum of the signal, or both. In any case, it is frequently possible to describe the coded input in a finite-state Markov framework.

We provide the details of the derivation of the power spectrum for the general case in Appendix 3A3, but summarize the method here and emphasize application of the result. We view the signal as a sample function from a random process, induced by the driving sequence $\{x_n\}$. A general representation for the modulator output is given in (3.7.1). However, this random process is not in general wide sense stationary, but *wide sense cyclostationary*, meaning that the mean and autocorrelation function are periodic with period T_s in this case. By computing the period-averages of these, we obtain the usual mean and autocorrelation that would result from a time randomization in the definition of the process. Fourier transformation then yields the desired power spectrum.

In Appendix 3A3 the general result for M-ary modulation of the form (3.7.1), when the digital input is Markovian, is, from (3A3.12),

$$\begin{aligned} G_s(f) &= \frac{1}{T_s^2} \sum_{n=-\infty}^{\infty} \left| \sum_{i=0}^{M-1} P_i S_i\left(\frac{n}{T_s}\right) \right|^2 \left[\delta\left(\frac{f-n}{T_s}\right) \right] \\ &\quad + \frac{1}{T_s} \sum_{i=0}^{M-1} \sum_{j=0}^{M-1} P_i S_i(f) S_j^*(f) \left[\sum_{m=-\infty}^{\infty} \left(a_{ij}^{(m)} - P_j\right) e^{-j2\pi f m T_s} \right], \end{aligned} \qquad (3.7.2)$$

where $S_i(f)$ is the Fourier transform of the ith signal, $s_i(t)$, P_i is the marginal probability of the ith signal, and $a_{ij}^{(m)}$ is the probability that index $s_j(t)$ is transmitted m time units following transmission of $s_i(t)$. These m-step transition probabilities are entry (i, j) in the matrix $\mathbf{A}^m$ for a regular Markov sequence. Tausworthe and Welch apparently first produced this general result [36], although Bennett [37] earlier derived a special case.

The signals employed in this formulation are arbitrary; however, the representation of (3.7.1) must be valid, and there is a bit of subtlety involved. If the modulation is bandpass, and the carrier frequency is not synchronous with the symbol rate, then it is necessary to first derive the power spectrum of the complex envelope signal $\tilde{s}(t)$,

[27] Note that the fact that a modulator is memoryless does not imply that the input symbols are statistically independent.

defined in

$$s(t) = \mathrm{Re}\{\tilde{s}(t)e^{j\omega_c t + \Theta}\}, \tag{3.7.3}$$

and then apply the shifting principle for power spectra:

$$G_s(f) = \frac{1}{4}G_{\tilde{s}}(f - f_c) + \frac{1}{4}G_{\tilde{s}}(-f - f_c). \tag{3.7.4}$$

Assuming that the baseband equivalent signal has spectrum confined to $[-f_c, f_c]$, the resulting bandpass spectrum is basically a replica of the baseband power spectrum.

The first term in (3.7.2) represents possible spectral line components arising from periodicities in the autocorrelation function, while the second term represents a continuum spectrum. Such spectral lines may contain useful timing information; for example, in carrier transmission a spectral line at the carrier frequency can be used to extract signal phase in the demodulator. Others may be employed to extract symbol timing. In any case, these lines must be understood as otherwise wasteful of signal power, and the spectral concentration of power may be a source of strong narrow-band interference to other users.

An important special case of this general expression is that for which the input sequence is *independent and equiprobable*. Then, since

$$a_{ij}^{(n)} = \begin{cases} \delta_{ij}, & n = 0, \\ P_j, & n \neq 0, \end{cases} \tag{3.7.5}$$

we obtain

$$\begin{aligned} G_s(f) = {} & \frac{1}{M^2T_s^2}\sum_{n=-\infty}^{\infty}\left|\sum_i S_i\left(\frac{n}{T_s}\right)\right|^2 \delta\left(f - \frac{n}{T_s}\right) \\ & + \frac{1}{T_s}\left[\sum_i \frac{1}{M}\left|S_i(f)\right|^2 - \left|\sum_i \frac{1}{M}S_i(f)\right|^2\right], \end{aligned} \tag{3.7.6}$$

which depends only on the Fourier transforms of the various signals.

In any case, spectral lines *may* exist only at multiples of the symbol rate, R_s, as indicated in (3.7.2), and will be present at $f = n/T_s = nR_s$ unless the Fourier transforms evaluated at that same frequency sum to zero. A sufficient (and necessary) condition for *all spectral lines to vanish is*

$$\sum_{i=0}^{M-1} P_i s_i(t) = 0, \qquad \text{for } \textit{all } t, \tag{3.7.7}$$

which is a common symmetry condition, for example, equiprobable signals with antipodal, biorthogonal, and M-PSK/QAM signal sets. Notice, however, that symmetric signal sets with nonequiprobable selection may produce spectral lines.

The following example illustrates the general solution, and important special cases follow.

Example 3.13 Power Spectrum for 4-ary PPM

Suppose the modulation is baseband rectangular pulse PPM with $M = 4$ signals. Let the amplitude of each pulse be A and the pulse duration be $T_s/4$, where $T_s = 2T_b$ is the symbol interval. We define the basic signals on $[0, T_s]$. The Fourier transforms of the four signals are given by

$$S_i(f) = \frac{AT_s}{4}\frac{\sin^2(\pi f T_s/4)}{(\pi f T_s/4)^2}e^{-j2\pi i T_s/4}e^{-j2\pi T_s/8}, \qquad i = 0, 1, 2, 3, \tag{3.7.8}$$

and the magnitude-squared term in the line spectrum portion of (3.7.2) at frequency m/T_s has a scaling factor

$$\left|1 + e^{-jm\pi/2} + e^{-jm\pi} + e^{-jm3\pi/2}\right|^2. \tag{3.7.9}$$

This factor is seen to be zero, however, for all $m \neq 0$, and we find therefore a single spectral line at $f = 0$ with power $A^2/16$, which is just the squared average value of the signal set, or the d.c. value squared.

The continuum contribution is similarly determined and after a bit of manipulation becomes

$$G_c(f) = \frac{A^2T_s}{16}\frac{\sin^2(\pi f T_s/4)}{(\pi f T_s/4)^2}\left[1 - \cos^2\left(\frac{\pi f T_s}{2}\right)\cos^2\left(\frac{\pi f T_s}{4}\right)\right]. \tag{3.7.10}$$

Thus, the power spectrum has a $\text{sinc}^2(x)$ shape with first null at $f = 4/T_s = 2/T_b$, modulated in frequency by the term in brackets. Figure 3.7.1 presents the result for a signal with unit average power. One-fourth of the total power resides in the spectral line and three-fourths in the continuum component. For large frequency, the power spectrum is similar to that of

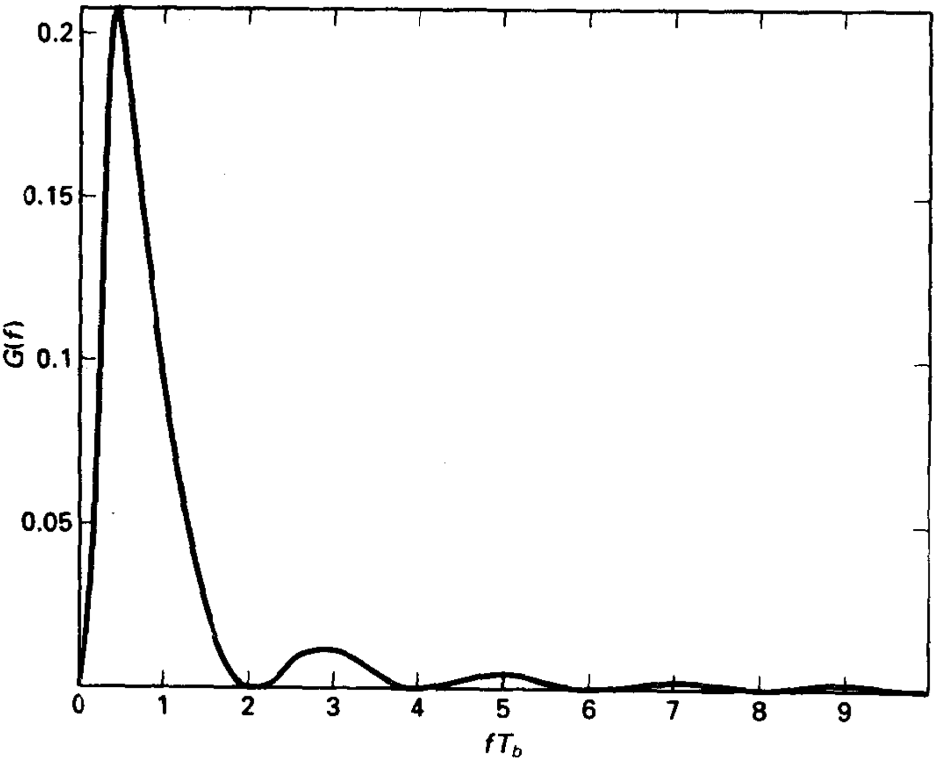

Figure 3.7.1 Continuum power spectrum for 4-ary PPM, average power = 1.

a binary NRZ signal with bit rate $2R_b$, showing bandwidth expansion for the 4-ary orthogonal set. In general, the bandwidth expansion is roughly $M/\log_2 M$.

3.7.3 Baseband Pulse-amplitude Signaling

Consider the simple, but widely applicable, case in which all signals are scalar multiples of some common waveform, with possible inclusion of a common bias term, or offset, in the multipliers. That is, we assume as in Section 3.3.5 that

$$s_i(t) = A[2x_i - (M-1) - B]\phi_0(t) \equiv a_i\phi_0(t), \tag{3.7.11}$$

where x_i are signal coefficients in the set $\{0, 1, 2, \ldots, (M-1)\}$ and $\phi_0(t)$ is a *unit-energy* baseband waveform, not necessarily limited to the interval $[0, T_s]$. In the communication literature, this is generally referred to as pulse-amplitude modulation (PAM). A is merely a scale factor related to energy normalization, and B is a possible bias. If $B = 0$, we have symmetric M-ary amplitude modulation, and all spectral lines vanish in (3.7.6) by symmetry, while if $B \neq 0$, the spectral line contribution to the total power spectrum is

$$G_l(f) = \frac{B^2A^2}{T_s^2}\sum_{n=-\infty}^{\infty}\left|\Phi_0\left(\frac{n}{T_s}\right)\right|^2\delta\left(f - \frac{n}{T_s}\right). \tag{3.7.12}$$

Notice that the spectral line contribution has an envelope that depends on the pulse shape adopted.

