$$z(k) = \begin{cases} 1 & \text{for } k = 0 \\ 0 & \text{for } k = \pm 1, \pm 2, \ldots, \pm N \end{cases} \qquad (3.90)$$

Equation (3.89) is used to solve the $2N + 1$ simultaneous equations for the set of $2N + 1$ weights $\{c_n\}$. The required length of the filter (number of tap weights) is a function of how much smearing the channel may introduce. For such an equalizer with finite length, the peak distortion is guaranteed to be minimized only if the eye pattern is initially open. However, for high-speed transmission and channels introducing much ISI, the eye is often closed before equalization [8]. Since the zero-forcing equalizer neglects the effect of noise, it is not always the best system solution.

Example 3.5 A Zero-Forcing Three-Tap Equalizer

Consider that the tap weights of an equalizing transversal filter are to be determined by transmitting a single impulse as a training signal. Let the equalizer circuit in Figure 3.26 be made up of just three taps. Given a received distorted set of pulse samples $\{x(k)\}$, with voltage values 0.0, 0.2, 0.9, –0.3, 0.1, as shown in Figure 3.25, use a zero-forcing solution to find the weights $\{c_{-1}, c_0, c_1\}$ that reduce the ISI so that the equalized pulse samples $\{z(k)\}$ have the values, $\{z(-1) = 0, z(0) = 1, z(1) = 0\}$. Using these weights, calculate the ISI values of the equalized pulse at the sample times $k = \pm 2, \pm 3$. What is the largest magnitude sample contributing to ISI, and what is the sum of all the ISI magnitudes?

Solution

For the channel impulse response specified, Equation (3.89) yields

$$\mathbf{z} = \mathbf{x}\,\mathbf{c}$$

or

$$\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} x(0) & x(-1) & x(-2) \\ x(1) & x(0) & x(-1) \\ x(2) & x(1) & x(0) \end{bmatrix} \begin{bmatrix} c_{-1} \\ c_0 \\ c_1 \end{bmatrix}$$

$$= \begin{bmatrix} 0.9 & 0.2 & 0 \\ -0.3 & 0.9 & 0.2 \\ 0.1 & -0.3 & 0.9 \end{bmatrix} \begin{bmatrix} c_{-1} \\ c_0 \\ c_1 \end{bmatrix}$$

Solving these three simultaneous equations results in the following weights:

$$\begin{bmatrix} c_{-1} \\ c_0 \\ c_1 \end{bmatrix} = \begin{bmatrix} -0.2140 \\ 0.9631 \\ 0.3448 \end{bmatrix}$$

DELL-1022, Part 2
9,584,262

The values of the equalized pulse samples $\{z(k)\}$ corresponding to sample times $k = -3, -2, -1, 0, 1, 2, 3$ are computed by using the preceding weights in Equation (3.89a), yielding

$$0.0000,\ -0.0428,\ 0.0000,\ 1.0000,\ 0.0000,\ -0.0071,\ 0.0345$$

The sample of greatest magnitude contributing to ISI equals 0.0428, and the sum of all the ISI magnitudes equals 0.0844. It should be clear that this three-tap equalizer has forced the sample points on either side of the equalized pulse to be zero. If the equalizer is made longer than three taps, more of the equalized sample points can be forced to a zero value.

Minimum MSE Solution. A more robust equalizer is obtained if the $\{c_n\}$ tap weights are chosen to minimize the mean-square error (MSE) of all the ISI terms plus the noise power at the output of the equalizer [9]. MSE is defined as the expected value of the squared difference between the desired data symbol and the estimated data symbol. One can use the set of overdetermined equations to obtain a minimum MSE solution by multiplying both sides of Equation (3.89a) by $\mathbf{x}^T$, which yields [10]

$$\mathbf{x}^T\mathbf{z} = \mathbf{x}^T\mathbf{x}\mathbf{c} \tag{3.91a}$$

and

$$\mathbf{R}_{xz} = \mathbf{R}_{xx}\mathbf{c} \tag{3.91b}$$

where $\mathbf{R}_{xz} = \mathbf{x}^T\mathbf{z}$ is called the *cross-correlation* vector and $\mathbf{R}_{xx} = \mathbf{x}^T\mathbf{x}$ is called the *autocorrelation* matrix of the input noisy signal. In practice, $\mathbf{R}_{xz}$ and $\mathbf{R}_{xx}$ are unknown *a priori,* but can be approximated by transmitting a test signal over the channel and using time average estimates to solve for the tap weights from Equation (3.91), as follows:

$$\mathbf{c} = \mathbf{R}_{xx}^{-1}\mathbf{R}_{xz} \tag{3.92}$$

In the case of the deterministic zero-forcing solution, the $\mathbf{x}$ matrix must be square. But to achieve the minimum MSE (statistical) solution, one starts with an overdetermined set of equations and hence a *nonsquare* $\mathbf{x}$ matrix, which then gets transformed to a *square* autocorrelation matrix $\mathbf{R}_{xx} = \mathbf{x}^T\mathbf{x}$, yielding a set of $2N + 1$ simultaneous equations, whose solution leads to tap weights that minimize the MSE. The size of the vector $\mathbf{c}$ and the number of columns of the matrix $\mathbf{x}$ correspond to the number of taps in the equalizing filter. Most high-speed telephone-line modems use an MSE weight criterion because it is superior to a zero-forcing criterion; it is more robust in the presence of noise and large ISI [8].

Example 3.6 A Minimum MSE 7-Tap Equalizer

Consider that the tap weights of an equalizing transversal filter are to be determined by transmitting a single impulse as a training signal. Let the equalizer circuit in Figure 3.26 be made up of seven taps. Given a received distorted set of pulse samples $\{x(k)\}$, with values 0.0108, −0.0558, 0.1617, 1.0000, −0.1749, 0.0227, 0.0110, use a minimum MSE solution to find the value of the weights $\{c_n\}$ that will minimize the ISI. With these weights, calculate the resulting values of the equalized pulse samples at the fol-

lowing times: $\{k = 0, \pm1, \pm2, \ldots, \pm6\}$. What is the largest magnitude sample contributing to ISI, and what is the sum of all the ISI magnitudes?

Solution

For a seven-tap filter ($N = 3$), one can form the **x** matrix in Equation (3.88) that has dimensions $4N + 1$ by $2N + 1 = 13 \times 7$:

$$\mathbf{x} = \begin{bmatrix} 0.0110 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.0227 & 0.0110 & 0 & 0 & 0 & 0 & 0 \\ -0.1749 & 0.0227 & 0.0110 & 0 & 0 & 0 & 0 \\ 1.0000 & -0.1749 & 0.0227 & 0.0110 & 0 & 0 & 0 \\ 0.1617 & 1.0000 & -0.1749 & 0.0227 & 0.0110 & 0 & 0 \\ -0.0558 & 0.1617 & 1.0000 & -0.1749 & 0.0227 & 0.0110 & 0 \\ 0.0108 & -0.0558 & 0.1617 & 1.0000 & -0.1749 & 0.0227 & 0.0110 \\ 0 & 0.0108 & -0.0558 & 0.1617 & 1.0000 & -0.1749 & 0.0227 \\ 0 & 0 & 0.0108 & -0.0558 & 0.1617 & 1.0000 & -0.1749 \\ 0 & 0 & 0 & 0.0108 & -0.0558 & 0.1617 & 1.0000 \\ 0 & 0 & 0 & 0 & 0.0108 & -0.0558 & 0.1617 \\ 0 & 0 & 0 & 0 & 0 & 0.0108 & -0.0558 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.0108 \end{bmatrix}$$

Using this **x** matrix, one can form the autocorrelation matrix $\mathbf{R}_{xx}$ and the cross-correlation vector $\mathbf{R}_{xz}$, defined in Equation (3.91). With the help of a computer to invert $\mathbf{R}_{xx}$ and perform matrix multiplication, the solution for the tap weights $\{c_{-3}, c_{-2}, c_{-1}, c_0, c_1, c_2, c_3\}$ shown in Equation (3.92) yields

$$-0.0116,\ 0.0108,\ 0.1659,\ 0.9495,\ -0.1318,\ 0.0670,\ -0.0269$$

Using these weights in Equation (3.89a), we solve for the 13 equalized samples $\{z(k)\}$ at times $k = -6, -5, \ldots, 5, 6$:

$$-0.0001,\ -0.0001,\ \ 0.0041,\ 0.0007,\ \ 0.0000,\ -0.0000,\ 1.0000,\ 0.0003,$$
$$-0.0007,\ \ 0.0015,\ -0.0095,\ 0.0022,\ -0.0003$$

The largest magnitude sample contributing to ISI equals 0.0095, and the sum of all the ISI magnitudes equals 0.0195.

3.4.3.2 Decision Feedback Equalizer

The basic limitation of a linear equalizer, such as the transversal filter, is that it performs poorly on channels having spectral nulls [11]. Such channels are often encountered in mobile radio applications. A decision feedback equalizer (DFE) is a nonlinear equalizer that uses previous detector decisions to eliminate the ISI on pulses that are currently being demodulated. The ISI being removed was caused by the tails of previous pulses; in effect, the distortion on a current pulse that was caused by previous pulses is subtracted.

Figure 3.27 shows a simplified block diagram of a DFE where the forward filter and the feedback filter can each be a linear filter, such as a transversal filter. The figure also illustrates how the filter tap weights are updated adaptively. (See the following section.) The nonlinearity of the DFE stems from the nonlinear

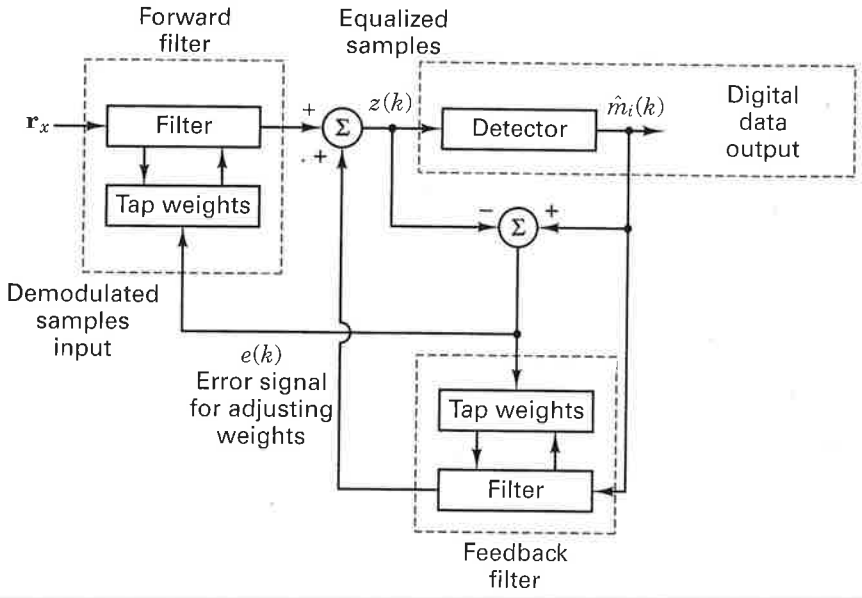


Figure 3.27 Decision Feedback Equalizer.

characteristic of the detector that provides an input to the feedback filter. The basic idea of a DFE is that if the values of the symbols previously detected are known (past decisions are assumed to be correct), then the ISI contributed by these symbols can be canceled out exactly at the output of the forward filter by subtracting past symbol values with appropriate weighting. The forward and feedback tap weights can be adjusted simultaneously to fulfill a criterion such as minimizing the MSE.

When only a forward filter is used, the output of the filter contains channel noise contributed from every sample in the filter. The advantage of a DFE implementation is that the feedback filter, which is additionally working to remove ISI, operates on noiseless quantized levels, and thus its output is free of channel noise.

3.4.4 Preset and Adaptive Equalization

On channels whose frequency responses are known and time invariant, the channel characteristics can be measured and the filter's tap weights adjusted accordingly. If the weights remain fixed during transmission of data, the equalization is called *preset* equalization; one very simple method of preset equalization consists of setting the tap weights $\{c_n\}$ according to some average knowledge of the channel. This was used for data transmission over voice-grade telephone lines at less than 2400 bit/s. Another preset method consists of transmitting a training sequence that is compared at the receiver with a locally generated sequence. The differences between the two sequences are used to set $\{c_n\}$. The significant aspect of any preset method

is that it is done once at the start of transmission or seldom (when transmission is broken and needs to be reestablished).

Another type of equalization, capable of tracking a slowly time-varying channel response, is known as *adaptive* equalization. It can be implemented to perform tap-weight adjustments periodically or continually. Periodic adjustments are accomplished by periodically transmitting a preamble or short training sequence of digital data that is known in advance by the receiver. The receiver also uses the preamble to detect start of transmission, to set the automatic gain control (AGC) level, and to align internal clocks and local oscillator with the received signal. Continual adjustments are accomplished by replacing the known training sequence with a sequence of data symbols estimated from the equalizer output and treated as known data. When performed continually and automatically in this way, the adaptive procedure (the most popular) is referred to as *decision directed* [11]. The name "decision directed" is not to be confused with decision feedback (DFE). Decision directed only addresses how filter tap weights are adjusted—that is, with the help of a signal from the detector. DFE, however, refers to the fact that there exists an additional filter that operates on the detector output and recursively feeds back a signal to the detector input. Thus, with DFE there are two filters, a feed-forward filter and a feedback filter that process the data and help mitigate the ISI.

A disadvantage of preset equalization is that it requires an initial training period that must be invoked at the start of any new transmission. Also, a time-varying channel can degrade system performance due to ISI, since the tap weights are fixed. Adaptive equalization, particularly *decision-directed* adaptive equalization, successfully cancels ISI when the initial probability of error exceeds one percent, (rule of thumb). If the probability of error exceeds one percent, the decision directed equalizer might not converge. A common solution to this problem is to initialize the equalizer with an alternate process, such as a preamble to provide good channel-error performance, and then switch to the decision-directed mode. To avoid the overhead represented by a preamble, many systems designed to operate in a continuous broadcast mode use *blind equalization* algorithms to form initial channel estimates. These algorithms adjust filter coefficients in response to sample statistics rather than in response to sample decisions [11].

Automatic equalizers use iterative techniques to estimate the optimum coefficients. The simultaneous equations described in Equation (3.89) do not include the affects of channel noise. To obtain a stable solution to the filter weights, it is necessary that the data be averaged to obtain stable signal statistics, or the noisy solutions obtained from the noisy data must be averaged. Considerations of algorithm complexity and numerical stability most often lead to algorithms that average noisy solutions. The most robust of this class of algorithm is the least-mean-square (LMS) algorithm. Each iteration of this algorithm uses a noisy estimate of the error *gradient* to adjust the weights in the direction to reduce the average mean-square error. The noisy gradient is simply the product $e(k)\ \mathbf{r}_x$ of an error scalar $e(k)$ and the data vector $\mathbf{r}_x$. The vector $\mathbf{r}_x$ is the vector of noise-corrupted channel samples residing in the equalizer filter at time k. Earlier, an impulse was transmitted and the equalizing filter operated on a sequence of samples (a vector) that represented

the impulse response of the channel. We displayed these received samples (in time-shifted fashion) as the matrix **x.** Now, rather than dealing with the response to an impulse, consider that data is sent and thus the vector of received samples $\mathbf{r}_x$ at the input to the filter (Figure 3.27) represents the data response of the channel. The error is formed as the difference between the desired output signal and the filter output signal and is given by

$$e(k) = z(k) - \hat{z}(k) \tag{3.93}$$

where $z(k)$ is the desired output signal (a sample free of ISI) and $\hat{z}(k)$ is an estimate of $z(k)$ at time k out of the filter (into the quantizer of Figure 3.27), which is obtained as follows:

$$\hat{z}(k) = \mathbf{c}^T\mathbf{r}_x = \sum_{n=-N}^{N} x(k-n)c_n \tag{3.94}$$

In Equation (3.94), the summation represents a convolution of the input data samples with the $\{c_n\}$ tap weights, where c_n refers to the nth tap weight at time k, and $\mathbf{c}^T$ is the transpose of the weight vector at time k. We next show the iterative process that updates the set of weights at each time k as

$$\mathbf{c}(k+1) = \mathbf{c}(k) + \Delta e(k)\mathbf{r}_x \tag{3.95}$$

where $\mathbf{c}(k)$ is the vector of filter weights at time k, and Δ is a small term that limits the coefficient step size and thus controls the rate of convergence of the algorithm as well as the variance of the steady state solution. This simple relationship is a consequence of the orthogonality principle that states that the error formed by an optimal solution is orthogonal to the processed data. Since this is a recursive algorithm (in the weights), care must be exercised to assure algorithm stability. Stability is assured if the parameter Δ is smaller than the reciprocal of the energy of the data in the filter. When stable, this algorithm converges in the mean to the optimal solution but exhibits a variance proportional to the parameter Δ. Thus, while we want the convergence parameter Δ to be large for fast convergence but not so large as to be unstable, we also want it to be small enough for low variance. The parameter Δ is usually set to a fixed small amount [12] to obtain a low-variance steady-state tap-weight solution. Schemes exist that permit Δ to change from large values during initial acquisition to small values for stable steady-state solutions [13].

Note that Equations (3.93) through (3.95) are shown in the context of real signals. When the receiver is implemented in quadrature fashion, such that the signals appear as real and imaginary (or inphase and quadrature) ordered pairs, then each line in Figure 3.27 actually consists of two lines, and Equations (3.93) through (3.95) need to be expressed with complex notation. (Such quadrature implementation is treated in greater detail in Sections 4.2.1 and 4.6.)

3.4.5 Filter Update Rate

Equalizer filters are classified by the rate at which the input signal is sampled. A transversal filter with taps spaced T seconds apart, where T is the symbol time, is called a *symbol-spaced* equalizer. The process of sampling the equalizer output at a rate $1/T$ causes aliasing if the signal is not strictly bandlimited to $1/T$ hertz—that is, the signal's spectral components spaced $1/T$ hertz apart are folded over and superimposed. The aliased version of the signal may exhibit spectral nulls [8]. A filter update rate that is greater than the symbol rate helps to mitigate this difficulty. Equalizers using this technique are called *fractionally-spaced* equalizers. With a fractionally spaced equalizer, the filter taps are spaced at

$$T' \leq \frac{T}{(1 + r)} \tag{3.96}$$

seconds apart, where r denotes the excess bandwidth. In other words, the received signal bandwidth is

$$W \leq \frac{(1 + r)}{T} \tag{3.97}$$

The goal is to choose T' so that the equalizer transfer function $H_e(f)$ becomes sufficiently broad to accommodate the whole signal spectrum. Note that the signal at the output of the equalizer is still sampled at a rate $1/T$, but since the tap weights are spaced T' seconds apart (the equalizer input signal is sampled at a rate $1/T'$), the equalization action operates on the received signal before its frequency components are aliased. Equalizer simulations over voice-grade telephone lines, with $T' = T/2$, confirm that such fractionally-spaced equalizers outperform symbol-spaced equalizers [14].

3.5 CONCLUSION

In this chapter, we described the detection of binary signals plus Gaussian noise in terms of two basic steps. In the first step the received waveform is reduced to a single number $z(T)$, and in the second step a decision is made as to which signal was transmitted, on the basis of comparing $z(T)$ to a threshold. We discussed how to best choose this threshold. We also showed that a linear filter known as a matched filter or correlator is the optimum choice for maximizing the output signal-to-noise ratio and thus minimizing the probability of error.

We defined intersymbol interference (ISI) and explained the importance of Nyquist's work in establishing a theoretical minimum bandwidth for symbol detection without ISI. We partitioned error-performance degradation into two main types. The first is a simple loss in signal-to-noise ratio. The second, resulting from distortion, is a bottoming-out of the error probability versus the E_b/N_0 curve.

Finally, we described equalization techniques that can be used to mitigate the effects of ISI.

REFERENCES

1. Nyquist, H., "Thermal Agitation of Electric Charge in Conductors," *Phys. Rev.*, vol. 32, July 1928, pp. 110–113.
2. Van Trees, H. L., *Detection, Estimation, and Modulation Theory,* Part 1, John Wiley & Sons, Inc., New York, 1968.
3. Arthurs, E., and Dym, H., "On the Optimum Detection of Digital Signals in the Presence of White Gaussian Noise—A Geometric Interpretation of Three Basic Data Transmission Systems," *IRE Trans. Commun. Syst.*, December 1962.
4. Wozencraft, J. M. and Jacobs, I. M., *Principles of Communication Engineering,* John Wiley & Sons, Inc., New York, 1965.
5. Borjesson, P. O., and Sundberg, C. E., "Simple Approximations of the Error Function $Q(x)$ for Communications Applications," *IEEE Trans. Commun.*, vol. COM27, Mar. 1979, pp. 639–642.
6. Nyquist, H., "Certain Topics of Telegraph Transmission Theory," *Trans. Am. Inst. Electr. Eng.*, vol. 47, Apr. 1928, pp. 617–644.
7. Hanzo, L. and Stefanov, J., "The Pan-European Digital Cellular Mobile Radio System—Known as GSM," *Mobile Radio Communications,* edited by R. Steele, Chapter 8, Pentech Press, London, 1992.
8. Qureshi, S. U. H., "Adaptive Equalization," *Proc. IEEE,* vol. 73, no. 9, September 1985, pp. 1340–1387.
9. Lucky, R. W., Salz, J., and Weldon, E. J., Jr., *Principles of Data Communications,* McGraw Hill Book Co., New York, 1968.
10. Harris, F., and Adams, B., "Digital Signal Processing to Equalize the Pulse Response of Non Synchronous Systems Such as Encountered in Sonar and Radar," *Proc. of the Twenty-Fourth Annual ASILOMAR Conference on Signals, Systems, and Computers,* Pacific Grove, California, November 5–7, 1990.
11. Proakis, J. G., *Digital Communications,* McGraw-Hill Book Company, New York, 1983.
12. Feuer, A., and Weinstein, E., "Convergence Analysis of LMS Filters with Uncorrelated Gaussian Data, "*IEEE Trans. on ASSP,* vol. V-33 pp. 220–230, 1985.
13. Macchi, O., *Adaptive Processing: Least Mean Square Approach With Applications in Transmission,* John Wiley & Sons, New York, 1995.
14. Benedetto, S., Biglieri, E., and Castellani, V., *Digital Transmission Theory,* Prentice Hall, 1987.

PROBLEMS

3.1. Determine whether or not $s_1(t)$ and $s_2(t)$ are orthogonal over the interval $(-1.5T_2 < t < 1.5T_2)$, where $s_1(t) = \cos(2\pi f_1 t + \phi_1)$, $s_2(t) = \cos(2\pi f_2 t + \phi_2)$, and $f_2 = 1/T_2$ for the following cases.

(a) $f_1 = f_2$ and $\phi_1 = \phi_2$

(b) $f_1 = \frac{1}{3} f_2$ and $\phi_1 = \phi_2$

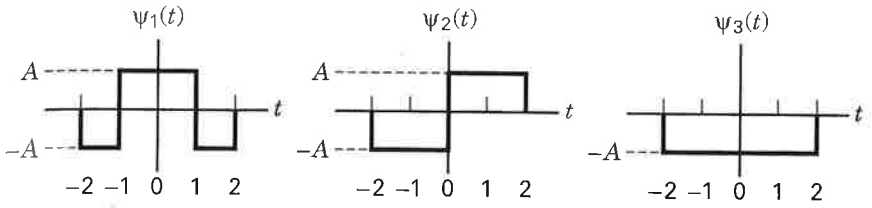


Figure P3.1

(c) $f_1 = 2f_2$ and $\phi_1 = \phi_2$
(d) $f_1 = \pi f_2$ and $\phi_1 = \phi_2$
(e) $f_1 = f_2$ and $\phi_1 = \phi_2 + \pi/2$
(f) $f_1 = f_2$ and $\phi_1 = \phi_2 + \pi$

3.2. **(a)** Show that the three functions illustrated in Figure P3.1 are pairwise orthogonal over the interval (–2, 2).
(b) Determine the value of the constant, A, that makes the set of functions in part (a) an orthonormal set.
(c) Express the following waveform, $x(t)$, in terms of the orthonormal set of part (b).

$$x(t) = \begin{cases} 1 & \text{for } 0 \le t \le 2 \\ 0 & \text{otherwise} \end{cases}$$

3.3. Consider the functions

$$\psi_1(t) = \exp(-|t|) \quad \text{and} \quad \psi_2 = 1 - A\exp(-2|t|)$$

Determine the constant, A, such that $\psi_1(t)$ and $\psi_2(t)$ are orthogonal over the interval $(-\infty, \infty)$.

3.4. Assume that in a binary digital communication system, the signal component out of the correlator receiver is $a_i(T) = +1$ or -1 V with equal probability. If the Gaussian noise at the correlator output has unit variance, find the probability of a bit error.

3.5. A bipolar binary signal, $s_i(t)$, is a +1- or –1-V pulse during the interval $(0, T)$. Additive white Gaussian noise having two-sided power spectral density of 10^{-3} W/Hz is added to the signal. If the received signal is detected with a matched filter, determine the maximum bit rate that can be sent with a bit error probability of $P_B \le 10^{-3}$.

3.6. Bipolar pulse signals, $s_i(t)$ $(i = 1, 2)$, of amplitude ±1 V are received in the presence of AWGN that has a variance of 0.1 V^2. Determine the optimum (minimum probability of error) detection threshold, γ_0, for matched filter detection if the a priori probabilities are: **(a)** $P(s_1) = 0.5$; **(b)** $P(s_1) = 0.7$; **(c)** $P(s_1) = 0.2$. **(d)** Explain the effect of the a priori probabilities on the value of γ_0. [*Hint:* Refer to Equations (B.10) to (B.12).]

3.7. A binary communication system transmits signals $s_i(t)$ $(i = 1, 2)$. The receiver test statistic $z(T) = a_i + n_0$, where the signal component a_i is either $a_1 = +1$ or $a_2 = -1$ and the noise component n_0 is uniformly distributed, yielding the conditional density functions $p(z|s_i)$ given by

$$p(z|s_1) = \begin{cases} \frac{1}{2} & \text{for } -0.2 \le z \le 1.8 \\ 0 & \text{otherwise} \end{cases}$$

and

$$p(z|s_2) = \begin{cases} \frac{1}{2} & \text{for } -1.8 \le z \le 0.2 \\ 0 & \text{otherwise} \end{cases}$$

Find the probability of a bit error, P_B, for the case of equally likely signaling and the use of an optimum decision threshold.

3.8. **(a)** What is the theoretical minimum system bandwidth needed for a 10-Mbits/s signal using 16-level PAM without ISI?

(b) How large can the filter roll-off factor be if the allowable system bandwidth is 1.375 MHz?

3.9. A voice signal (300 to 3300 Hz) is digitized such that the quantization distortion $\le \pm 0.1\%$ of the peak-to-peak signal voltage. Assume a sampling rate of 8000 samples/s and a multilevel PAM waveform with $M = 32$ levels. Find the theoretical minimum system bandwidth that avoids ISI.

3.10. Binary data at 9600 bits/s are transmitted using 8-ary PAM modulation with a system using a raised cosine roll-off filter characteristic. The system has a frequency response out to 2.4 kHz.

(a) What is the symbol rate?

(b) What is the roll-off factor of the filter characteristic?

3.11. A voice signal in the range 300 to 3300 Hz is sampled at 8000 samples/s. We may transmit these samples directly as PAM pulses or we may first convert each sample to a PCM format and use binary (PCM) waveforms for transmission.

(a) What is the minimum system bandwidth required for the detection of PAM with no ISI and with a filter roll-off characteristic of $r = 1$?