The continuum spectrum contribution is

$$G_c(f) = \frac{|\Phi_0(f)|^2}{T_s}\left(\frac{1}{M}\sum_{i=0}^{M-1}a_i^2 - \left|\frac{1}{M}\sum_{i=0}^{M-1}a_i\right|^2\right), \tag{3.7.13}$$

where a_i is defined in (3.7.11). The term in parentheses clearly does not involve frequency and is related only to the signal coefficients. In fact, this quantity is just the variance of the signal coefficient set. Thus,

$$G_c(f) = \frac{|\Phi_0(f)|^2}{T_s}\frac{M^2-1}{3}A^2 = \frac{E_s}{T_s}|\Phi_0(f)|^2, \tag{3.7.14}$$

and the spectral shape is (not surprisingly) purely determined by the pulse-shaping function $\phi_0(t)$. Remember that this result pertains to modulation with independent input symbols.

Example 3.14 Polar NRZ (Nonreturn to Zero) Baseband Transmission

A binary signal is assumed to be either A or $-A$ volts for a duration of T_s seconds and is the signal model adopted in the binary random wave process of Example 2.18. Thus, the power spectral density result will not be new, but the example provides a consistency check. The pulse $\phi_0(t)$ can be expressed as

$$\phi_0(t) = \left(\frac{1}{T_s}\right)^{1/2}, \qquad 0 \le t < T_s, \tag{3.7.15a}$$

and its Fourier transform is

$$\Phi_0(f) = T_s^{1/2}e^{-j\pi f T_s}\frac{\sin(\pi f T_s)}{\pi f T_s}. \tag{3.7.15b}$$

Furthermore, the modulation coefficients are $a_i = \pm A T_s^{1/2}$. By symmetry, all spectral lines vanish, and from (3.7.14) the power spectrum is

$$G_s(f) = A^2 T_s \frac{\sin^2(\pi f T_s)}{(\pi f T_s)^2}, \tag{3.7.16}$$

in agreement with the Fourier transform of (2.5.11). Figure 3.7.2 provides a logarithmic plot of this spectrum, with frequency normalized to the symbol (or bit) rate, R_s. We remark that the first and second spectral sidelobes are approximately 13.5 and 17 dB below the power spectral density at zero frequency and that the higher-order sidelobes decrease at a rate of 6 dB per octave,[28] a rather slow decay rate. Also, the main lobe can be shown, by integration, to contain roughly 90% of the total signal power (see Exercise 2.5.4.).

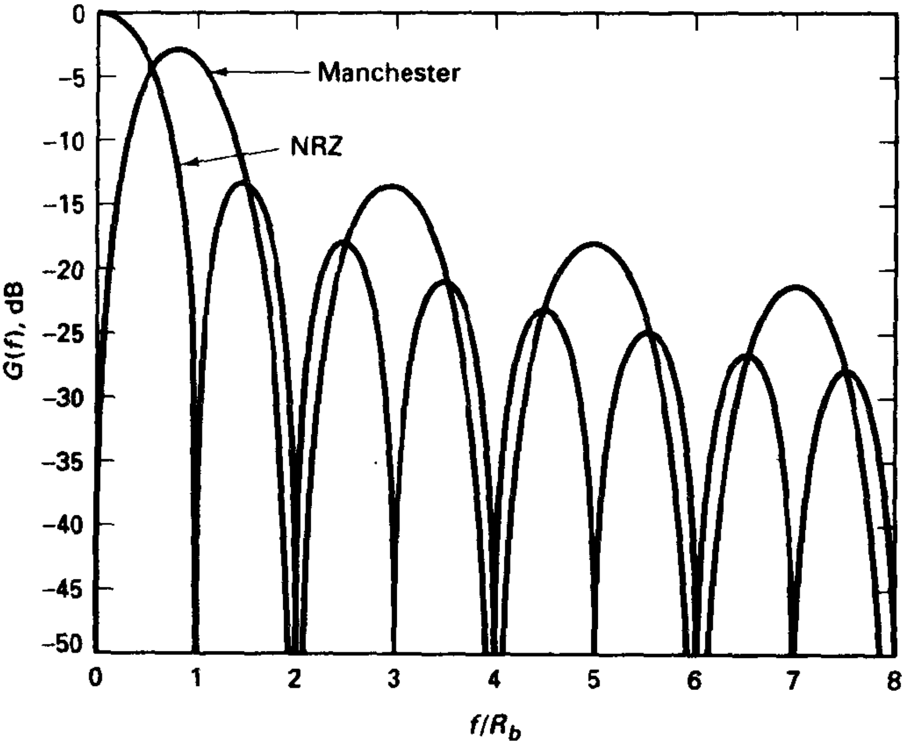

Figure 3.7.2 Power spectra for NRZ and Manchester signals.

If a bias were added to the signal so that the signal switched between 0 and $2A$ (producing unipolar NRZ), we would find that the resultant spectrum would have a continuum component identical with the previous case and a *single* spectral line at zero frequency, corresponding to the d.c. component. The power in this component is $B^2 = A^2$. Other spectral lines are absent because the Fourier transform $\Phi_0(f)$ happens to have zero magnitude at the possible spectral line frequencies, that is, at all multiples of the bit rate. Thus, the average power is a factor of 2 larger, while the instantaneous power is increased fourfold. Recall that this on–off modulation set was in fact a factor of 2 (or 4) less efficient than antipodal transmission under an average (or peak) power constraint.

[28] A frequency octave is a factor of 2 change in frequency.

Example 3.15 Binary Transmission with Manchester, or Biphase, or Split-phase Format

Suppose the two signals available in any interval are $A\phi_0(t)$ and $-A\phi_0(t)$, with $\phi_0(t)$ shown in Figure 3.7.3. The Fourier transform of the pulse is

$$\Phi_0(f) = \frac{T_s^{1/2}}{2}\frac{\sin(\pi f T_s/2)}{(\pi f T_s/2)}[e^{-j\pi f T_s/2} - e^{-j3\pi f T_s/2}]. \qquad (3.7.17)$$

It is then straightforward, using the Euler trigonometric identity for $\sin(x)$, to show that the power spectrum for biphase signaling is

$$G_s(f) = A^2 T_s \frac{\sin^4(\pi f T_s/2)}{(\pi f T_s/2)^2}, \qquad (3.7.18)$$

again absent of spectral lines by symmetry. The power spectrum is also shown in Figure 3.7.2, where we see that the spectral density at zero frequency is zero, and the first null in the spectrum is at $f = 2R_s$. If we define bandwidth as the location of the first null in the power spectrum, we would say the biphase signal's bandwidth is twice that of the NRZ signal; this is not at all surprising, given the fact that in a sequence of transmissions from either format the minimal dwell time at either polarity with biphase signaling is half that appearing in the NRZ stream. On the other hand, the spectral null at zero frequency is predictable from the fact that in any string of symbols the average value is zero, in contrast to the NRZ case, where arbitrarily long runs of A or $-A$ voltage levels are possible. The power spectrum of Manchester signals is apparently suited to channels with poor low-frequency response, for example, magnetic recording, where it has seen widespread use. If a bias term is added to the signal to provide unipolar signaling, we will see spectral lines at zero frequency (as with NRZ), as well as at *odd* multiples of the symbol rate. These spectral lines represent periodicities containing timing information for bit synchronizers, and unipolar biphase transmission is sometimes said to be *self-clocking* as a result.

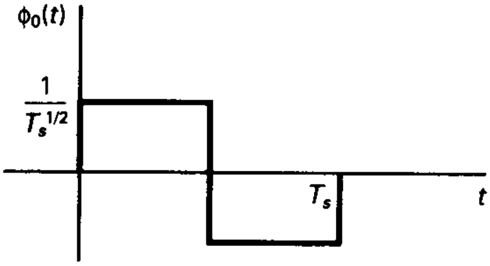

Figure 3.7.3 $\phi_0(t)$ for Manchester, or biphase, signals.

Example 3.16 *M*-ary Raised-cosine Signaling

Instead of the time-limited pulses assumed previously, we may adopt the pulse shape often referred to as the ***raised-cosine pulse***, so known not for its time-domain shape but for its Fourier transform. Given any $0 < \beta \leq 1$, we define the Fourier transform of the pulse be

$$\Phi_0(f) = \begin{cases} T_s^{1/2}, & 0 \leq |f| < \frac{1-\beta}{2T_s}, \\ T_s^{1/2}\cos^2\left(\frac{\pi T_s}{2\beta}\left[|f| - \frac{1-\beta}{2T_s}\right]\right), & \frac{1-\beta}{2T_s} \leq |f| < \frac{1+\beta}{2T_s}, \\ 0, & \frac{1+\beta}{2T_s} < |f|. \end{cases} \qquad (3.7.19)$$

which is illustrated in Figure 3.7.4. The name derives from the fact that, in the transition between passband and stopband of the frequency response, the characteristic is a raised-cosine characteristic. The spectrum is zero outside the frequency range $[0, (1+\beta)/2T_s]$. $1+\beta$ is known as the ***excess bandwidth factor***, since Nyquist [38] showed that the smallest bandwidth consistent with zero interpulse interference is the Nyquist bandwidth, $\frac{1}{2}T_s$.

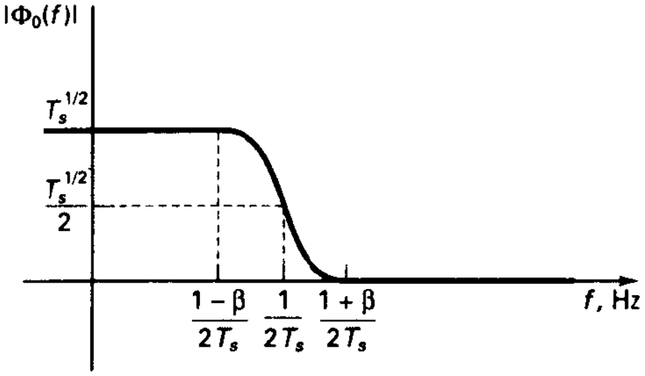

Figure 3.7.4 Frequency response for raised cosine Nyquist pulse; $1+\beta$ is excess bandwidth factor.