(b) Using the same filter roll-off characteristic, what is the minimum bandwidth required for the detection of binary (PCM) waveforms if the samples are quantized to eight levels?

(c) Repeat part (b) using 128 quantization levels.

3.12. An analog signal is PCM formatted and transmitted using binary waveforms over a channel that is bandlimited to 100 kHz. Assume that 32 quantization levels are used and that the overall equivalent transfer function is of the raised cosine type with roll-off $r = 0.6$.

(a) Find the maximum bit rate that can be used by this system without introducing ISI.

(b) Find the maximum bandwidth of the original analog signal that can be accommodated with these parameters.

(c) Repeat parts (a) and (b) for transmission with 8-ary PAM waveforms.

3.13. Assume that equally-likely RZ binary pulses are coherently detected over a Gaussian channel with $N_0 = 10^{-8}$ Watt/Hz. Assume that synchronization is perfect, and that the received pulses have an amplitude of 100 mV. If the bit-error probability specification is $P_B = 10^{-3}$, find the largest data rate that can be transmitted using this system.

3.14. Consider that NRZ binary pulses are transmitted along a cable that attenuates the signal power by 3 dB (from transmitter to receiver). The pulses are coherently detected at the receiver, and the data rate is 56 kbit/s. Assume Gaussian noise with $N_0 = 10^{-6}$ Watt/Hz. What is the minimum amount of power needed at the transmitter in order to maintain a bit-error probability of $P_B = 10^{-3}$?

3.15. Show that the Nyquist minimum bandwidth for a random binary sequence sent with ideal-shaped bipolar pulses is the same as the noise equivalent bandwidth. Hint: the power spectral density for a random bipolar sequence is given in Equation (1.38) and the noise equivalent bandwidth is defined in Section 1.7.2.

3.16. Consider the 4-ary PAM-modulated sequence of message symbols {+1 +1 −1 +3 +1 +3}, where the members of the alphabet set are: {±1, ±3}. The pulses have been shaped with a root-raised cosine filter such that the support time of each filtered pulse is 6-symbol times, and the transmitted sequence is the analog waveform shown in Figure 3.23a. Note that the waveform appears "smeared" due to the filter-induced ISI. Show how a bank of N correlators can be implemented to perform matched-filter demodulation of the received pulse sequence, $r(t)$, where N corresponds to the number of symbols in the pulse-support time. [*Hint:* For the bank of correlators, use reference signals of the form $s_1(t - kT)$, where $k = 0, \ldots, 5$ and T is the symbol time.]

3.17. A desired impulse response of a communication system is the ideal $h(t) = \delta(t)$, where $\delta(t)$ is the impulse function. Assume that the channel introduces ISI so that the overall impulse response becomes $h(t) = \delta(t) + \alpha\delta(t - T)$, where $\alpha < 1$, and T is the symbol time. Derive an expression for the impulse response of a zero-forcing filter that will equalize the effects of ISI. Demonstrate that this filter suppresses the ISI. If the resulting suppression is deemed inadequate, how can the filter design be modified to increase the ISI suppression further?

3.18. The result of a single pulse (impulse) transmission is a received sequence of samples (impulse response), with values 0.1, 0.3, −0.2, 1.0, 0.4, −0.1, 0.1, where the leftmost sample is the earliest. The value 1.0 corresponds to the mainlobe of the pulse, and the other entries correspond to adjacent samples. Design a 3-tap transversal equalizer that forces the ISI to be zero at one sampling point on each side of the mainlobe. Calculate the values of the equalized output pulses at times $k = 0, \pm1, \ldots, \pm3$. After equalization, what is the largest magnitude sample contributing to ISI, and what is the sum of all the ISI magnitudes?

3.19. Repeat problem 3.18 for the case of a channel impulse response described by the following received samples: 0.01, 0.02, −0.03, 0.1, 1.0, 0.2, −0.1, 0.05, 0.02. Use a computer to find the weights of a nine-tap transversal equalizer to meet the minimum MSE criterion. Calculate the values of the equalized output pulses at times $k = 0, \pm1, \ldots, \pm8$. After equalization, what is the largest magnitude sample contributing to ISI, and what is the sum of all the ISI magnitudes?

3.20. In this chapter, it has been emphasized that signal-processing devices, such as multipliers and integrators, typically deal with signals having units of *volts*. Therefore, the transfer functions of such processors must accommodate these units. Draw a block diagram of a product-integrator showing the signal units on each of the wires, and the device transfer functions in each of the blocks (Hint: see Section 3.2.5.1).

QUESTIONS

3.1. In the case of *baseband* signaling, the received waveforms are already in a pulse-like form. Why then, is a demodulator needed to recover the pulse waveform? (See Chapter 3, introduction.)

3.2. Why is E_b/N_0 a natural figure-of-merit for digital communication systems? (See Section 3.1.5.)

3.3. When representing timed events, what dilemma can easily result in confusing the most-significant bit (MSB) and the least-significant bit (LSB)? (See Section 3.2.3.1.)

3.4. The term *matched-filter* is often used synonymously with *correlator*. How is that possible when their mathematical operations are different? (See Section 3.2.3.1.)

3.5. Describe the two fair ways of comparing different curves that depict bit-error probability versus E_b/N_0. (See Section 3.2.5.3.)

3.6. Are there other pulse-shaping filter functions, besides the *raised-cosine*, that exhibit zero ISI? (See Section 3.3.)

3.7. Describe a reasonable goal in endeavoring to *compress bandwidth* to the minimum possible, without incurring ISI. (See Section 3.3.1.1.)

3.8. The error performance of digital signaling suffers primarily from two degradation types: *loss* in signal-to-noise ratio, and *distortion* resulting in an irreducible bit-error probability. How do they differ? (See Section 3.3.2.)

3.9. Often times, providing more E_b/N_0 will not mitigate the degradation due to *intersymbol interference* (ISI). Explain why this is the case. (See Section 3.3.2.)

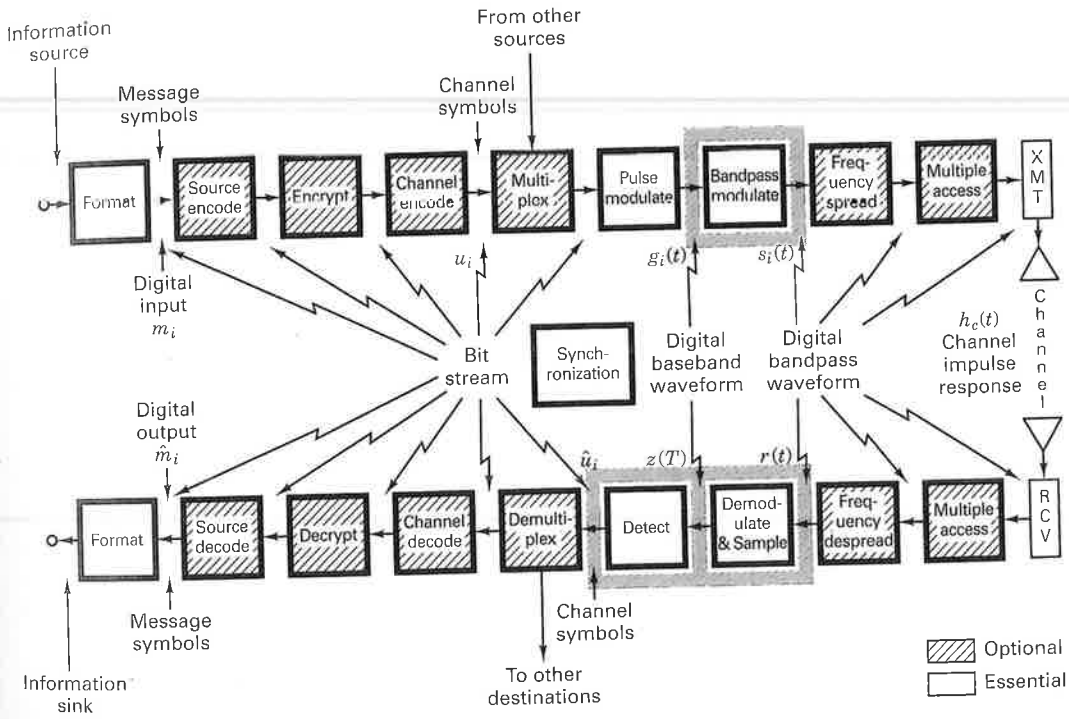
3.10. Describe the difference between equalizers that use a *zero-forcing* solution, and those that use a *minimum mean-square error* solution? (See Section 3.4.3.1.)

EXERCISES

Using the Companion CD, run the exercises associated with Chapter 3.

CHAPTER 4

Bandpass Modulation and Demodulation



4.1 WHY MODULATE?

Digital modulation is the process by which digital symbols are transformed into waveforms that are compatible with the characteristics of the channel. In the case of baseband modulation, these waveforms usually take the form of shaped pulses. But in the case of *bandpass modulation* the shaped pulses modulate a sinusoid called a *carrier wave,* or simply a *carrier;* for radio transmission the carrier is converted to an electromagnetic (EM) field for propagation to the desired destination. One might ask why it is necessary to use a carrier for the radio transmission of baseband signals. The answer is as follows. The transmission of EM fields through space is accomplished with the use of antennas. The size of the antenna depends on the wavelength λ and the application. For cellular telephones, antennas are typically $\lambda/4$ in size, where wavelength is equal to c/f, and c, the speed of light, is 3×10^8 m/s. Consider sending a baseband signal (say, $f = 3000$ Hz) by coupling it to an antenna directly without a carrier wave. How large would the antenna have to be? Let us size it by using the telephone industry benchmark of $\lambda/4$ as the antenna dimension. For the 3,000 Hz baseband signal, $\lambda/4 = 2.5 \times 10^4$ $m \approx 15$ miles. To transmit a 3,000 Hz signal through space, *without carrier-wave modulation,* an antenna that spans 15 miles would be required. However, if the baseband information is first modulated on a higher frequency carrier, for example a 900 MHz carrier, the equivalent antenna diameter would be about 8 cm. For this reason, carrier-wave or bandpass modulation is an essential step for all systems involving radio transmission.

Bandpass modulation can provide other important benefits in signal transmission. If more than one signal utilizes a single channel, modulation may be used to separate the different signals. Such a technique, known as *frequency-division multiplexing,* is discussed in Chapter 11. Modulation can be used to minimize the effects of interference. A class of such modulation schemes, known as *spread-spectrum modulation,* requires a system bandwidth much larger than the minimum bandwidth that would be required by the message. The trade-off of bandwidth for interference rejection is considered in Chapter 12. Modulation can also be used to place a signal in a frequency band where design requirements, such as filtering and amplification, can be easily met. This is the case when radio-frequency (RF) signals are converted to an intermediate frequency (IF) in a receiver.

4.2 DIGITAL BANDPASS MODULATION TECHNIQUES

Bandpass modulation (either analog or digital) is the process by which an information signal is converted to a sinusoidal waveform; for digital modulation, such a sinusoid of duration T is referred to as a digital symbol. The sinusoid has just three features that can be used to distinguish it from other sinusoids: amplitude, frequency, and phase. Thus bandpass modulation can be defined as the process whereby the amplitude, frequency, or phase of an RF carrier, or a combination of them, is varied in accordance with the information to be transmitted. The general form of the carrier wave is

$$s(t) = A(t) \cos \theta(t) \tag{4.1}$$

where $A(t)$ is the time-varying amplitude and $\theta(t)$ is the time-varying angle. It is convenient to write

$$\theta(t) = \omega_0 t + \phi(t) \tag{4.2}$$

so that

$$s(t) = A(t) \cos [\omega_0 t + \phi(t)] \tag{4.3}$$

where ω_0 is the *radian frequency* of the carrier and $\phi(t)$ is the *phase.* The terms f and ω will each be used to denote frequency. When f is used, frequency in hertz is intended; when ω is used, frequency in radians per second is intended. The two frequency parameters are related by $\omega = 2\pi f$.

The basic *bandpass modulation/demodulation* types are listed in Figure 4.1. When the receiver exploits knowledge of the carrier's phase to detect the signals, the process is called *coherent detection;* when the receiver does not utilize such phase reference information, the process is called *noncoherent detection.* In digital communications, the terms *demodulation* and *detection* are often used interchangeably, although demodulation emphasizes waveform recovery, and detection entails the process of symbol decision. In ideal coherent detection, there is available at the receiver a prototype of each possible arriving signal. These prototype waveforms attempt to duplicate the transmitted signal set in every respect, even RF phase. The

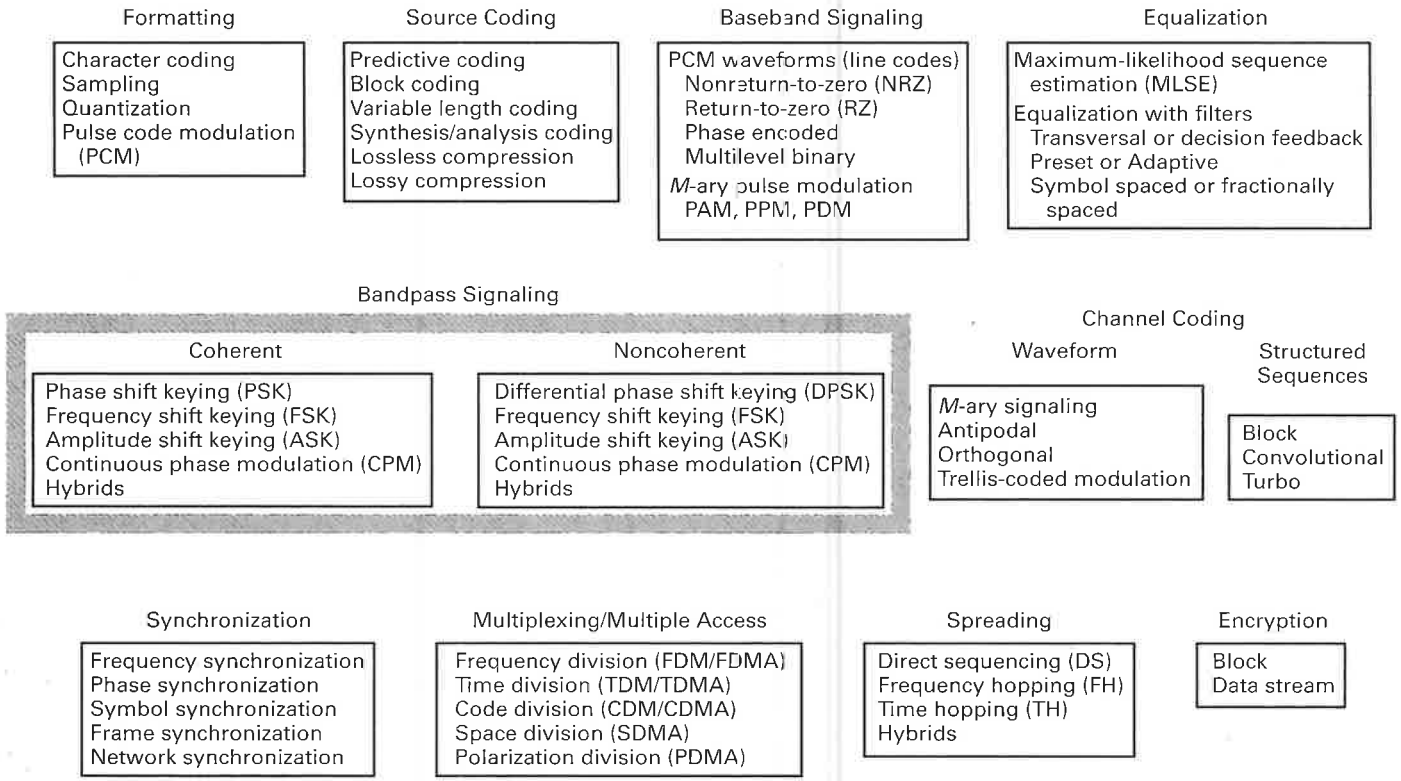


Figure 4.1 Basic digital communication transformations.

receiver is then said to be *phase locked* to the incoming signal. During demodulation, the receiver multiplies and integrates (correlates) the incoming signal with each of its prototype replicas. Under the heading of coherent modulation/demodulation in Figure 4.1 are listed phase shift keying (PSK), frequency shift keying (FSK), amplitude shift keying (ASK), continuous phase modulation (CPM), and hybrid combinations. The basic bandpass modulation formats are discussed in this chapter. Some specialized formats, such as offset quadrature PSK (OQPSK), minimum shift keying (MSK) belonging to the CPM class, and quadrature amplitude modulation (QAM), are treated in Chapter 9.

Noncoherent demodulation refers to systems employing demodulators that are designed to operate without knowledge of the absolute value of the incoming signal's phase; therefore, phase estimation is not required. Thus the advantage of noncoherent over coherent systems is reduced complexity, and the price paid is increased probability of error (P_E). In Figure 4.1 the modulation/demodulation types that are listed in the noncoherent column, DPSK, FSK, ASK, CPM, and hybrids, are similar to those listed in the coherent column. We had implied that phase information is not used for noncoherent reception; how do you account for the fact that there is a form of phase shift keying under the noncoherent heading? It turns out that an important form of PSK can be classified as noncoherent (or differentially coherent) since it does not require a reference in phase with the received carrier. This "pseudo-PSK," termed *differential PSK* (DPSK), utilizes phase information of the prior symbol as a phase reference for detecting the current symbol. This is described in Sections 4.5.1 and 4.5.2.

4.2.1 Phasor Representation of a Sinusoid

Using a well-known trigonometric identity called Euler's theorem, we introduce the complex notation of a sinusoidal carrier wave as follows:

$$e^{j\omega_0 t} = \cos \omega_0 t + j \sin \omega_0 t \tag{4.4}$$

One might be more comfortable with the simpler, more straightforward notation $\cos \omega_0 t$ or $\sin \omega_0 t$. What possible benefit can there be with the complex notation? We will see (in Section 4.6) that this notation facilitates our description of how real-world modulators and demodulators are implemented. For now, let us point to the general benefits of viewing a carrier wave in the complex form of Equation (4.4).

First, within this compact from, $e^{j\omega_0 t}$, is contained the two important quadrature components of any sinusoidal carrier wave, namely the inphase (real) and the quadrature (imaginary) components that are orthogonal to each other. Second, the unmodulated carrier wave is conveniently represented in a polar coordinate system as a unit vector or phasor rotating counterclockwise at the constant rate of ω_0 radians/s, as depicted in Figure 4.2. As time is increasing (i.e., from t_0 to t_1) we can visualize the time-varying projections of the rotating phasor on the inphase (I) axis and the quadrature (Q) axis. These cartesian axes are usually referred to as the I channel and Q channel respectively, and the projections on them represent the

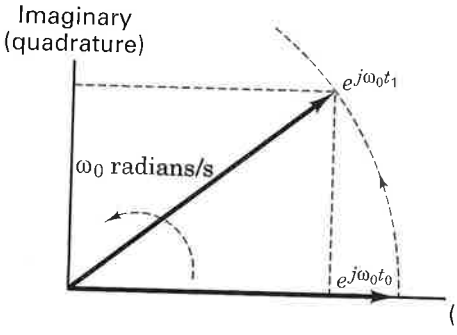


Figure 4.2 Phasor representation of a sinusoid.

signal components (orthogonal to each other) associated with those channels. Third, when it comes time to modulate the carrier wave with information, we can view this modulation as a methodical perturbation of the rotating phasor (and its projections).

For example, consider a carrier wave that is *amplitude modulated* (AM) with a sinusoid having an amplitude of unity and a frequency ω_m, where $\omega_m \ll \omega_0$. The analytical form of the transmitted waveform is

$$s(t) = \mathrm{Re}\left\{ e^{j\omega_0 t}\left(1 + \frac{e^{j\omega_m t}}{2} + \frac{e^{-j\omega_m t}}{2}\right)\right\} \tag{4.5}$$

where Re{x} is the real part of the complex quantity {x}. Figure 4.3 illustrates that the rotating phasor $e^{j\omega_0 t}$ of Figure 4.2 is now perturbed by two sideband terms—$e^{j\omega_m t}/2$ rotating counterclockwise and $e^{-j\omega_m t}/2$ rotating clockwise. The sideband phasors are rotating at a much slower speed than the carrier-wave phasor. The net result of the composite signal is that the rotating carrier-wave phasor now appears to be growing longer and shorter pursuant to the dictates of the sidebands, but its frequency stays constant—hence, the term "amplitude modulation."

Another example to reinforce the usefulness of the phasor view is that of *frequency modulating* (FM) the carrier wave with a similar sinusoid having a frequency of ω_m radians/s. The analytical representation of *narrowband* FM (NFM) has an appearance similar to AM and is represented by

$$s(t) = \mathrm{Re}\left\{ e^{j\omega_0 t}\left(1 - \frac{\beta}{2} e^{-j\omega_m t} + \frac{\beta}{2} e^{j\omega_m t}\right)\right\} \tag{4.6}$$

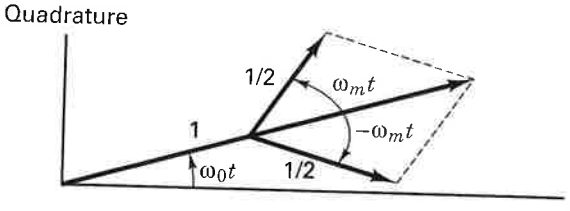


Figure 4.3 Amplitude modulation.

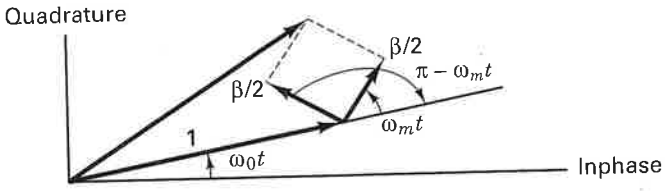


Figure 4.4 Narrowband frequency modulation.

where β is the modulation index [1]. Figure 4.4 illustrates that the rotating carrier-wave phasor is again perturbed by two sideband terms, but because one of the sideband terms carries a minus sign in Equation (4.6), the clockwise and counterclockwise rotating sideband phasors have a different symmetry than in the case of AM. In the case of AM the sideband symmetry results in the carrier-wave phasor growing longer and shorter with time. In NFM, the sideband symmetry (90° different than AM) results in the carrier-wave phasor speeding up and slowing down according to the dictates of the sidebands, but the amplitude says essentially constant—hence, the term "frequency modulation."

Figure 4.5 illustrates examples of the most common digital modulation formats: PSK, FSK, ASK, and a hybrid combination of ASK and PSK (ASK/PSK or APK). The first column lists the analytic expression, the second is a typical pictorial of the waveform versus time, and the third is a vector (or phasor) schematic, with the orthogonal axes labeled $\{\psi_j(t)\}$. In the general M-ary signaling case, the processor accepts k source bits (or channel bits if there is coding) at a time and instructs the modulator to produce one of an available set of $M = 2^k$ waveform types. Binary modulation, where $k = 1$, is just a special case of M-ary modulation.

In Figure 4.2, we represented a carrier wave as a phasor rotating in a plane at the speed of the carrier-wave frequency ω_0 radians/s. In Figure 4.5, the phasor schematic for each digital-modulation example represents a constellation of information signals (vectors or points in the signaling space), where time is not represented. In other words, the constantly rotating aspect of the unmodulated carrier wave has been removed, and only the information-bearing phasor positions, relative to one another, are presented. Each example in Figure 4.5 uses a particular value of M, the set size.

4.2.2 Phase Shift Keying

Phase shift keying (PSK) was developed during the early days of the deep-space program; PSK is now widely used in both military and commercial communications systems. The general analytic expression for PSK is

$$s_i(t) = \sqrt{\frac{2E}{T}} \cos\left[\omega_0 t + \phi_i(t)\right] \qquad \begin{matrix} 0 \le t \le T \\ i = 1, \ldots, M \end{matrix} \tag{4.7}$$

where the phase term, $\phi_i(t)$, will have M discrete values, typically given by

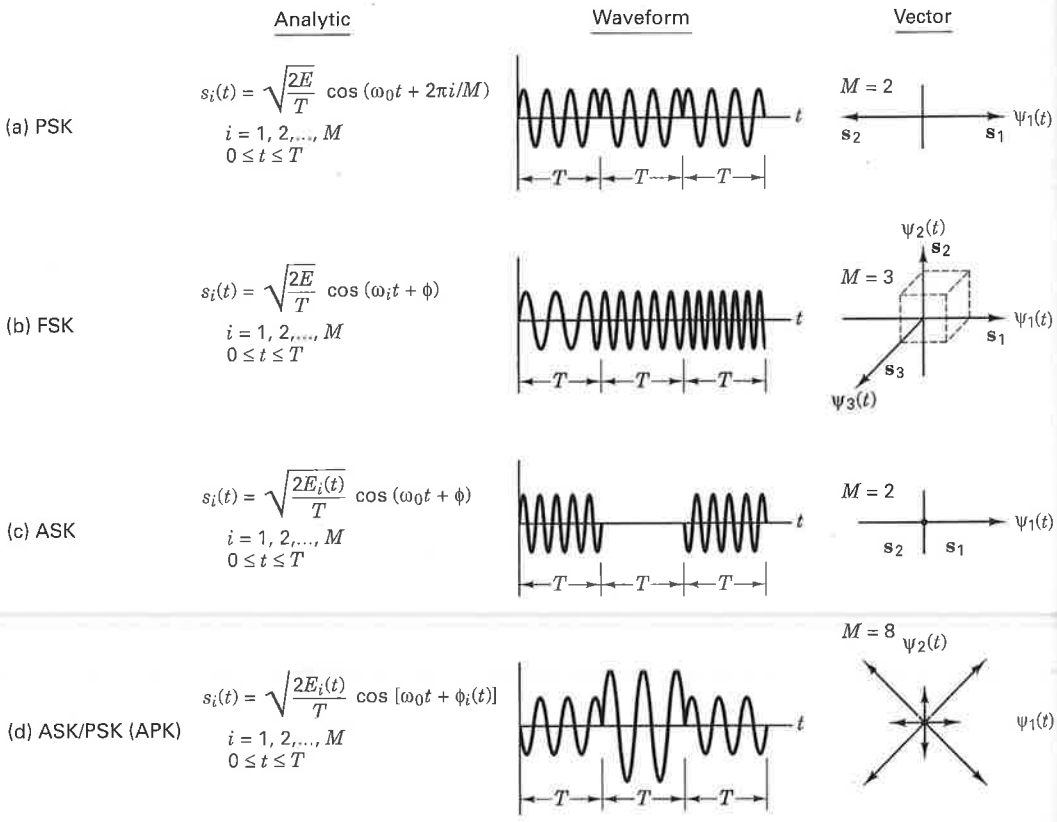
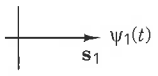


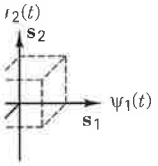
Figure 4.5 Digital modulations. (a) PSK. (b) FSK. (c) ASK. (d) ASK/PSK (APK).