The time-domain expression for $\phi_0(t)$ is

$$\phi_0(t) = \frac{1}{T_s^{1/2}} \frac{\sin(\pi t/T_s)}{(\pi t/T_s)} \left[\frac{\cos(\beta\pi t/T_s)}{1 - 4\beta^2 t^2/T_s^2} \right], \tag{3.7.20}$$

which may be seen by recognizing that (3.7.19) is the frequency-domain convolution of a rectangular spectrum and a half-cycle cosinusoidal spectrum and then multiplying the respective inverse Fourier transforms. The time function is shown in Figure 3.7.5 for representative values of β. Because the Fourier transform is defined to be strictly band-limited, the signal $\phi_0(t)$ must have infinite time duration; in practice, some truncation could be utilized to approximate the ideal case. Each pulse carries unit energy.

The power spectrum for symmetric M-ary PAM transmission with this pulse is

$$G_s(f) = \frac{E_s}{T_s} |\Phi_0(f)|^2, \tag{3.7.21}$$

so the power spectrum is also strictly band-limited.

Actually, it is more common to utilize a modulator pulse whose Fourier magnitude spectrum is the square root of (3.7.19), which occupies the same transmission bandwidth, which when properly (matched) filtered produces zero intersymbol interference. This goes under the name ***square-root raised-cosine filtering***.

3.7.4 Spectra for M-PSK and M-QAM Modulation

Although more general pulse shaping can be applied, we assume that the modulated signal is, in the case of M-PSK,

$$s_i(t) = \left(\frac{2E_s}{T_s}\right)^{1/2} \cos\left(\omega_c t + \frac{2\pi i}{M} + \theta\right), \qquad 0 \le t < T_s. \tag{3.7.22}$$

In the case of M-QAM

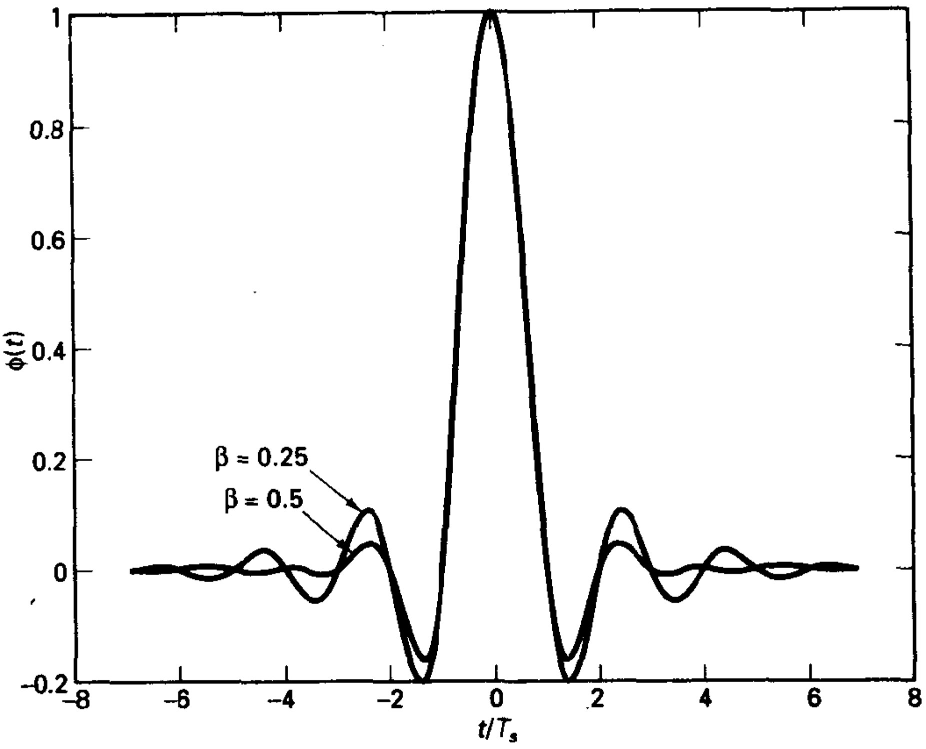

Figure 3.7.5 Nyquist pulses.

$$s_i(t) = a_i \left(\frac{2}{T_s}\right)^{1/2} \cos(\omega_c t + \theta) + b_i \left(\frac{2}{T_s}\right)^{1/2} \sin(\omega_c t + \theta), \qquad 0 \le t < T_s, \tag{3.7.23}$$

where the coefficients are selected from an M-ary QAM constellation. It is typically the case that the carrier frequency is large relative to the symbol rate, but is not synchronous, and the formulation of (3.7.1) is not strictly valid. In this case, both preceding modulations may be represented in complex envelope notation as

$$s_i(t) = \mathrm{Re}\{c_i \phi_0(t) e^{j\omega_c t} e^{j\theta}\}, \tag{3.7.24}$$

where $c_i \phi_0(t) e^{j\theta}$ is the complex envelope of the ith waveform, and $\phi_0(t)$ is a rectangular pulse. Here c_i are complex numbers of the form $a_i + jb_i$.

In both cases we have enough signal set symmetry[29] so that spectral lines vanish in (3.7.2). Computing the Fourier transforms, $S_i(f)$, and noting that all are related to each other by a complex number c_i, yields the bandpass spectrum

$$G_s(f) = \frac{E_s}{2}\left[\frac{\sin^2(\pi(f - f_c)T_s)}{(\pi(f - f_c)T_s)^2} + \frac{\sin^2(\pi(f + f_c)T_s)}{(\pi(f + f_c)T_s)^2}\right] \tag{3.7.25}$$

Except for a scale factor related to average energy, the spectrum expression is identical for all schemes having a common *symbol rate* R_s. At this point it is important to remember that M-ary schemes convey $\log_2 M$ bits per symbol, so $R_s = R_b / \log_2 M$, and

[29] At least for the constellations presented in Section 3.3.5.

the spectral widths scale down in frequency according to $\log_2 M$ for a given *bit* rate. Likewise, the symbol energy scales in a similar manner with M, relative to the bit energy E_b. In Figure 3.7.6, we show the *one-sided* power spectrum, relative to f_c, for M-ary PSK (or QAM) signaling, wherein we normalize frequency to the bit rate and normalize so that the energy per bit is 1. We remark that the null-to-null definition of bandwidth would yield $B = 2R_s$, assuming a rectangular pulse shape.

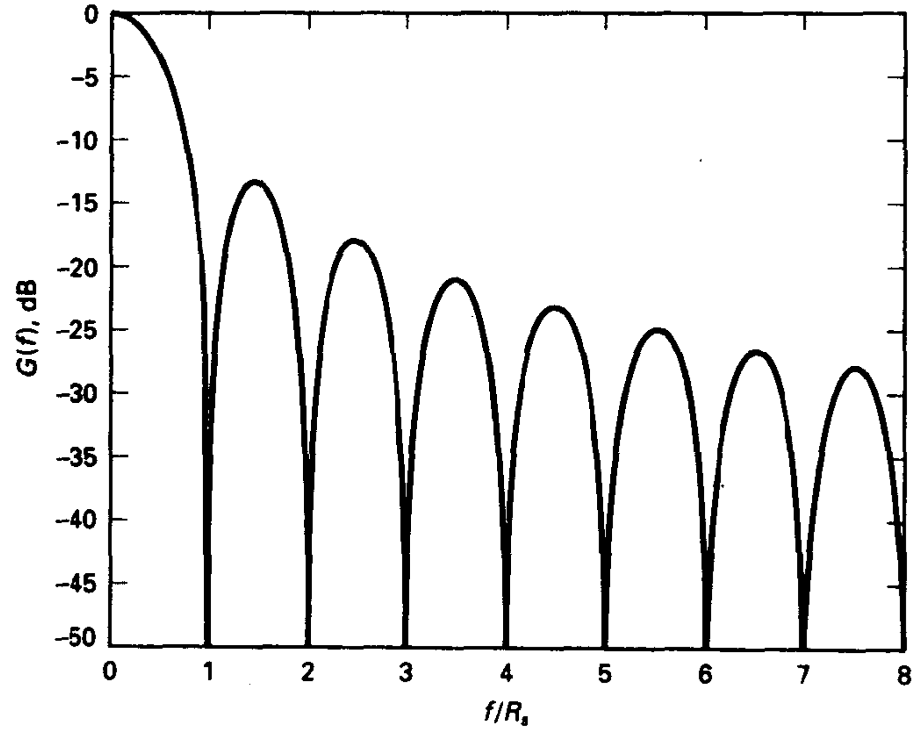

Figure 3.7.6 Power spectra for M-PSK/M-QAM.

Other pulse shapes may be chosen for spectrum control; biphase pulse (Manchester) shaping would widen the total spectrum, but would provide a region near the carrier frequency with near-zero spectral density. This is occasionally useful for purposes of adding a pilot carrier in transmission to assist with attaining a coherent phase reference in the receiver. On the other hand, raised-cosine shaping is a possibility for carrier transmission just as for baseband signaling.

Example 3.17 Satellite Transmission Using Pulse-shaped QPSK

Suppose it is required to transmit a 140-Mbps binary message stream through a satellite transponder whose nominal bandwidth is 72 MHz. If we adopt 8-PSK modulation, the symbol rate is $R_s = R_b/3 = 46.7$ MHz. Use of rectangular NRZ pulses (the easiest to implement) would produce a power spectral density having a null-to-null bandwidth of 93.3 MHz. Rather severe amplitude and phase distortion would occur in the transponder as a result. If, however, we adopt square-root, raised-cosine shaping with $\beta = 0.3$, the signal's power spectrum can be completely confined to a bandwidth of $2(1.3)R_s/2 \approx 60.7$ MHz. (The leading factor of 2 accounts for the two-sided nature of the bandpass spectrum centered at f_c.) Presumably, this signal is degraded less by the amplitude and delay distortion of the satellite transponder and the resulting intersymbol interference at the demodulator output.

We conclude with some rule-of-thumb relationships for power spectra that often give a rough assessment of the power spectrum.