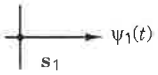
$$\phi_i(t) = \frac{2\pi i}{M} \qquad i = 1, \ldots, M$$

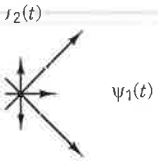
For the binary PSK (BPSK) example in Figure 4.5a, M is 2. The parameter E is symbol energy, T is symbol time duration, and $0 \le t \le T$. In BPSK modulation, the modulating data signal shifts the phase of the waveform $s_i(t)$ to one of two states, either zero or π (180°). The waveform sketch in Figure 4.5a shows a typical BPSK waveform with its abrupt phase changes at the symbol transitions; if the modulating data stream were to consist of alternating ones and zeros, there would be such an abrupt change at each transition. The signal waveforms can be represented as vectors or phasors on a polar plot; the vector length corresponds to the signal amplitude, and the vector direction for the general M-ary case corresponds to the signal phase relative to the other $M - 1$ signals in the set. For the BPSK example, the vector picture illustrates the two 180° opposing vectors. Signal sets that can be depicted with such opposing vectors are called *antipodal signal sets.*

ector









SK (APK).

arameter E is
odulation, the
of two states,
typical BPSK
ıe modulating
ld be such an
epresented as
to the signal
sponds to the
PSK example,
ts that can be

on Chap. 4

4.2.3 Frequency Shift Keying

The general analytic expression for FSK modulation is

$$s_i(t) = \sqrt{\frac{2E}{T}} \cos(\omega_i t + \phi) \qquad \begin{matrix} 0 \le t \le T \\ i = 1, \ldots, M \end{matrix} \tag{4.8}$$

where the frequency term ω_i has M discrete values, and the phase term ϕ is an arbitrary constant. The FSK waveform sketch in Figure 4.5b illustrates the typical frequency changes at the symbol transitions. At the symbol transitions, the figure depicts a gentle shift from one frequency (tone) to another. This behavior is only true for a special class of FSK called continuous-phase FSK (CPFSK) which is described in Section 9.8. In the general MFSK case, the change to a different tone can be quite abrupt, because there is no requirement for the phase to be continuous. In this example, M has been chosen equal to 3, corresponding to the same number of waveform types (3-ary); note that this $M = 3$ choice for FSK has been selected to emphasize the mutually perpendicular axes. In practice, M is usually a nonzero power of 2 (2, 4, 8, 16, ...). The signal set is characterized by Cartesian coordinates, such that each of the mutually perpendicular axes represents a sinusoid with a different frequency. As described earlier, signal sets that can be characterized with such mutually perpendicular vectors are called *orthogonal* signals. Not all FSK signaling is orthogonal. For any signal set to be orthogonal, it must meet the criterion set forth in Equation (3.69). For an FSK signal set, in the process of meeting this criterion, a condition arises on the spacing between the tones in the set. The necessary frequency spacing between tones to fulfill the orthogonality requirement is discussed in Section 4.5.4.

4.2.4 Amplitude Shift Keying

For the ASK example in Figure 4.5c, the general analytic expression is

$$s_i(t) = \sqrt{\frac{2E_i(t)}{T}} \cos(\omega_0 t + \phi) \qquad \begin{matrix} 0 \le t \le T \\ i = 1, \ldots, M \end{matrix} \tag{4.9}$$

where the amplitude term $\sqrt{2E_i(t)/T}$ will have M discrete values, and the phase term ϕ is an arbitrary constant. In Figure 4.5c, M has been chosen equal to 2, corresponding to two waveform types. The ASK waveform sketch in the figure can describe a radar transmission example, where the two signal amplitude states would be $\sqrt{2E/T}$ and zero. The vector picture utilizes the same phase–amplitude polar coordinates as the PSK example. Here we see a vector corresponding to the maximum-amplitude state, and a point at the origin corresponding to the zero-amplitude state. Binary ASK signaling (also called on–off keying) was one of the earliest forms of digital modulation used in radio telegraphy in the early 1900s. Simple ASK is no longer widely used in digital communications systems, and thus it will not be treated in detail here.

4.2.5 Amplitude Phase Keying

For the combination of ASK and PSK (APK) example in Figure 4.5d, the general analytic expression

$$s_i(t) = \sqrt{\frac{2E_i(t)}{T}} \cos\left[\omega_0 t + \phi_i(t)\right] \qquad \begin{matrix} 0 \le t \le T \\ i = 1, \ldots, M \end{matrix} \tag{4.10}$$

illustrates the indexing of both the signal amplitude term and the phase term. The APK waveform picture in Figure 4.5d illustrates some typical simultaneous phase and amplitude changes at the symbol transition times. For this example, M has been chosen equal to 8, corresponding to eight waveforms (8-ary). The figure illustrates a hypothetical eight-vector signal set on the phase-amplitude plane. Four of the vectors are at one amplitude, and the other four vectors are at a different amplitude. Each of the vectors is separated by 45°. When the set of M symbols in the two-dimensional signal space are arranged in a rectangular constellation, the signaling is referred to as quadrature amplitude modulation (QAM); examples of QAM are considered in Chapter 9.

The vector picture for each of the modulation types described in Figure 4.5 (except the FSK case) is characterized on a plane whose *polar* coordinates represent signal *amplitude* and *phase*. The FSK case assumes orthogonal FSK (see Section 4.5.4) and is characterized in a *Cartesian* coordinate space, with each axis representing a *frequency tone* ($\cos \omega_i t$) from the M-ary set of orthogonal tones.

4.2.6 Waveform Amplitude Coefficient

The waveform amplitude coefficient appearing in Equations (4.7) to (4.10) has the same general form $\sqrt{2E/T}$ for all modulation formats. The derivation of this expression begins with

$$s(t) = A \cos \omega t \tag{4.11}$$

where A is the peak value of the waveform. Since the peak value of a sinusoidal waveform equals $\sqrt{2}$ times the root-mean-square (rms) value, we can write

$$\begin{aligned} s(t) &= \sqrt{2} A_{\text{rms}} \cos \omega t \\ &= \sqrt{2A_{\text{rms}}^2} \cos \omega t \end{aligned}$$

Assuming the signal to be a voltage or a current waveform, A_{rms}^2 represents average power P (normalized to 1 Ω). Therefore, we can write

$$s(t) = \sqrt{2P} \cos \omega t \tag{4.12}$$

Replacing P watts by E joules/T seconds, we get

$$s(t) = \sqrt{\frac{2E}{T}} \cos \omega t \tag{4.13}$$

We shall use either the amplitude notation A in Equation (4.11) or the designation $\sqrt{2E/T}$ in Equation (4.13). Since the *energy* of a received signal is the key parameter in determining the error performance of the detection process, it is often more convenient to use the amplitude notation in Equation (4.13) because it facilitates solving directly for the probability of error P_E as a function of signal energy.

4.3 DETECTION OF SIGNALS IN GAUSSIAN NOISE

The bandpass model of the detection process is virtually identical to the baseband model considered in Chapter 3. That is because a received bandpass waveform is first transformed to a baseband waveform before the final detection step takes place. For linear systems, the mathematics of detection is unaffected by a shift in frequency. In fact, we can define an *equivalence theorem* as follows: Performing bandpass linear signal processing, followed by heterodyning the signal to baseband yields the same results as heterodyning the bandpass signal to baseband, followed by baseband linear signal processing. The term "heterodyning" refers to a frequency *conversion* or *mixing* process that yields a spectral shift in the signal. As a result of this equivalence theorem, all linear signal-processing simulations can take place at baseband (which is preferred for simplicity), with the same results as at bandpass. This means that the performance of most digital communication systems will often be described and analyzed as if the transmission channel is a baseband channel.

4.3.1 Decision Regions

Consider that the two-dimensional signal space in Figure 4.6 is the locus of the noise-perturbed prototype binary vectors $(\mathbf{s}_1 + \mathbf{n})$ and $(\mathbf{s}_2 + \mathbf{n})$. The noise vector, $\mathbf{n}$, is a zero-mean random vector; hence the received signal vector, $\mathbf{r}$, is a random vector with mean $\mathbf{s}_1$ or $\mathbf{s}_2$. The detector's task after receiving $\mathbf{r}$ is to decide which of the signals ($\mathbf{s}_1$ or $\mathbf{s}_2$,) was actually transmitted. The method is usually to decide on the signal classification that yields the minimum expected P_E, although other strategies are possible [2]. For the case where M equals 2, with $\mathbf{s}_1$ and $\mathbf{s}_2$ being equally likely and with the noise being an additive white Gaussian noise (AWGN) process, we will see that the minimum-error decision rule is equivalent to choosing the signal class such that the distance $d(\mathbf{r}, \mathbf{s}_i) = \| \mathbf{r} - \mathbf{s}_i \|$ is minimized, where $\| \mathbf{x} \|$ is called the *norm* or *magnitude* of vector $\mathbf{x}$. This rule is often stated in terms of decision regions. In Figure 4.6, let us construct decision regions in the following way. Draw a line connecting the tips of the prototype vectors $\mathbf{s}_1$ and $\mathbf{s}_2$. Next, construct the perpendicular bisector of the connecting line. Notice that this bisector passes through the origin of the space if $\mathbf{s}_1$ and $\mathbf{s}_2$ are equal in amplitude. For this $M = 2$ example, in Figure 4.6, the constructed perpendicular bisector represents the locus of points equidistant between $\mathbf{s}_1$ and $\mathbf{s}_2$; hence, the bisector describes the boundary between decision region 1 and decision region 2. The *decision rule* for the detector, stated in terms of *decision regions,* is as follows: Whenever the received signal $\mathbf{r}$ is located

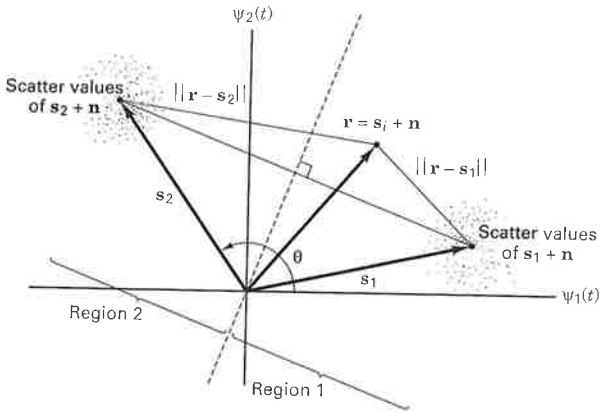


Figure 4.6 Two-dimensional signal space, with arbitrary equal-amplitude vectors $\mathbf{s}_1$ and $\mathbf{s}_2$.

in region 1, choose signal $\mathbf{s}_1$; when it is located in region 2, choose signal $\mathbf{s}_2$. In Figure 4.6, if the angle θ equals 180°, then the signal set $\mathbf{s}_1$ and $\mathbf{s}_2$ represents BPSK. However, in this figure, θ was purposely chosen to be less than 180° in order to emphasize the idea of decision regions in general.

4.3.2 Correlation Receiver

In Section 3.2 we treated the detection of *baseband* binary signals in Gaussian noise. Since the detection of *bandpass* signals employs the same concepts, we shall summarize the key findings of that section. We focus particularly on that realization of a matched filter known as a *correlator*. In addition to binary detection, we also consider the more general case of M-ary detection. We assume that the only performance degradation is due to AWGN. The received signal is the sum of the transmitted prototype signal plus the random noise:

$$r(t) = s_i(t) + n(t) \qquad \begin{matrix} 0 \le t \le T \\ i = 1, \ldots, M \end{matrix} \tag{4.14}$$

Given such a received signal, the detection process consists of *two basic steps* as shown in Figure 3.1. In the first step, the received waveform, $r(t)$, is reduced to a *single random variable* $z(T)$, or to a *set of random variables* $z_i(T)$ ($i = 1, \ldots, M$), formed at the output of the demodulator and sampler at time $t = T$, where T is the symbol duration. In the second step, a symbol decision is made on the basis of comparing $z(T)$ to a threshold or on the basis of choosing the maximum $z_i(T)$. Step 1 can be thought of as transforming the waveform into a point in the decision space. We call this point the *predetection point,* and view it as the "location" of the received E_b/N_0 (or other related receiving-system SNR parameter); beyond here, signals are replaced by bits. Often, such SNR parameters are modeled with reference to the output of the receiving antenna. Note that although the signal power and noise power is each different at different points in the system, a receiving-system SNR can be modeled to yield the same value at various reference points (see Section 5.5.5). At the predetection point, there will be a new sample of a baseband pulse during each symbol

time. Note that E_b/N_0 is defined at a place where there are no bits yet. Bits will only appear after completion of the detection process. Perhaps a better name for E_b/N_0 would be energy per *effective* bit versus N_0. Step 2 can be thought of as determining *in which decision region* the point is located. For the detector to be optimized (in the sense of minimizing the error probability), it is necessary to optimize the waveform-to-random-variable transformation, by using matched filters or correlators in step 1, and by also optimizing the decision criterion in step 2.

In Sections 3.2.2 and 3.2.3 we found that the matched filter provides the maximum signal-to-noise ratio at the filter output at time $t = T$. We described a correlator as one realization of a matched filter. We can define a *correlation receiver* comprised of M correlators, as shown in Figure 4.7a, that transforms a received waveform, $r(t)$, to a sequence of M numbers or correlator outputs, $z_i(T)$ $(i = 1, \ldots, M)$. Each correlator output is characterized by the following product integration or correlation with the received signal:

$$z_i(T) = \int_0^T r(t)s_i(t)\,dt \qquad i = 1, \ldots, M \tag{4.15}$$

The verb "to correlate" means "to match." The correlators attempt to match the incoming received signal, $r(t)$, with each of the candidate prototype waveforms, $s_i(t)$, known a priori to the receiver. A reasonable decision rule is to choose the waveform, $s_i(t)$, that *matches best* or has the *largest correlation* with $r(t)$. In other words, the decision rule is

$$\begin{array}{l}\text{Choose the } s_i(t) \text{ whose index} \\ \text{corresponds to the max } z_i(T)\end{array} \tag{4.16}$$

Following Equation (3.10), any signal set, $\{s_i(t)\}$ $(i = 1, \ldots, M)$, can be expressed in terms of some set of basis functions, $\{\psi_j(t)\}$ $(j = 1, \ldots, N)$, where $N \leq M$. Then the bank of M correlators in Figure 4.7a may be replaced with a bank of N correlators, shown in Figure 4.7b, where the set of basis functions $\{\psi_j(t)\}$ form *reference signals.* The decision stage of this receiver consists of logic circuitry for choosing the signal, $s_i(t)$. The choice of $s_i(t)$ is made according to the best match of the coefficients, a_{ij}, seen in Equation (3.10), with the set of outputs $\{z_j(T)\}$. When the prototype waveform set, $\{s_i(t)\}$, is an orthogonal set, the receiver implementation in Figure 4.7a is identical to that in Figure 4.7b (differing perhaps by a scale factor). However, when $\{s_i(t)\}$ is *not* an orthogonal set, the receiver in Figure 4.7b, using N correlators instead of M, with reference signals $\{\psi_j(t)\}$, can represent a cost-effective implementation. We examine such an application for the detection of multiple phase shift keying (MPSK) in Section 4.4.3.

In the case of *binary detection,* the correlation receiver can be configured as a single matched filter or product integrator, as shown in Figure 4.8a, with the reference signal being the difference between the binary prototype signals, $s_1(t) - s_2(t)$. The output of the correlator, $z(T)$, is fed directly to the decision stage.

For binary detection, the correlation receiver can also be drawn as two matched filters or product integrators, one of which is matched to $s_1(t)$, and the

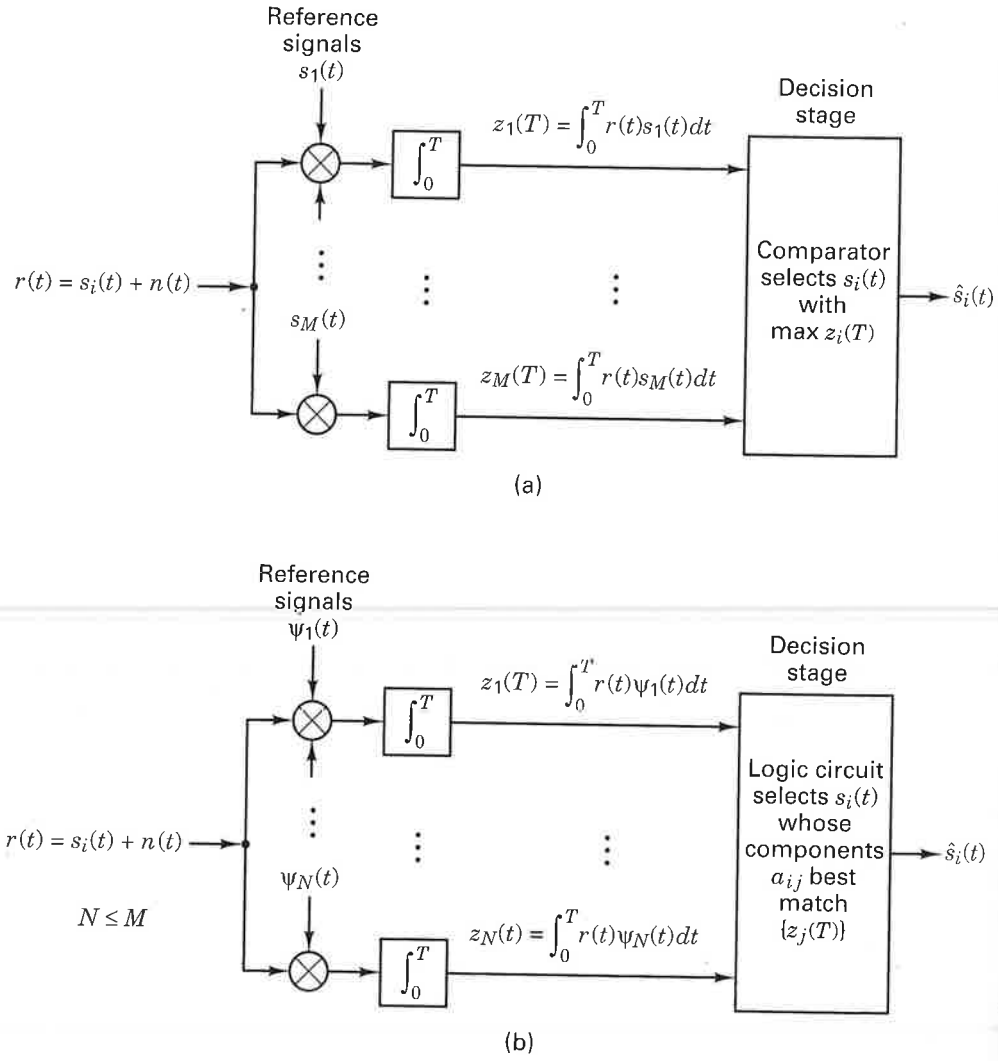


Figure 4.7 (a) Correlator receiver with reference signals $\{s_i(t)\}$. (b) Correlator receiver with reference signals $\{\psi_j(t)\}$.

other is matched to $s_2(t)$. (See Figure 4.8b.) The decision stage can then be configured to follow the rule in Equation 4.16, or the correlator outputs $z_i(T)$ $(i = 1, 2)$ can be differenced to form

$$z(T) = z_1(T) - z_2(T) \tag{4.17}$$

as shown in Figure 4.8b. Then, $z(T)$, called the *test statistic,* is fed to the decision stage, as in the case of the single correlator. In the *absence of noise,* an input waveform $s_i(t)$ yields the output $z(T) = a_i(T)$, a signal-only component. The input noise

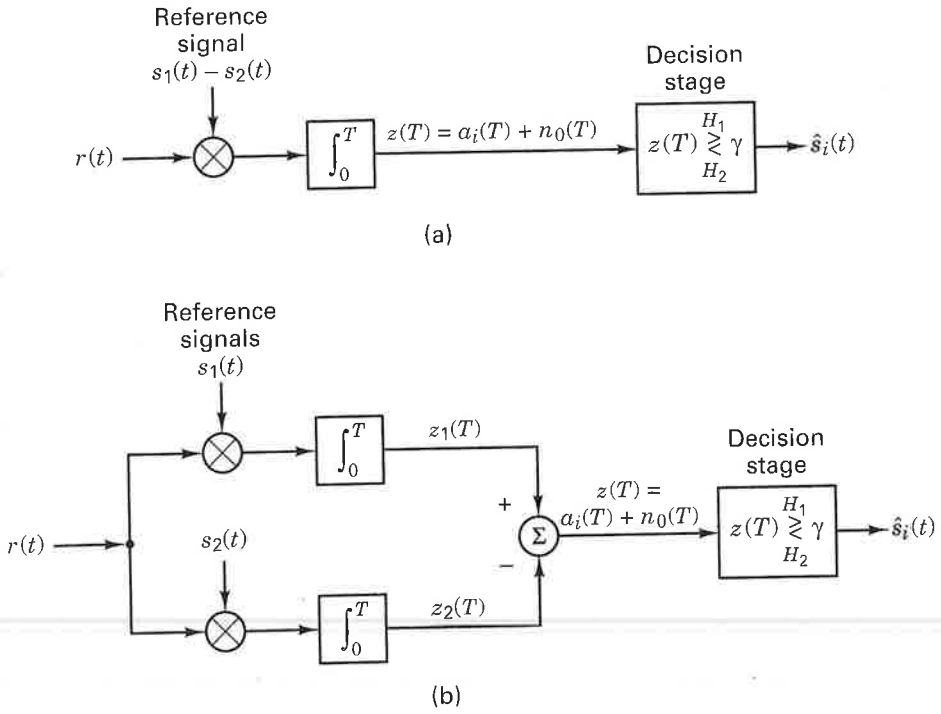


Figure 4.8 Binary correlator receiver. (a) Using a single correlator. (b) Using two correlators.

$n(T)$ is a Gaussian random process. Since the correlator is a *linear* device, the output noise is also a Gaussian random process [2]. Thus, the output of the correlator, sampled at $t = T$, yields

$$z(T) = a_i(T) + n_0(T) \qquad i = 1, 2$$

where $n_0(T)$ is the noise component. To shorten the notation we sometimes express $z(t)$ as $a_i + n_0$. The noise component n_0 is a zero-mean *Gaussian random variable,* and thus $z(T)$ is a *Gaussian random variable* with a mean of either a_1 or a_2, depending on whether a binary one or binary zero was sent.

4.3.2.1 Binary Decision Threshold

For the random variable $z(T)$, Figure 4.9 illustrates the two conditional probability density functions (pdfs), $p(z|s_1)$ and $p(z|s_2)$, with mean value of a_1 and a_2, respectively. These pdfs, also called the *likelihood* of s_1 and the *likelihood* of s_2, respectively, were presented in Section 3.1.2, and are rewritten as

$$p(z|s_1) = \frac{1}{\sigma_0\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{z - a_1}{\sigma_0}\right)^2\right] \tag{4.18a}$$

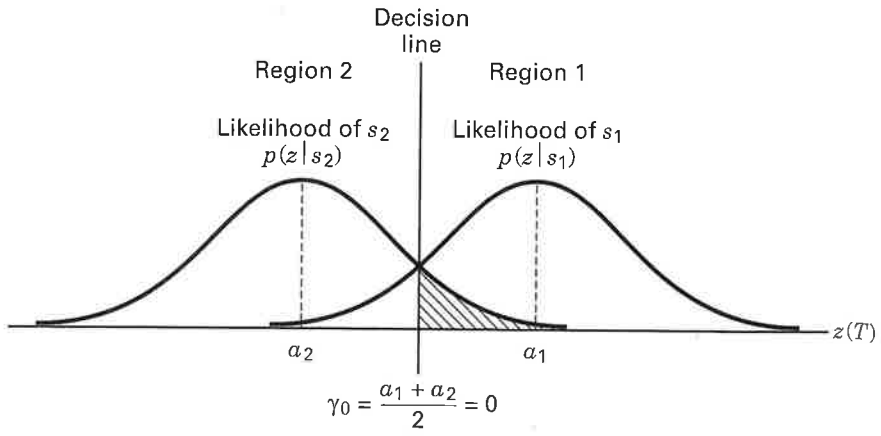


Figure 4.9 Conditional probability density functions: $p(z/s_1)$, $p(z/s_2)$.

and

$$p(z|s_2) = \frac{1}{\sigma_0\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{z - a_2}{\sigma_0}\right)^2\right] \tag{4.18b}$$

where σ_0^2 is the noise variance. In Figure 4.9 the rightmost likelihood $p(z|s_1)$ illustrates the probability density of the detector output $z(T)$, given that $s_1(t)$ was transmitted. Similarly, the leftmost likelihood $p(z|s_2)$ illustrates the probability density of $z(T)$, given that $s_2(t)$ was transmitted. The abscissa $z(T)$ represents the full range of possible sample output values from the correlation receiver shown in Figure 4.8.

With regard to optimizing the binary decision threshold for deciding in which region a received signal is located, we found in Section 3.2.1 that the *minimum error* criterion for equally likely binary signals corrupted by Gaussian noise can be stated as

$$z(T) \underset{H_2}{\overset{H_1}{\gtrless}} \frac{a_1 + a_2}{2} = \gamma_0 \tag{4.19}$$

where a_1 is the signal component of $z(T)$ when $s_1(t)$ is transmitted, and a_2 is the signal component of $z(T)$ when $s_2(t)$ is transmitted. The threshold level γ_0 represented by $(a_1 + a_2)/2$ is the *optimum threshold* for minimizing the probability of making an incorrect decision given equally likely signals and symmetrical likelihoods. The decision rule in Equation (4.19) states that hypothesis H_1 should be selected [equivalent to deciding that signal $s_1(t)$ was sent] if $z(T) > \gamma_0$, and hypothesis H_2 should be selected [equivalent to deciding that $s_2(t)$ was sent] if $z(T) < \gamma_0$. If $z(T) = \gamma_0$, the decision can be an arbitrary one. For equal-energy, equally likely antipodal signals, where $s_1(t) = -s_2(t)$ and $a_1 = -a_2$, the optimum decision rule becomes

$$z(T) \underset{H_2}{\overset{H_1}{\gtrless}} 0 \tag{4.20a}$$

or

$$\begin{array}{ll} \text{decide } s_1(t) & \text{if } z_1(T) > z_2(T) \\ \text{decide } s_2(t) & \text{otherwise} \end{array} \tag{4.20b}$$

4.4 COHERENT DETECTION

4.4.1 Coherent Detection of PSK

The detector shown in Figure 4.7 can be used for the coherent detection of any digital waveforms. Such a correlating detector is often referred to as a *maximum likelihood detector.* Consider the following binary PSK (BPSK) example: Let

$$s_1(t) = \sqrt{\frac{2E}{T}} \cos(\omega_0 t + \phi) \qquad 0 \le t \le T \tag{4.21a}$$

$$\begin{aligned} s_2(t) &= \sqrt{\frac{2E}{T}} \cos(\omega_0 t + \phi + \pi) \\ &= -\sqrt{\frac{2E}{T}} \cos(\omega_0 t + \phi) \qquad 0 \le t \le T \end{aligned} \tag{4.21b}$$

and

$$n(t) = \text{zero-mean white Gaussian random process}$$

where the phase term ϕ is an arbitrary constant, so that the analysis is unaffected by setting $\phi = 0$. The parameter E is the signal energy per symbol, and T is the symbol duration. For this antipodal case, only a single basis function is needed. If an orthonormal signal space is assumed in Equations (3.10) and (3.11) (i.e., $K_j = 1$), we can express a basis function $\psi_1(t)$ as

$$\psi_1(t) = \sqrt{\frac{2}{T}} \cos \omega_0 t \qquad \text{for } 0 \le t \le T \tag{4.22}$$

Thus, we may express the transmitted signals $s_i(t)$ in terms of $\psi_1(t)$ and the coefficients $a_{i1}(t)$ as follows:

$$s_i(t) = a_{i1}\psi_1(t) \tag{4.23a}$$

$$s_1(t) = a_{11}\psi_1(t) = \sqrt{E}\,\psi_1(t) \tag{4.23b}$$

$$s_2(t) = a_{21}\psi_1(t) = -\sqrt{E}\,\psi_1(t) \tag{4.23c}$$

Assume that $s_1(t)$ was transmitted. Then the expected values of the product integrators in Figure 4.7b, with reference signal $\psi_1(t)$, are found as

$$\mathbf{E}\{z_1|s_1\} = \mathbf{E}\left\{ \int_0^T \sqrt{E}\ \psi_1^2(t) + n(t)\psi_1(t)\ dt \right\} \tag{4.24a}$$

$$\mathbf{E}\{z_2|s_1\} = \mathbf{E}\left\{ \int_0^T -\sqrt{E}\ \psi_1^2(t) + n(t)\psi_1(t)\ dt \right\} \tag{4.24b}$$

$$\mathbf{E}\{z_1|s_1\} = \mathbf{E}\left\{ \int_0^T \frac{2}{T}\sqrt{E}\cos^2 \omega_0 t + n(t)\sqrt{\frac{2}{T}}\cos \omega_0 t\ dt \right\} = \sqrt{E} \tag{4.25a}$$

and

$$\mathbf{E}\{z_2|s_1\} = \mathbf{E}\left\{ \int_0^T -\frac{2}{T}\sqrt{E}\cos^2 \omega_0 t + n(t)\sqrt{\frac{2}{T}}\cos \omega_0 t\ dt \right\} = -\sqrt{E} \tag{4.25b}$$

where $\mathrm{E}\{\cdot\}$ denotes the ensemble average, referred to as the *expected value.* Equation (4.25) follows because $\mathbf{E}\{n(t)\} = 0$. The decision stage must decide which signal was transmitted by determining its location within the signal space. For this example, the choice of $\psi_1(t) = \sqrt{2/T}\ \cos \omega_0 t$ normalizes $\mathbf{E}\{z_i(T)\}$ to be $\pm \sqrt{E}$. The prototype signals $\{s_i(t)\}$ are the same as the reference signals $\{\psi_j(t)\}$ except for the normalizing scale factor. The decision stage chooses the signal with the largest value of $z_i(T)$. Thus, the received signal in this example is judged to be $s_1(t)$. The error performance for such coherently detected BPSK systems is treated in Section 4.7.1.