3.7.5 Asymptotic Behavior of Power Spectrum; Role of Dimensionality

First, consider an arbitrary concatenation of signals selected from the modulator set, which we again express as

$$s(t) = \sum_n s_{x_n}(t - nT_s). \tag{3.7.26}$$

We view this signal as a deterministic signal produced by some message sequence. Suppose this signal extends over some large number of symbols, and let $S(f)$ be the Fourier transform of this waveform. (We assume that the transform exists.) The energy spectrum of the signal is, by definition, $|S(f)|^2$. It may be that the signal set is a baseband set or perhaps a set of carrier-modulated signals whose individual Fourier spectra are centered previously some frequency f_c. The fine structure of this energy spectrum depends strongly on the exact properties of the signal set, as outlined previously but the behavior of the spectrum for large frequency f (relative to the center frequency) is governed only by the smoothness properties of the signal set. Specifically, consider the time derivatives of the signals $s_i(t)$, and let N_d be the smallest order of derivative (or antiderivative) that is not continuous. Then it is known (see, for example, Bracewell [39]) that the envelope of the energy spectrum at large f behaves as

$$|S(f)|^2 \approx O\left(|f|^{-2(N_d+2)}\right). \tag{3.7.27}$$

(The "big O" notation connotes the dominating functional dependence for large values of the argument and is read "on the order of"; it does not convey the absolute level, but the functional behavior.) In other words, the asymptotic rate of decay of the energy spectrum in the high-frequency region (relative to the carrier frequency if any) is $2(N_d + 2) \cdot 6$ dB/octave. Notice that we have claimed nothing about the absolute level of power spectral density in these large-frequency sidelobes; in some sense this depends on the size and richness of discontinuities of the (N_d)th derivative.

Example 3.14 Revisited: Asymptotic Spectrum for NRZ Transmission

Suppose the modulation is the binary NRZ format. An arbitrary concatenation of the binary signals will have discontinuity at the symbol boundaries. However, the integral, or the -1st-derivative is everywhere continuous. Thus, $N_d = 0$ in the preceding terminology, and this implies that the energy spectrum for an arbitrary concatenation of bits decays as f^{-2}. This is consistent with our earlier determination that the *power* spectrum of the random binary wave decays as f^{-2}, since the power and energy spectra differ only by a time normalization.

The same result pertains to any transmission scheme that has similar discontinuities, such as the Manchester format. We have seen that for a given bit rate the Manchester spectrum is in some sense twice as wide as the NRZ spectrum; nonetheless, they have the same asymptotic rate of decay.

A simple change in the signal description, letting the two pulses be half-cycle sinusoids, renders the asymptotic rate of decay to be as f^{-4}, for now the first derivative is the smallest order of derivative that is not everywhere continuous.

Example 3.18 Distinction between Carrier-synchronous Modulation and Asynchronous Modulation

Let's consider two on–off signaling techniques, as shown in Figure 3.7.7a. For bit 0, the carrier is absent, and for bit 1, the carrier is turned on for a nominal $1\frac{1}{2}$ cycles (this small number is for illustrative purposes), and the signal begins each repetition at zero. In the first case, we draw the transmitted signal for the pattern 1101. Notice that this signal can be represented in the framework of (3.7.1); that is, the signal in interval n is a translation of the set available in interval 0. An arbitrary concatenation of such signals is everywhere continuous, and so the spectrum decays as $O(f^{-4})$. The actual power spectrum would be given by (3.7.2), needing only the Fourier transform of the one basic signal. A rather simple change in the formulation, retaining synchronism, but defining the 1 signal to begin and finish at a maximum, implies discontinuity and thus decay only as $O(f^{-2})$. Thus, the *starting phase* in such cases is crucial. A similar situation occurs with synchronism, but with, say, $1\frac{3}{4}$ cycles per bit.

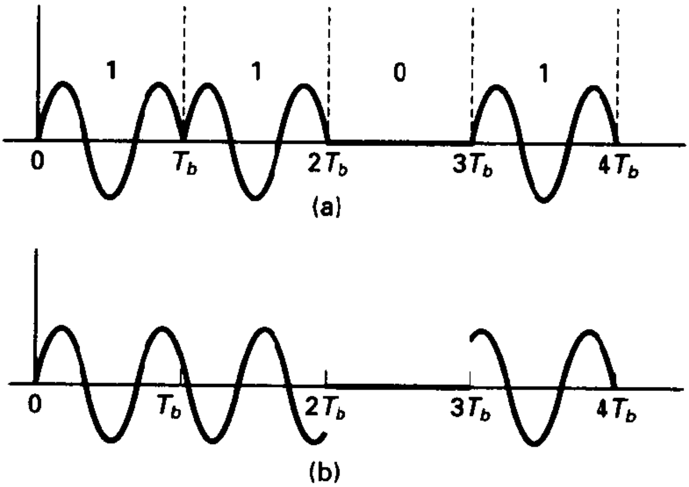

Figure 3.7.7 Modulator output for two cases of on–off signaling: (a) $f_c = 3R_b/2$, synchronous; (b) $f_c = 3R_b/2 - \epsilon$, no phase reset.

In contrast, suppose that the carrier frequency is nominally $1\frac{1}{2}$ times the bit rate, but we do not reset the signal at the start of each bit. Instead, we merely amplitude-modulate a carrier whose nominal frequency is about $1\frac{1}{2}$ times the bit rate. The same bit sequence might produce the signal pattern shown in Figure 3.7.7b. There are small but important differences, and the power spectra are slightly different. In particular, the signal is no longer continuous (in the practical sense it will be of course), and we would say the asymptotic decay is $O(f^{-2})$. The proper means of finding the exact power spectrum is to find the power spectrum of the baseband complex envelope, which is just the random binary wave in this case, and then use (3.7.4) to translate to the actual carrier frequency.

Still another variation on this would have exact synchronism, but such that discontinuities exist. For example, let the carrier frequency be exactly $1\frac{3}{4}$ cycles per bit. The method outlined is applicable, but the results should decay only as f^{-2}. Further discussion of these issues is found in Appendix 3A3.

Dimensionality and Bandwidth

Consider the modulator output over pT_s seconds, corresponding to p message symbols. If the signals available to the modulator extend longer than a symbol time, some truncation is involved, but with p large, this is a negligible effect. Suppose the signal over this interval is essentially band-limited to W hertz. The number of real orthonormal functions that can occupy this time interval and also have frequency confined to W hertz is [40]

$$N^* \leq 2WpT_s. \tag{3.7.28}$$

This must be loosely interpreted, since signals cannot be simultaneously exactly time limited and frequency limited; more precise statements can be found in [40]. Its validity is more solid for when the time–bandwidth product is large.

Now consider a modulator set having N orthonormal dimensions per symbol. If we wish p successive transmissions to not corrupt each other, then we would wish that time translates of the orthonormal basis functions also be orthogonal with each other. Thus, in p modulator intervals, lasting roughly pT_s seconds, we are seeking to define Np orthonormal functions having bandwidth confined to W hertz. By (3.7.28), we must have $N^* = Np \leq 2WpT_s$, or that the *minimum* bandwidth consistent with a signal set having dimensionality N dimensions/symbol be

$$W > \frac{N}{2T_s} = \frac{NR_s}{2} \quad \text{hertz} \tag{3.7.29a}$$

or

$$\frac{W}{R_b} > \frac{N}{2\log_2 M}, \tag{3.7.29b}$$

since each M-ary symbol conveys $\log_2 M$ bits. We can define $D = N/\log_2 M$ as the signal-space dimensionality per bit and then claim that

$$W > \frac{D}{2} \quad \text{hertz.} \tag{3.7.29c}$$

This signal theory result places a *lower bound* on the bandwidth occupancy of a digital signal, which is only a function of signal-space dimensionality per bit. The expression applies equally well to baseband and bandpass transmission.

For example, if we adopt a 64-QAM carrier modulation, having dimensionality per bit $\frac{2}{6}$ (two real orthonormal functions define the signal set), then the minimum ratio of bandwidth to bit rate is $\frac{1}{6}$, by (3.7.29c). We know this is approachable by use of pulse-shaped 64-QAM, wherein $\sin(t)/t$ pulse shaping is employed in each quadrature arm. The symbol rate in each quadrature channel is $R_b/6$, and the baseband signal prior to modulation can have bandwidth as small as $R_b/12$ hertz, while still maintaining orthogonality between successive symbols, or zero intersymbol interference. Modulation to a carrier frequency doubles the bandwidth to $W = R_b/6$ hertz.

On the other hand, suppose we use baseband 16-ary orthogonal signaling with PPM. Here, the signal-space dimensionality is 16 dimensions/4 bits, and the minimum bandwidth consistent with this signal set is $2R_b$, or the bandwidth expansion ratio is at least 2. Of course, if we use rectangular pulses to construct the basis (and thus the signal), we will find that the actual bandwidth is roughly twice as large and is hardly band-limited anyway.

The dimensionality theory can be badly abused in measuring signal bandwidth, for it neglects the spectral properties of the actual functions used to define the signal set. A prime example surfaces in Section 3.8. We can use antipodal modulation, with signals defined by binary-coded patterns with, say, 15 chips per bit. This pattern forms the single basis function used to describe the signal set. Thus, the dimensionality/bit is 1, indicating the *potential* for small bandwidth. In actuality, the true signal bandwidth is much wider, by design. In effect, the spectral properties of the basis functions used to construct the set are important, as well as the number of them.

3.7.6 Power Spectrum for Markov-Input Modulation

We return briefly to the general result presented in (3.7.2), derived in Appendix 3A3, showing one application of how precoding the modulator input can significantly shape the signal power spectrum. To apply the method, we need a valid state description for the modulator input sequence and need to find the steady-state probabilities of the various signals. Such Markovian dependencies may be introduced for error control coding purposes (improving the energy efficiency of the channel) or can be introduced specifically for spectral shaping. The case of alternative-mark-inversion transmission is an example of a three-level signaling technique, discussed at the beginning of this chapter, for which one of two possible waveforms is present in any interval, dependent on the previous selections.