4.4.2 Sampled Matched Filter

In Section 3.2.2, we discussed the basic characteristic of the matched filter—namely, that its impulse response is a delayed version of the mirror image (rotated on the $t = 0$ axis) of the input signal waveform. Therefore, if the signal waveform is $s(t)$, its mirror image is $s(-t)$, and the mirror image delayed by T seconds is $s(T-t)$. The impulse response $h(t)$ of a filter matched to $s(t)$ is then described by

$$h(t) = \begin{cases} s(T-t) & 0 \le t \le T \\ 0 & \text{elsewhere} \end{cases} \tag{4.26}$$

Figures 4.7 and 4.8 illustrate the basic function of a correlator to product-integrate the received noisy signal with each of the candidate reference signals and determine the best match. The schematics in these figures imply the use of analog hardware (multipliers and integrators) and continuous signals. They do not reflect the way that the correlator or matched filter (MF) can be implemented using digital techniques and sampled waveforms. Figure 4.10 shows how an MF can be implemented using digital hardware. The input signal $r(t)$ comprises a prototype signal $s_i(t)$, plus noise $n(t)$, and the bandwidth of the signal is $W = 1/2T$, where T is the symbol time. Thus, the minimum Nyquist sampling rate is $f_s = 2W = 1/T$, and the sampling time T_s needs to be equal to or less than the symbol time. In other words, there must be at least one sample per symbol. In real systems, such sampling is usually performed at a rate that exceeds the Nyquist minimum by a factor of 4 or more. The only cost is processor speed, not transmission bandwidth. At the clock

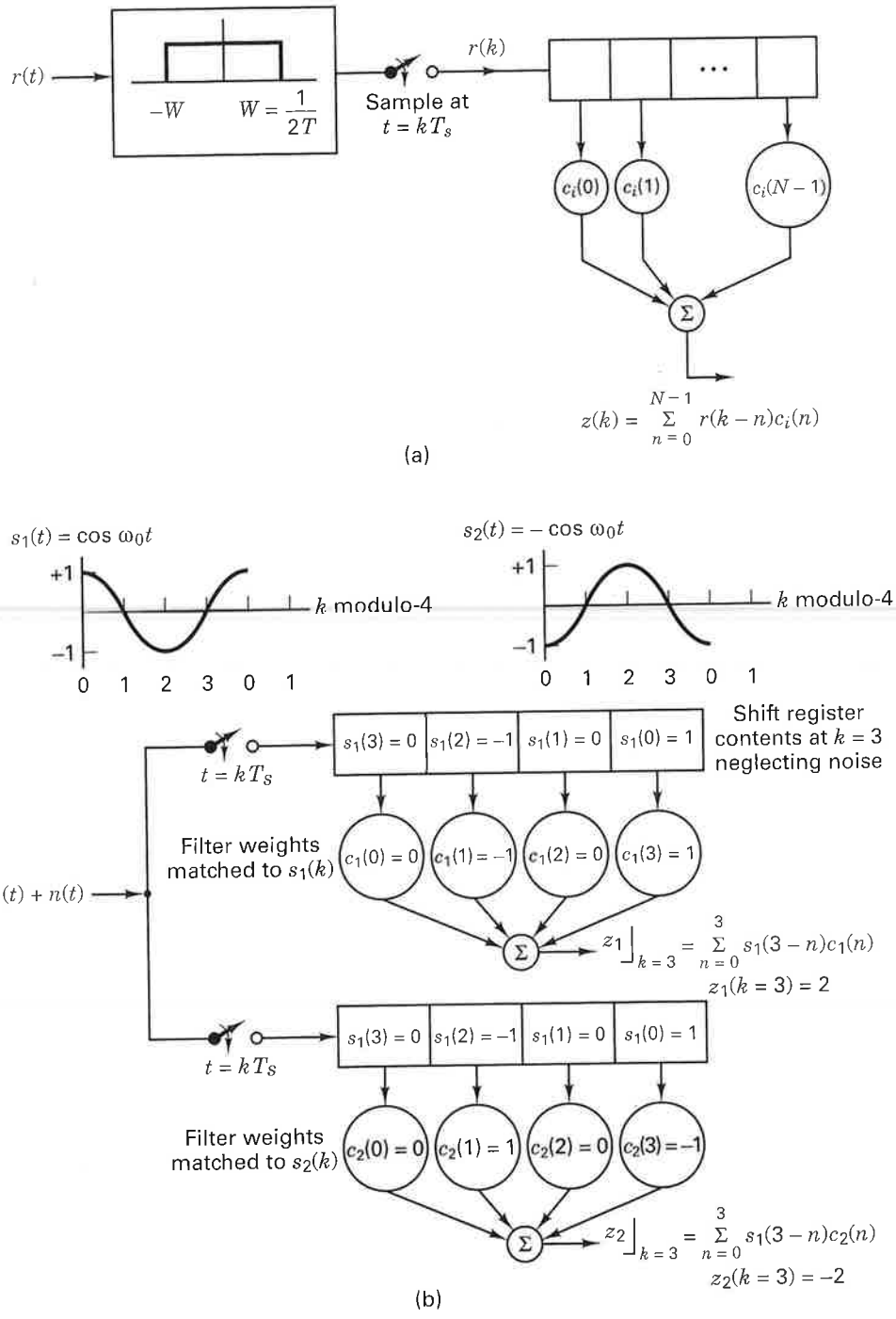


Figure 4.10 (a) Sampled matched filter. (b) Sampled matched filter detection example, neglecting noise.

times of $t = kT_s$, the samples are shifted into the register (as shown in Figure 4.10a) so that earlier samples are located to the right of later samples. Once the received signal has been sampled, the continuous time notation t is changed to kT_s or simply to k to allow use of the simple discrete notation

$$r(k) = s_i(k) + n(k) \qquad i = 1, 2 \qquad k = 0, 1, \ldots$$

where the index i identifies a particular symbol out of the M-ary set (binary in this example), and k is the sampling-time index. In Figure 4.10, the shift register, with its coefficients or weights $c_i(n)$ approximate an MF, where $n = 0, \ldots, N-1$ is the time index of the weights and register stages. In this example, $N = 4$ represents the number of stages in the register and the number of samples per symbol. Hence, the summation shown in the figure takes place over the time $n = 0$ to $n = 3$. Following the summer, a symbol decision will be made after 4 time samples have entered the registers. Note that for simplicity, the example in Figure 4.10b uses only 3-level amplitude values (0, ±1) for the $s_i(k)$ samples. In real systems, each sample (and weight) would be represented by 6–10 bits, depending on the density of the signal constellation. The set of filter weights $\{c_i(n)\}$ constitutes the filter impulse response; the weights are matched to signal samples according to a discrete form of Equation (4.26), as follows:

$$c_i(n) = s_i[(N-1) - n] = s_i(3 - n) \tag{4.27}$$

By using a discrete form of the *convolution integral* in Equation (A.44b), the output at a time corresponding to the kth sample can be expressed as

$$z_i(k) = \sum_{n=0}^{N-1} r(k-n)\, c_i(n) \qquad k = 0, 1, \ldots, \text{modulo-}N \tag{4.28}$$

where x modulo-y is defined as the remainder of dividing x by y, the index k represents time for both the received samples and the filter output, and n is a dummy time variable. In the expression $r(k-n)$ of Equation (4.28), it is helpful to think of n as the "age" of the sample (how long has it been sitting in the filter). In the expression $c_i(n)$, it is helpful to think of n as the address of the weight. We assume that because the system is synchronized, the symbol timing is known. Also, we assume the noise to have a zero mean, so that the expected value of a received sample is

$$\mathbf{E}\{r(k)\} = s_i(k) \qquad i = 1, 2$$

Thus, if $s_i(t)$ is transmitted, the expected matched filter output is

$$\mathbf{E}\{z_i(k)\} = \sum_{n=0}^{N-1} s_i(k-n)\, c_i(n) \qquad k = 0, 1, \ldots, \text{modulo-}N \tag{4.29}$$

In Figure 4.10b, where the prototype waveforms are plotted as a function of time, we see that the leftmost sample (amplitude equals + 1) of the $s_1(t)$ plot represents the earliest sample at time $k = 0$. Assuming that $s_1(t)$ was sent, and the noise is neglected for notational simplicity, we then denote the received samples of $r(k)$ as $s_1(k)$. As these samples fill the stages of the MF, at the end of each symbol time the

$k = 0$ sample is located in the rightmost stage of each register. In Equations (4.28) and (4.29), notice that the time indexes n of the reference weights are in reverse order compared with the time index $k - n$ of the samples, which is a key aspect of the convolution integral. The fact that the earliest time sample now corresponds to the rightmost weight will ensure a meaningful correlation. Even though we describe the mathematical operation of an MF to be *convolution* of a signal with the impulse response of the filter, the end result appears to be the *correlation* of a signal with a replica of that same signal. That is why it is valid to describe a correlator as an implementation of a matched filter.

In Figure 4.10b, detection will follow the MF in the usual way. For the binary decision, the $z_i(k)$ outputs are examined at each value of $k = N - 1$ corresponding to the end of a symbol. Under the condition that $s_1(t)$ had been transmitted and noise is neglected, we combine Equations (4.27) through (4.29) to express the correlator outputs at time $k = N - 1 = 3$ as

$$z_1(k=3) = \sum_{n=0}^{3} s_1(3 - n)\, c_1(n) = 2 \tag{4.30a}$$

and

$$z_2(k=3) = \sum_{n=0}^{3} s_1(3 - n)\, c_2(n) = -2 \tag{4.30b}$$

Since $z_1(k = 3)$ is greater than $z_2(k = 3)$, the detector chooses $s_1(t)$ as the transmitted symbol.

One might ask, "What is the difference between the MF in Figure 4.10b and the correlator in Figure 4.8?" In the case of the MF, a new output value is available in response to each new input sample; thus the output will be a time series such as the MF output seen in Figure 3.7b (a succession of increasing positive and negative correlations to an input sine wave). Such an MF output sequence can be equated to several correlators operating at different starting points of the input time series. Note that a correlator only computes an output once per symbol time, such as the value of the peak signal at time T in Figure 3.7b. If the timing of the MF and correlator are aligned, then their outputs at the end of a symbol time are identical. An important distinction between the MF and correlator is that since the correlator yields a single output value per symbol, it must have side information, such as the start and stop times over which the product integration should take place. If there are timing errors in the correlator, then the sampled output fed to the detector may be badly degraded. On the other hand, since the MF yields a *time series* of output values (reflecting time-shifted input samples multiplied by fixed weights), then with the use of additional circuitry, the best time for sampling the MF output can be learned.

Example 4.1 Sampled Matched Filter

Consider the waveform set

$$s_1(t) = At \qquad 0 \le t \le kT$$

and

$$s_2(t) = -At \qquad 0 \le t \le kT$$

where $k = 0, 1, 2, 3$.

Illustrate how a *sampled* matched filter as shown in Figure 4.10 can be used to detect a received signal, say $s_1(t)$, from this sawtooth waveform set in the absence of noise.

Solution

First, the waveform is sampled so that $s_1(t)$ is transformed into the set of samples $\{s_1(k)\}$. The sampled matched filter receiver will be shown with two branches, following the implementation in Figure 4.10b. The top branch is made up of shift registers and coefficients matched to the $\{s_1(k)\}$ sample points. The bottom branch is similarly matched to the $\{s_2(k)\}$ sample points. The four equally spaced sample points ($k = 0, 1, 2, 3$) for each of the $\{s_i(k)\}$ are as follows:

$$s_1(k=0) = 0 \qquad s_1(k=1) = \frac{A}{4} \qquad s_1(k=2) = \frac{A}{2} \qquad s_1(k=3) = \frac{3A}{4}$$

$$s_2(k=0) = 0 \qquad s_2(k=1) = -\frac{A}{4} \qquad s_2(k=2) = -\frac{A}{2} \qquad s_2(k=3) = -\frac{3A}{4}$$

The $c_i(n)$ coefficients represent the delayed mirror-image rotation of the signal to which the filter is matched. Therefore, $c_i(n) = s_i(N-1-n)$, where $n = 0, \ldots, N-1$, and we can write $c_i(0) = s_i(3)$, $c_i(1) = s_i(2)$, $c_i(2) = s_i(1)$, $c_i(3) = s_i(0)$.