Specifically, let the levels be 0, A, $-A$, and the selection rules be described by the modulation state transition matrix

$$\mathbf{A} = \begin{bmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix} \tag{3.7.30}$$

where A_{ij} represents the probability of selecting signal $s_j(t)$ immediately following transmission of $s_i(t)$.

By symmetry of the transition structure and the symmetry of the signal set, we have from (3.7.2) that all spectral lines vanish. Furthermore, the marginal probabilities for signals are given by $P_0 = \frac{1}{2}$, $P_1 = P_2 = \frac{1}{4}$.

3.8 SPREAD-SPECTRUM MODULATION

The focus on power spectra in the previous section probably suggests that frequency spectrum is a precious commodity and that designers generally seek to minimize the bandwidth occupied by a digitally modulated signal. This is often the case, but there are situations where the signal's bandwidth is intentionally made much larger, perhaps 1000 times larger, than the bandwidth implied by the basic message symbol rate $1/T_s$. Such a modulation process is known generically by the apt name ***spread-spectrum modulation.***

There are several reasons why exorbitant use of bandwidth may be tolerable or useful. The principal benefits are the following:

1. Spreading a fixed amount of transmitter power over a wide bandwidth lowers the power spectral density, inducing less interference to another narrowband signal occupying the same frequency range and making the presence of the signal less detectable by an eavesdropper.
2. By having various users employ proper modulation formats (spreading codes), we are able to achieve near orthogonality of waveforms despite the fact that many users share the same spectrum. This orthogonality, if strict, would allow multiple users to coexist in a given frequency range without mutual interference, providing multiple access through what is known as ***code-division multiple access*** (*CDMA*). The same principle makes spread-spectrum systems less vulnerable to intentional or unintentional interference.
3. Wide-bandwidth signals can provide precise time-of-arrival measurements for range determination and position location; this derives from the possibility of narrow autocorrelation responses attached to wideband signals.
4. Spread-spectrum signals enjoy a resistance to multipath interference, again owing to the narrow autocorrelation responses.

A vast literature on the topic of spread spectrum exists, including entire texts. Dixon's book [41] is a introductory treatment of the main themes, and the three volume set of Simon et al. [42] is perhaps the current ultimate account. Holmes's text [43] is another good presentation, although restricted to coherent spread-spectrum techniques. Given this situation, ts well as the practical interest in the applications of spread spectrum today, it may seem odd that the presentation here does not even achieve chapter status. This is by design—the basics of spread-spectrum transmission and reception are not essentially different from the material we have already seen, and the material is best understood in a unified presentation, rather than being perceived as exotic and fundamentally different. Some mistakenly regard spread-spectrum transmission as a form of coding, but we shall see it is actually a form of memoryless modulation of a carrier, albeit a rather nonstandard carrier.

Two principal forms of spread-spectrum modulation are encountered in practice: **direct sequence** (DS) **spread spectrum** and **frequency-hopping** (FH) **spread spectrum**. Hybrids of these exist, and other forms such as time hopping and chirp modulation have been studied for similar purposes. However, these are not common and will not be studied here.

3.8.1 Direct Sequence Spread Spectrum

DS spread-spectrum modulation is illustrated in Figure 3.8.1. A binary information sequence at rate R_b is modulo 2 added with a higher-speed binary ***pseudorandom code sequence*** $\{c_n\}$, often called a ***chip sequence***, producing a high-speed random sequence $\{m_n\}$, which in turn phase shift keys a carrier. The clock rate, or ***chip rate***, R_c, of the code sequence is B times faster than the information rate, and normally the respective clocks

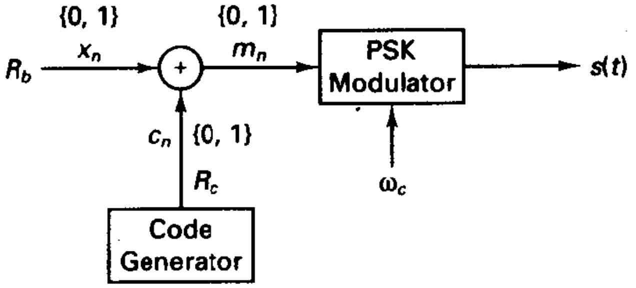

Figure 3.8.1 Direct sequence spread spectrum modulator.

are synchronously related at the transmitter. The synchronism is only a convenience for implementation (and analysis perhaps), but is not essential in principle.

A suitable chip sequence should have several attributes. It should appear, to a naive observer, much like a random binary sequence, although it must be deterministic in any practical setting so that cooperative communication can ensue. Specifically, the sequence should be balanced between 0's and 1's and should exhibit favorable autocorrelation properties, that is, low autocorrelation at all nonzero shifts of the sequence. Furthermore, in the multiuser CDMA setting, the codes act as signatures for the various users sharing the same channel, and reduction of mutual interference hinges on small cross-correlation among different pairs of sequences. Production of good code sequences has been the subject of much study in the past 30 years, much of it driven by military systems requirements. It is not within the scope of this text to develop this material, but it suffices to say that code generators are some form of shift-register network with output-feedback capable of producing a sequence with long period and perhaps low susceptibility to structural identification by an eavesdropper. The sequences are called pseudo-random because the sequence is deterministic and completely predictable by an informed party, while to a naive observer the sequences appear random.

The ***maximal-length sequences*** are generated by linear feedback shift register mechanisms and suffice for our understanding here. Superb treatments of these sequences are found in Golomb [44] and MacWilliams and Sloane [45]. It is known that, for any binary register length L, feedback connections exist for producing a code sequence with period $2^L - 1$, which is the maximal period for such a finite-state machine. The shift register encoders have a strong connection with finite field theory, taken up in Chapter 5, and in particular the proper feedback connections are provided by coefficients of primitive polynomials. Figure 3.8.2 illustrates a shift register encoder for a length-63 sequence, along with feedback connections for other length sequences. Maximal length codes have interesting properties: a balance (within 1) of 0's and 1's, proper frequency of strings of various types, and a (deterministic) autocorrelation function that has the desirable "thumbtack" shape shown in Figure 3.8.2c. There are relatively few maximal-length

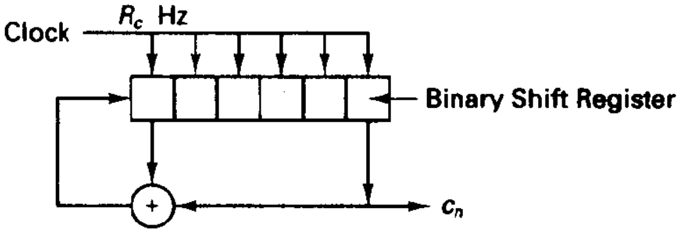

Figure 3.8.2a Maximal length sequence generator, $L = 63$.

L	Tap Connections
7	1, 3
15	1, 4
31	2, 5
63	1, 6
127	3, 7
255	2, 3, 4, 8
511	4, 9
1023	3, 10
$2^{15}-1$	1, 15
$2^{31}-1$	1, 31

Figure 3.8.2b Feedback connections for maximal length sequence generators.

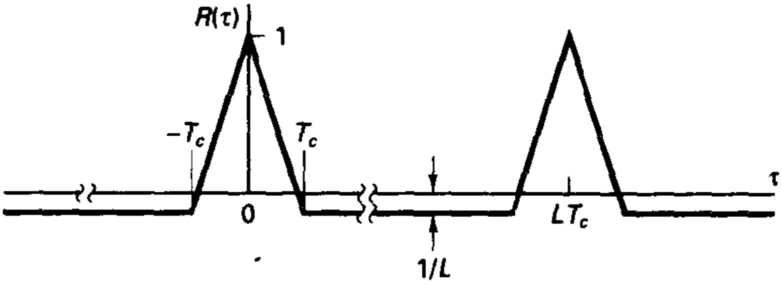

Figure 3.8.2c Normalized autocorrelation function for maximal length pseudorandom sequence, period is L.

sequences at any desired length, and to produce larger sets for CDMA applications, it is common to modulo 2 add the outputs of two preferred sequences with some designated phase shift, producing a ***Gold sequence*** [46], labeled by the adopted phase shift. These sequences have been shown to have good autocorrelation and cross-correlation properties. Exercise 3.8.2 examines these for length-15 sequences. Recently, nonbinary, for example, quadriphase, code sequences have been studied [47] as a means of further lowering the cross-correlation between signature sequences and have in fact achieved, asymptotically in N, the Welch bound [48] on the minimal cross-correlation achievable for M signals built from N-chip sequences.

In the binary DS case, the modulated signal is given by

$$s(t) = A \sin\left[\omega_c t + \theta + \frac{\pi}{2} m(t)\right] = -Am(t)\sin(\omega_c t + \theta), \qquad (3.8.1)$$

where $m(t)$ is the ±1 waveform related to m_n by mapping logical 0 to -1 and logical 1 to 1,[30] and θ is a random initial carrier phase. (It is sometimes helpful to regard the variables x_n, c_n, and m_n as having values ±1, in which case the modulo 2 addition can be exchanged for normal multiplication.)

For modeling purposes it is convenient to assume that the code sequence is a (fair) coin-flipping process so that m_n is a i.i.d. binary sequence for any underlying message. In this case the signal $s(t)$ is stochastically equivalent to a PSK signal modulated at

[30] The pulse waveform is usually the rectangular, or NRZ pulse, but generalization is possible.

rate $R_c = BR_b$. Correspondingly, the baseband equivalent spectrum of the modulated signal is

$$G_s(f) = A^2T_c\frac{\sin^2(\pi f T_c)}{(\pi f T_c)^2} = \frac{A^2T_b}{B}\frac{\sin^2(\pi f T_b/B)}{(\pi f T_b/B)^2}. \tag{3.8.2}$$

This illustrates both the spectral expansion by a factor of B and a lowering of the power spectral density, as shown in Figure 3.8.3.

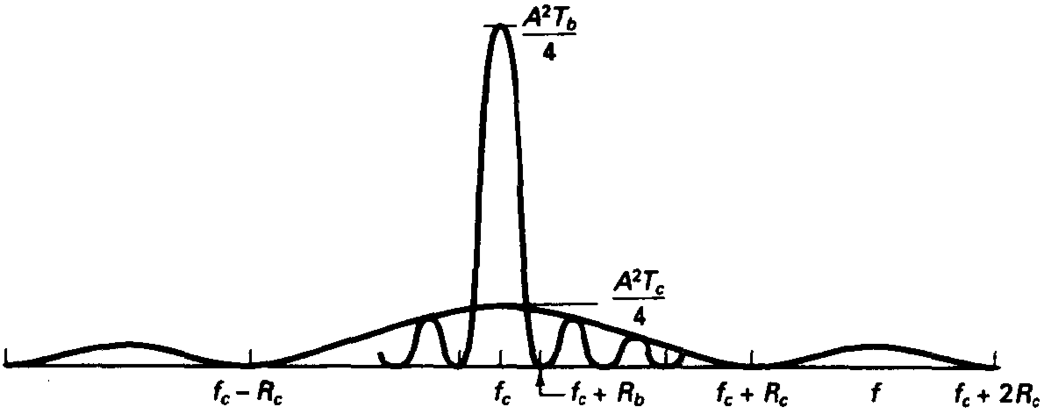

Figure 3.8.3 Power spectra for spread and non-spread modulation, positive frequency portion shown.

Although Figure 3.8.1 illustrates the typical implementation, it is clear that Figure 3.8.4 is equivalent, which highlights the fact that we are really just impressing the information sequence on a more exotic carrier, $c(t)\sin(\omega_c t + \theta)$, where $c(t)$ is the ±1 code sequence expressed as a function of time. In fact, the modulation of this nonstandard carrier is *antipodal* in DS spread spectrum, since over the nth message bit interval $s(t) = x_n c(t)\sin(\omega_c t + \theta)$. This should suggest certain equivalences with nonspread performance.

Detection of DS spread-spectrum modulation usually follows correlation receiver structures already developed. (The matched filter version of the receiver is less attractive here unless the code sequence repeats every message bit, although matched filters are

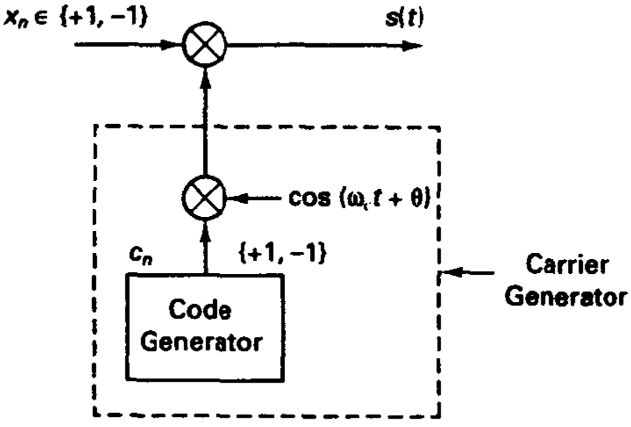

Figure 3.8.4 Alternative realization of direct sequence modulation; antipodal modulation of code-modulated sinusoid.

often found in the initial synchronization of the receiver code generator.) Figure 3.8.5 presents the coherent DS receiver, which in effect employs a scaled version of $c(t)$ $\cos(\omega_c t + \theta)$ as a basis function. The correlator integrates over one bit interval (not chip interval), and comparison with a zero threshold provides optimal data decisions. It is required that, as for coherent detection of nonspread signals, the carrier reference be properly phase aligned. Here, in addition, we require that the local code generator be synchronized to the incoming code sequence to within a small fraction of a *chip duration*, or else the correlator output will be small (see Figure 3.8.2c). Therein lies the primary complexity of the DS receiver, especially when synchronization must be established rapidly.

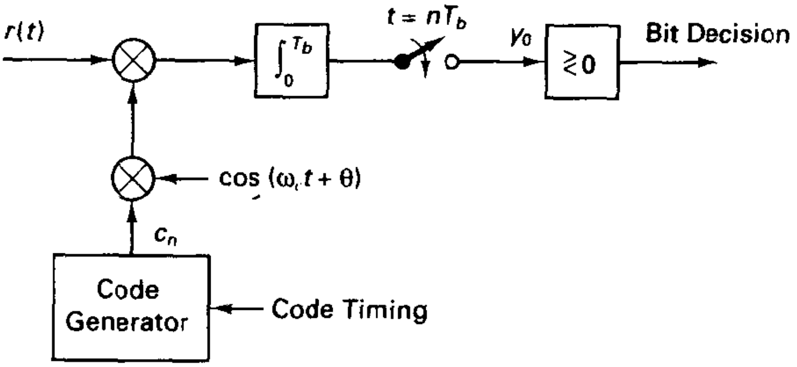

Figure 3.8.5 Direct sequence receiver in correlator form.

The error probability of the DS receiver in additive white Gaussian noise is easily found by recalling that for all antipodal detection situations

$$P_b = Q\left[\left(\frac{2E_b}{N_0}\right)^{1/2}\right]. \tag{3.8.3}$$

This result is counter to two spread-spectrum misunderstandings. First, we might expect DS spread spectrum to be less energy efficient than nonspread antipodal signaling because of the larger transmission bandwidth and hence larger noise power in the receiver. While it is true that the predetection signal-to-noise ratios may be poor in DS receivers,[31] the correlation process produces a decision SNR yielding identical error probability, expressed in terms of bit energy-to-noise power density ratio, to that of nonspread antipodal transmission. Conversely, we might expect that the large bandwidth expansion and apparent coding produces better energy efficiency, as happens in an analog FM receiver with wideband FM. This also is *not true*—again the message modulation is antipodal, and the spreading code does not provide energy efficiency gains. The reader is invited to see [57] for other discussion of other myths.

Where DS systems do shine is in providing high immunity to narrowband interferering signals or to other DS signals occupying the same frequency band. To illustrate

[31] It is not uncommon for this to be −20 dB.

this, consider the case of a sinusoidal interferer, located at the carrier frequency of the DS signal. Such a signal might either be an unintended spurious emission from an authorized transmitter or a ***tone jammer*** attempting to defeat communications of the DS users. We represent the interfering signal as $(2\gamma E_b/T_b)^{1/2}\cos(\omega_c t+\theta)$, so that γ is interpreted as the interferer's relative power level. (Notice that we give the interferer the best possible conditions by giving it the frequency and phase angle of the intended carrier.)

Analysis of the receiver in Figure 3.8.5 proceeds easily, assuming that the interfering signal does not corrupt the synchronization status of the receiver (this is perhaps the most vulnerable aspect of the receiver). By superposition, the output of the integrator in Figure 3.8.5 is the sum of the desired signal contribution, $\mu = \pm E_b^{1/2}$, a zero-mean Gaussian noise term with variance $N_0/2$, and the interference contribution, η. By realizing that this interference term is the sum of B chip-duration integration results, we find that

$$\eta = (E_b\gamma)^{1/2}\left[\frac{1}{B}\sum_{i=1}^{B} c_n\right]. \tag{3.8.4}$$

By modeling the code chips as an i.i.d. binary process, we have that the variance of the term in brackets is $1/B$, and that η has variance

$$\sigma_\eta^2 = \frac{E_b\gamma}{B}. \tag{3.8.5}$$

Furthermore by a central-limit theorem approximation, we may assume that if the band-spreading ratio B is large η is essentially Gaussian in distribution, and the error probability will be a function of the ratio of the square of the mean to total variance of error terms. After adding the variances of the additive noise and interference terms (by independence assumptions) we find that the ratio of the squared-mean to total variance is

$$\frac{\mu^2}{\sigma^2} = \frac{\mu^2}{(N_0/2)+\sigma_\eta^2} = \frac{2E_b}{N_0}\left[\frac{1}{1+(2E_b/N_0)(\gamma/B)}\right]. \tag{3.8.6}$$

Notice that with negligible thermal noise the detection SNR is

$$\frac{\mu^2}{\sigma^2} \approx \frac{\gamma}{B}, \tag{3.8.7a}$$

rather than γ obtained without spread-spectrum modulation. The effective power of the interferer is reduced by a factor of B, the band-spreading ratio, and $10\log_{10}B$ is commonly called the ***processing gain*** of a DS system. This processing gain effect holds as well for nonsinusoidal interferering signals, provided these are narrowband relative to the DS signal bandwidth.

In any case, provided the Gaussian approximation to interference holds, the error probability is

$$P_b = Q\left[\left(\frac{\mu^2}{\sigma^2}\right)^{1/2}\right], \tag{3.8.7b}$$

where the μ^2/σ^2 is given in (3.8.6).

Example 3.19 DS system with $B = 1023$

Suppose the task is to communicate binary data at a rate of 1000 bps. We select a length-1023 maximal-length code clocked at $R_c = 1.023$ MHz as a spreading code. Note that the code and data clocks can be synchronously derived and that the code sequence repeats exactly once per bit. This design provides bit timing once the code synchronization has been achieved by the receiver. Suppose the ratio of energy per bit-to-noise density is $E_b/N_0 = 10$ dB, and let a tone interferer have power 10 dB greater than the total power of the desired signal; that is, $\gamma = 10$. With a band-spreading ratio of 1023, the signal power density would be some 20 dB below the additive noise level, and with a spectrum analyzer, we would observe only a spectral line due to the interference above a background noise floor. Even though the predetection signal-to-noise ratio is quite poor, the postdetection SNR is, by (3.8.6),

$$\frac{\mu^2}{\sigma^2} = 20\left[\frac{1}{1 + 20(10/1023)}\right] \approx 16, \tag{3.8.8}$$

which is within 1 dB of the value obtained without interference. In other words, the roughly 30-dB processing gain of the spread-spectrum system has virtually negated the effect of an interfering signal 10 dB stronger than the desired signal. It should be obvious that the error probability of a nonspread system in this case is intolerably poor, although we will not analyze the specific effects of sinusoidal interferers on demodulation performance.

The SNR calculated here should be understood as a worst-case SNR, obtained with the interferer in phase with the desired carrier. If we average over a randomly chosen value for θ, the average decision quality is a factor of 2 greater, since the effective interference power is proportional to $\cos^2(\theta)$, and $E[\cos^2(\theta)] = \frac{1}{2}$.

It is illuminating to interpret the processing gain in terms of spectral bandwidths in the receiver. The interferer originates as a narrowband signal, but following multiplication by the reference in the receiver, it is converted into a wideband signal, with bandwidth proportional to R_c. (Think of the tone interferer as modulating the DS carrier in the receiver.) On the other hand, the desired DS signal at the input is "de-spread," its bandwidth shrinking to that proportional to the information rate R_b. The ratio of the postdetection bandwidths of these two signals is intuitively a measure of decision SNR, and this is exactly the processing gain defined previously.

Similar benefits accrue in cases where the interference is wideband—the interferer is left as a wideband signal, while the DS signal bandwidth collapses as before. To analyze how this sharing of the spectrum can occur through code orthogonality, suppose we have two users sending signals

$$s_1(t) = x_1(t)c_1(t)\cos(\omega_c t) \tag{3.8.9a}$$

and

$$s_2(t) = x_2(t)c_2(t)\cos(\omega_c t), \tag{3.8.9b}$$

where $x_i(t)$ is the ± 1 waveform equivalent of the message sequence $\{x_{i_n}\}$. We assume the received waveform is two signals plus noise:

$$\begin{aligned} r(t) = {} & \left(\frac{2E_s}{T_s}\right)^{1/2} x_1(t)c_1(t)\cos(\omega_c t) \\ & + \left(\frac{2E_s\gamma}{T_s}\right)^{1/2} x_2(t-\tau)c_2(t-\tau)\cos(\omega_c t + \theta) + n(t). \end{aligned} \tag{3.8.10}$$

Thus, the second signal is received at relative power level γ and delayed by some arbitrary amount τ due to propagation delay differences. Assume that we wish to recover the message sequence $\{x_{1_n}\}$. One standard approach is to employ a correlation receiver that is optimal in the no-interference case, correlating $r(t)$ with $c_1(t)\cos(\omega_c t)$ (this requires the usual carrier and code synchronization). At the output of the integrate-and-dump detector, the decision statistic for the nth data bit is, assuming that $x_1(t) = 1$,

$$z_n = E_s^{1/2} + (E_s\gamma)^{1/2}\cos(\theta)\int_{(n-1)T_s}^{nT_s} c_1(t)c_2(t-\tau)x_2(t-\tau)\,dt + n_n, \tag{3.8.11}$$

where the second term is the result of user-2 interference and n_n is the result of additive Gaussian noise, known as before to have zero-mean and variance $N_0/2$.

Notice that the interference term is a result of several factors: the relative carrier phase angle θ, the polarity of the message $x_2(t-\tau)$ over the integration interval, and most importantly the cross-correlation properties of the two code waveforms (or sequences). In the special case where the transmissions are synchronized so that $\tau = 0$ and where $c_1(t)$ and $c_2(t)$ repeat every message bit *and* are strictly orthogonal, the interference term in (3.8.11) is zero, irrespective of the message $x_2(t)$ or θ. Thus, in purely synchronous CDMA using DS spreading, many multiple users can share the channel spectrum without mutual interference, provided the sequences are mutually orthogonal. Finding large sets of orthogonal binary code sequences is not difficult; for example, rows of Hadamard matrices will suffice, although these are not so well modeled as random binary sequences for spectral purposes.

The practical situation, however, is that such synchronization can usually not be arranged, especially to chip-duration accuracy. (An interesting exception is in cellular CDMA networks wherein outbound links from the cell sites to remote terminals carry simultaneous messages to many users, and each user receives multiple synchronized messages.)

As soon as asynchronism enters the picture, performance analysis becomes more complicated, as does the code design problem. For a specific pair of codes, (3.8.11) gives the means to calculate performance of error probability, but there are many cases to consider, such as various relative delays τ, as well as whether $x_2(t-\tau)$ switches polarity in the middle of the integration interval. A code design problem appears then to be to find large sets of codes for which the worst-case, pairwise interference is minimized. There is no closed-form analytic solution of this problem, and designers normally resort to families of codes with good cross-correlation properties under asynchronism. The Gold sequences mentioned earlier are a prevalent choice, but quadriphase sequences offer some advantages in minimizing worst-case interference.

The interference rejection properties of DS spread spectrum can be assessed in the CDMA case by treating the integral in the second term of (3.8.11) as a Gaussian random variable[32] (this is more justifiable if we consider τ as a variable), with zero mean and variance $C(\tau)$, where $C(\tau)$ is the normalized cross-correlation at lag τ of the two sequences. (This cross-correlation should be near zero.) The interference then appears as a noise like perturbation to the decision statistic, where the interference noise has variance

$$\sigma_I^2 = E_s\gamma C(\tau). \tag{3.8.12}$$

[32] References [49] and [50] studied the Gaussian approximation and observed that it is optimistic.

Thus, $1/C(\tau)$ plays the role of processing gain and, in fact, under the random coin-flipping model for sequences, would become B, the band-spreading ratio. Processing gain is the amount of effective reduction of interference power by virtue of near orthogonality of codes. Notice that interferers 20 dB stronger than the desired signal can be tolerated if the processing gain is, say 30 dB. Whenever nonperfect orthogonality exists, however, the ***near–far*** problem eventually limits system performance. That is, even with a high degree of mutual orthogonality for all situations, if γ is large due to the nondesired signal's transmitter being much nearer than that of the desired signal, error probability eventually becomes unacceptably high. (Exercise 3.8.4 treats some numerical examples.)

Finally, we might ask about effects of interference from multiple simultaneous users. Superposition applies in the receiver analysis, and by usual statistical methods, the aggregate interference can be treated as Gaussian (the limit theorem becomes even more germane in the multiuser case), with variance obtained by adding the contributions of each user. Here, again. a sufficiently large number of weakly correlated interfering signals can degrade system performance.

3.8.2 Frequency-hopping Spread Spectrum

In FH systems, the digital modulation is performed as described earlier in this chapter, but the carrier frequency hops among frequency slots at a rate called the ***frequency-hopping rate***. The carrier frequency is established by a frequency synthesizer, in turn driven by a pseudorandom sequence generator, as shown in Figure 3.8.6. The carrier frequency is selected from a set of N possible frequencies, equally spaced by some amount Δf over a total ***hopping range*** of $W = N\Delta f$ hertz. If the length of the pseudorandom sequence is $2^L - 1$, then there are $N = 2^L - 1$ unique input vectors to the synthesizer.[33]

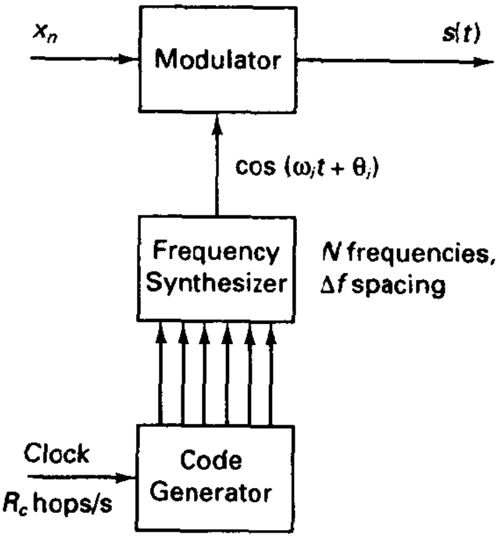

Figure 3.8.6 Frequency-hopping modulator.

[33] Although not indicated in Figure 3.8.6, the actual hopping of frequencies is often performed not at the modulator, but in some frequency up-conversion process.

In FH spread spectrum, both the hopping rate and the hopping range are important design parameters. (In DS spread spectrum on the other hand, the chip rate R_c is the only design parameter.) ***Slow hopping*** designates systems wherein the hop rate is slow relative to the message rate, so in effect the carrier dwells at any given frequency for many successive symbol durations. In such cases it is reasonable to contemplate coherent detection, or at least differentially coherent detection if DPSK modulation is selected. ***Fast hopping***, in contrast, connotes operation for which the carrier frequency is changed multiple times per modulator symbol. This may be desirable to prevent a hostile interferer from listening to the hopping pattern and simply frequency following. Also, if the channel were a frequency-selective fading channel, fast hopping can provide a means of mitigating the harmful effects of fading on any one frequency, yielding a frequency diversity benefit. Time/frequency signal occupancy patterns are illustrated for both cases in Figure 3.8.7.

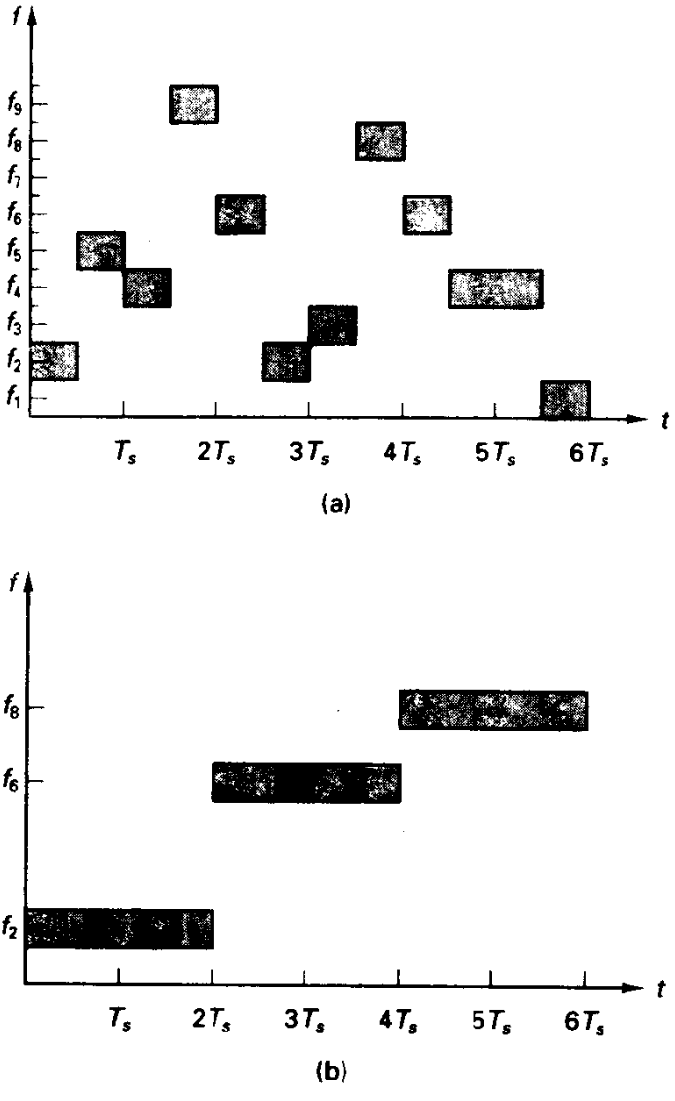

Figure 3.8.7 Carrier frequency patterns for frequency hopping. Modulation occurs relative to indicated frequency. (a) Two hops per symbol; (b) two symbols per hop.

For all but very slow hopping cases, detection is normally performed noncoherently because of the difficulty in maintaining an accurate carrier phase reference under changing frequency conditions. Thus, we typically encounter DPSK and M-ary orthogonal, for example, MFSK, modulation. If hopping is slow, then our previous theory of Sections 3.4 and 3.5 provides the receiver structure and the receiver performance, at least in the AWGN environment. We merely realize that the carrier will be hopping at some prescribed rate, according to a known pattern, and, presuming a synchronized code generator, the frequency in the receiver can be synthesized to properly compensate for the transmitter frequency, restoring the problem to one of nonspread communications. (If DPSK transmission is employed, an extra start-up symbol is necessary for each hop.)

For *slow-hopping* FH systems operating in a pure AWGN environment, the error probabilities are exactly those given earlier for the various modulation and detection cases. Here, again, FH systems neither gain nor lose in energy efficiency, and in fact the performance does not depend on hopping range or hopping rate, provided each symbol is contained within one hop.

Example 3.20 DPSK with FH Spread Spectrum

Consider transmission of digitally encoded speech, producing a bit rate of 2400 bps.[34] Suppose that we elect binary DPSK modulation for its relatively good energy efficiency and adopt a hopping pattern of one hop every four information bits. We must actually signal at a rate of 3000 bps to accommodate the overhead symbol in each hop. The hopping rate is 600 hops per second, rather leisurely with today's frequency-synthesizer technology.

Suppose that $P_r T_b/N_0 = E_b/N_0 = 10$ dB at the receiver. Realizing that a fifth symbol must be added at the beginning of each hop to act as a DPSK phase reference, we find that the effective E_b/N_0 is about 1 dB less, or 9 dB. Evaluating the DPSK error probability expression, we find

$$P_b = \frac{1}{2}e^{-7.94} = 1.8 \cdot 10^{-4}. \tag{3.8.13}$$

This is considered quite acceptable for most digital speech encoders, due to natural redundancy in speech and the ability of the auditory system to tolerate errors.

The hop interval and hopping range depend on other system considerations, particularly other interference scenarios discussed later. We might wish to maintain orthogonality between transmissions of other users who are randomly hopping in the same band. This would require that the hop spacing Δf be a multiple of 3000 Hz.[35] Choice of the minimum spacing and use of 255 slots would consume a bandwidth of about 0.75 MHz.

Whereas slow hopping does not induce any energy penalty (or gain) on the AWGN channel relative to nonspread modulation with the same basic signal set and detection strategy, fast-hopping noncoherent systems do suffer an energy penalty because of inability to coherently integrate, or combine, the data from the multiple hops involved in a symbol decision. To analyze this effect, we consider the case of binary FSK modulation with noncoherent detection when $H > 1$ hops per bit are used. The FH receiver is shown in Figure 3.8.8, where over each hop interval we form statistics y_{0_i} and y_{1_i} in each of the two channels, exactly as for nonhopped communication. The optimal manner

[34] This corresponds to U. S. military standard LPC-10.

[35] Orthogonality may occur only if other users are hop synchronous.

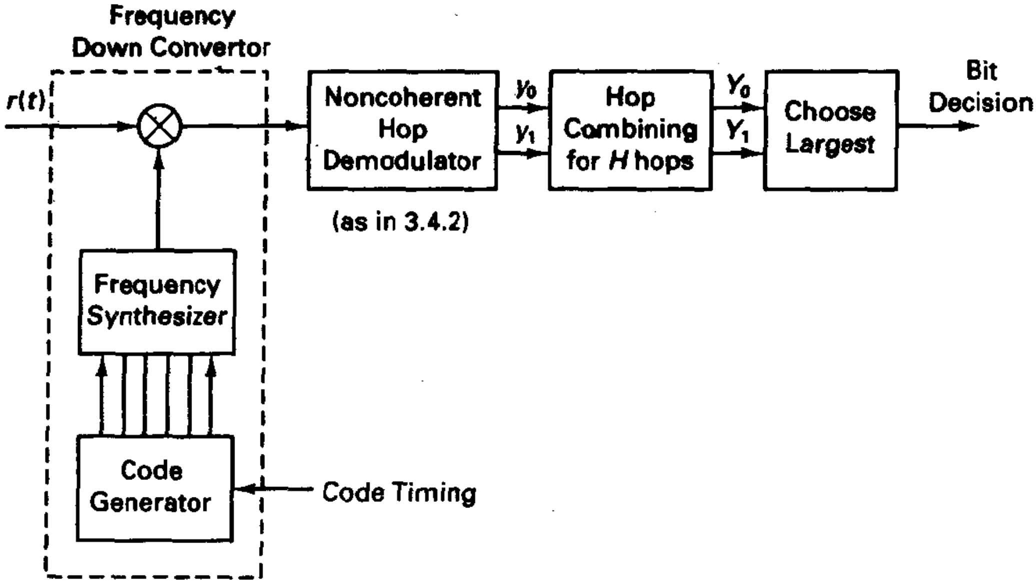

Figure 3.8.8 Frequency-hopping demodulator, binary modulation.

of combining a sequence of H such measurements follows from likelihood calculations and implies that we form the two statistics

$$Y_0 = \sum_{i=1}^{H} \log_e I_0\left(\frac{\mu y_{0_i}}{\sigma^2}\right), \qquad Y_1 = \sum_{i=1}^{H} \log_e I_0\left(\frac{\mu y_{1_i}}{\sigma^2}\right) \tag{3.8.14}$$

and decide in favor of the largest. In (3.8.14), $\mu = (E_b/H)^{1/2}$, the energy per hop, and $\sigma^2 = N_0/2$. Exercise 3.8.5 develops these results.

Partly for analytical convenience and partly for ease of implementation, we use the approximation $\log_e I_0(x) \approx x^2$, which is most accurate at low values of the argument x, or for small SNR. Assuming that μ and σ^2 do not change from hop to hop, we take as our decision variables

$$Z_0 = \sum_{i=1}^{H} y_{0_i}^2, \qquad Z_1 = \sum_{i=1}^{H} y_{1_i}^2 \tag{3.8.15}$$

This receiver combining policy is often referred to as ***square-law combining***. On the other hand, with large SNR, addition of the measurements directly *without squaring* provides a better approximation to the optimal statistics in (3.8.14). It is naturally wiser to have a good approximation for the small SNR regime; when SNR is high, the error probability is small anyway despite suboptimality.

Following our earlier analysis, the hop random variables Y_{0_i} and Y_{1_i} are Rician and Rayleigh distributed, respectively, assuming transmission of the 0 symbol, and independent. The square of a Rayleigh random variable has a chi-squared distribution with two degrees of freedom, as earlier described in Chapter 2. Furthermore, the sum of squares of H such Rayleigh variables has a chi-squared distribution with $2H$ degrees of freedom:

$$f_{Z_1|S_0}(z_1|S_0) = \frac{1}{\sigma^{2H}2^H\Gamma(H)} z_1^{H-1} e^{-z_1/2\sigma^2}, \qquad z_1 \geq 0. \tag{3.8.16}$$

where $\Gamma(H) = (H-1)!$ is the gamma function.

The sum of squares of Rician variates is not so simple to express, but has a noncentral chi-squared distribution. Proakis [31] provides a detailed derivation. The p.d.f. for Z_0, conditioned upon message 0 transmission, is

$$f_{Z_0|S_0}(z_0|S_0) = \frac{1}{2\sigma^2}\left(\frac{z_0}{s^2}\right)^{(H-1)/2} e^{-(s^2+z_0)/2\sigma^2} I_{H-1}\left(\frac{s z_0^{1/2}}{\sigma^2}\right), \qquad z_0 \geq 0, \tag{3.8.17a}$$

where

$$s^2 = \sum_{i=1}^{H} \mu_i^2 = \sum_{i=1}^{H} \frac{E_b}{H} = E_b \tag{3.8.17b}$$

is the ***noncentrality parameter***, and $I_P(x)$ is the modified Bessel function of the first kind with order P.

The probability of error is then the probability that Z_1 exceeds Z_0, which may be put into integral form by invoking (3.8.16), (3.8.17), and the independence of the two statistics. We shall not plot the resulting error probability, but instead display the loss in performance, relative to the case when $H = 1$ (or, more generally, when there are many symbols per hop). In Figure 3.8.9 we show the ***noncoherent combining loss*** as a function of H for differing values of E_b/N_0. The plot of Figure 3.8.9 allows us to construct the error probability plot for fast-hopping binary FSK by applying the indicated corrections to the curve for slow-hopping or standard nonhopped binary FSK with noncoherent detection. Recall that the latter has error probability given by

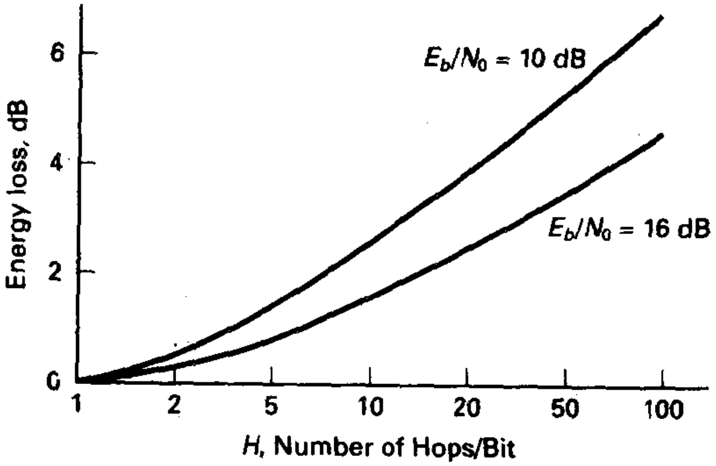

Figure 3.8.9 Noncoherent combining loss for fast frequency hopping, AWGN channel, binary orthogonal signals.