Consider the top branch in Figure 4.10b. At the $k = 0$ clock time, the first sample $s_1(k=0) = 0$ enters the leftmost stage of each register. At the next clock time, the second sample $s_1(k=1) = A/4$ enters the leftmost stage of each register; at this same time the first sample has been shifted to the next right stage in each register, and so on. At the $k = 3$ clock time the sample $s_1(k=3) = 3A/4$ enters the leftmost stage; by this time the first sample has been shifted into the rightmost stage. The four signal samples are now located in the registers in mirror-image arrangement compared with how they would be plotted in time. Hence, the convolution operation is an appropriate expression for describing the alignment of the incoming waveform samples with the reference coefficients to maximize the correlation in the proper branch.

4.4.3 Coherent Detection of Multiple Phase-Shift Keying

Figure 4.11 illustrates the signal space for a multiple phase-shift keying (MPSK) signal set; the figure describes a four-level (4-ary) PSK or quadriphase shift keying (QPSK) example ($M = 4$). At the transmitter, binary digits are collected two at a time, and for each symbol interval, the two sequential digits instruct the modulator as to which of the four waveforms to produce. For typical coherent M-ary PSK (MPSK) systems, $s_i(t)$ can be expressed as

$$s_i(t) = \sqrt{\frac{2E}{T}} \cos\left(\omega_0 t - \frac{2\pi i}{M}\right) \qquad \begin{matrix} 0 \le t \le T \\ i = 1, \ldots, M \end{matrix} \tag{4.31}$$

where E is the received energy of such a waveform over each symbol duration T, and ω_0 is the carrier frequency. If an orthonormal signal space is assumed in Equations (3.10) and (3.11), we can choose a convenient set of axes, such as

$$\psi_1(t) = \sqrt{\frac{2}{T}} \cos \omega_0 t \tag{4.32a}$$

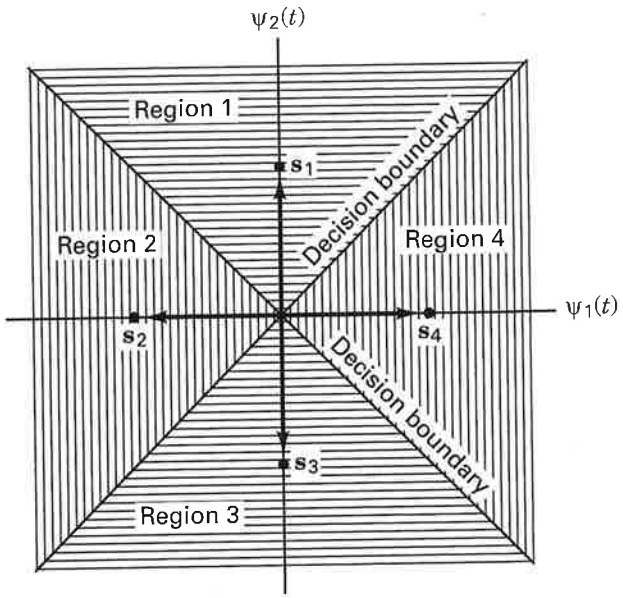


Figure 4.11 Signal space and decision regions for a QPSK system.

and

$$\psi_2(t) = \sqrt{\frac{2}{T}} \sin \omega_0 t \tag{4.32b}$$

where the amplitude $\sqrt{2/T}$ has been chosen to normalize the expected output of the detector, as was done in Section 4.4.1. Now $s_i(t)$ can be written in terms of these orthonormal coordinates, giving

$$s_i(t) = a_{i1}\psi_1(t) + a_{i2}\psi_2(t) \qquad \begin{matrix} 0 \le t \le T \\ i = 1, \ldots, M \end{matrix} \tag{4.33a}$$

$$= \sqrt{E} \cos\left(\frac{2\pi i}{M}\right) \psi_1(t) + \sqrt{E} \sin\left(\frac{2\pi i}{M}\right) \psi_2(t) \tag{4.33b}$$

Notice that Equation (4.33) describes a set of M multiple phase waveforms (intrinsically nonorthogonal) in terms of only two orthogonal carrier-wave components. The $M = 4$ (QPSK) case is unique among MPSK signal sets in the sense that the QPSK waveform set is represented by a combination of antipodal and orthogonal members. The decision boundaries partition the signal space into $M = 4$ regions; the construction is similar to the procedure outlined in Section 4.3.1 and Figure 4.6 for $M = 2$. The decision rule for the detector (see Figure 4.11) is to decide that $s_1(t)$ was transmitted if the received signal vector falls in region 1, that $s_2(t)$ was transmitted if the received signal vector falls in region 2, and so on. In other words, the decision rule is to choose the ith waveform if $z_i(T)$ is the largest of the correlator outputs (seen in Figure 4.7).

The form of the correlator shown in Figure 4.7a implies that there are always M product correlators used for the demodulation of MPSK signals. The figure infers that for each of the M branches, a reference signal with the appropriate phase shift is configured. In practice, the implementation of an MPSK demodulator follows Figure 4.7b, requiring only $N = 2$ product integrators regardless of the size of the signal set M. The savings in implementation is possible because any arbitrary integrable waveform set can be expressed as a linear combination of orthogonal waveforms, as shown in Section 3.1.3. Figure 4.12 illustrates such a demodulator. The received signal $r(t)$ can be expressed by combining Equations (4.32) and (4.33) as

$$r(t) = \sqrt{\frac{2E}{T}}(\cos \phi_i \cos \omega_0 t + \sin \phi_i \sin \omega_0 t) + n(t) \qquad \begin{matrix} 0 \le t \le T \\ i = 1, \ldots, M \end{matrix} \tag{4.34}$$

where $\phi_i = 2\pi i/M$, and $n(t)$ is a zero-mean white Gaussian noise process. Notice that in Figure 4.12, there are only two reference waveforms or basis functions, $\psi_1(t) = \sqrt{2/T} \cos \omega_0 t$ for the upper correlator and $\psi_2(t) = \sqrt{2/T} \sin \omega_0 t$ for the lower correlator. The upper correlator computes

$$X = \int_0^T r(t)\psi_1(t)\, dt \tag{4.35}$$

and the lower correlator computes

$$Y = \int_0^T r(t)\psi_2(t)\, dt \tag{4.36}$$

Figure 4.13 illustrates that the computation of the received phase angle $\hat{\phi}$ can be accomplished by computing the arctan of Y/X, where X can be thought of as the in-phase component of the received signal, Y is the quadrature component, and $\hat{\phi}$ is a

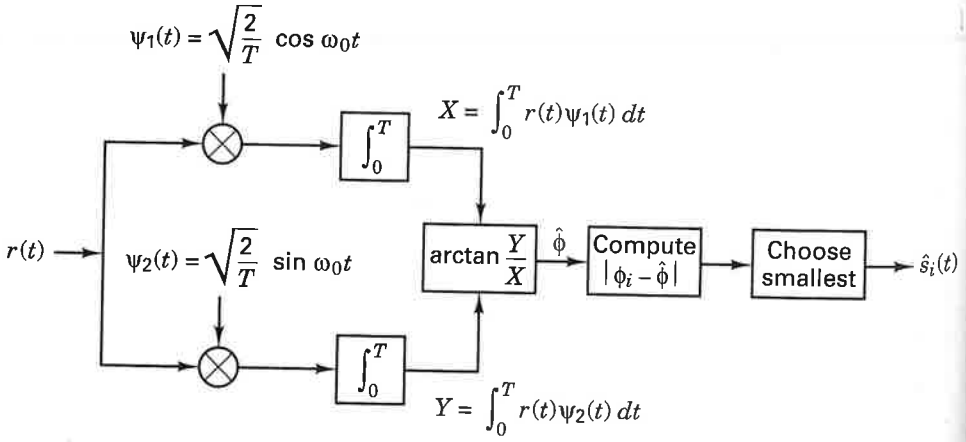


Figure 4.12 Demodulator for MPSK signals.

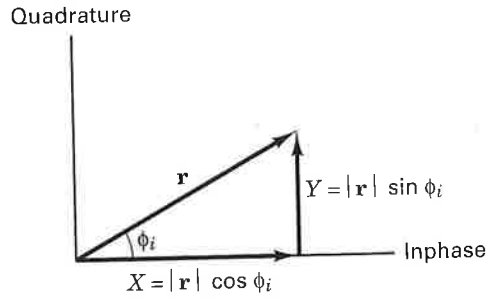


Figure 4.13 In-phase and quadrature components of the received signal vector **r**.

noisy estimate of the transmitted ϕ_i. In other words, the upper correlator of Figure 4.12 produces an output X, the magnitude of the in-phase projection of the vector **r**, and the lower correlator produces an output Y, the magnitude of the quadrature projection of the vector **r**, where **r** is the vector representation of $r(t)$. The X and Y outputs of the correlators feed into the block marked arctan (Y/X). The resulting value of the angle $\hat{\phi}$ is compared with each of the prototype phase angles, ϕ_i. The demodulator selects the ϕ_i that is closest to the angle $\hat{\phi}$. In other words, the demodulator computes $|\phi_i - \hat{\phi}|$ for each of the ϕ_i prototypes and chooses the ϕ_i yielding the smallest output.

4.4.4 Coherent Detection of FSK

FSK modulation is characterized by the information being contained in the frequency of the carrier. A typical set of FSK signal waveforms was described in Equation (4.8) as

$$s_i(t) = \sqrt{\frac{2E}{T}} \cos(\omega_i t + \phi) \qquad \begin{matrix} 0 \le t \le T \\ i = 1, \ldots, M \end{matrix}$$

where E is the energy content of $s_i(t)$ over each symbol duration T, and $(\omega_{i+1} - \omega_i)$ is typically assumed to be an integral multiple of π/T. The phase term ϕ is an arbitrary constant and can be set equal to zero. Assuming that the basis functions $\psi_1(t)$, $\psi_2(t), \ldots, \psi_N(t)$ form an orthonormal set, the most useful form for $\{\psi_j(t)\}$ is

$$\psi_j(t) = \sqrt{\frac{2}{T}} \cos \omega_j t \qquad j = 1, \ldots, N \tag{4.37}$$

where, as before, the amplitude $\sqrt{2/T}$ normalizes the expected output of the matched filter. From Equation (3.11), we can write

$$a_{ij} = \int_0^T \sqrt{\frac{2E}{T}} \cos(\omega_i t) \sqrt{\frac{2}{T}} \cos \omega_j t \, dt \tag{4.38}$$

Therefore,

$$a_{ij} = \begin{cases} \sqrt{E} & \text{for } i = j \\ 0 & \text{otherwise} \end{cases} \tag{4.39}$$

In other words, the ith prototype signal vector is located on the ith coordinate axis at a displacement $\sqrt{E}$ from the origin of the signal space. In this scheme, for the general M-ary case and a given E, the distance between any two prototype signal vectors $\mathbf{s}_i$ and $\mathbf{s}_j$ is constant:

$$d(\mathbf{s}_i, \mathbf{s}_j) = \|\mathbf{s}_i - \mathbf{s}_j\| = \sqrt{2E} \qquad \text{for } i \neq j \tag{4.40}$$

Figure 4.14 illustrates the prototype signal vectors and the decision regions for a 3-ary ($M = 3$) coherently detected orthogonal FSK signaling scheme. A natural choice for the size M of a signaling set is any power-of-two value. However in this case, the reasons for the unorthodox selection of $M = 3$, are: We desire to examine a signaling set that is greater than binary, and the signaling space for orthogonal signaling is best visualized as having mutually perpendicular axes. It is not possible, beyond 3-axes, to convey the notion of mutual perpendicularity in some visually accurate way. As in the PSK case, the signal space is partitioned into M distinct regions, each containing one prototype signal vector; in this example, because the decision region is three dimensional, the decision boundaries are planes instead of lines. The optimum decision rule is to decide that the transmitted signal belongs to the class whose index corresponds to the region where the received signal is

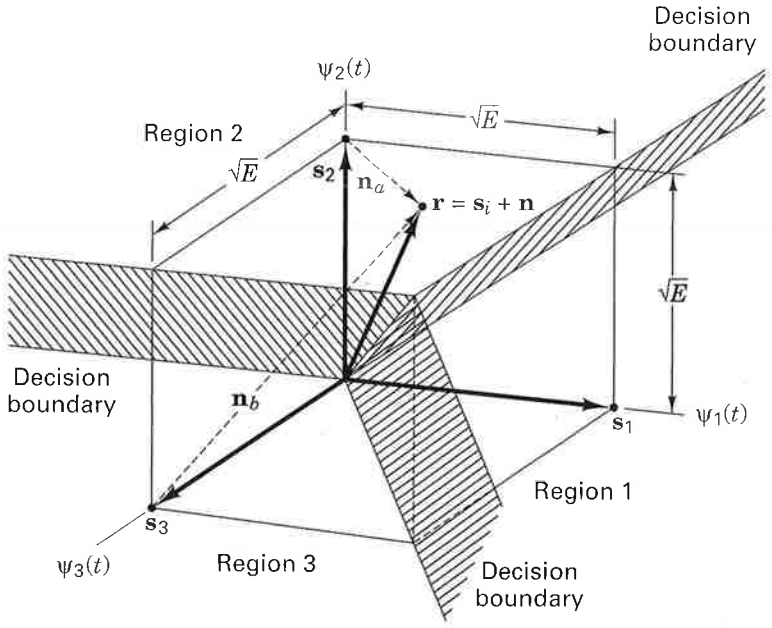


Figure 4.14 Partitioning the signal space for a 3-ary FSK signal.

found. In Figure 4.14, a received signal vector $\mathbf{r}$ is shown in region 2. Using the decision rule stated above, the detector classifies $\mathbf{r}$ as signal $\mathbf{s}_2$. Since the noise is a Gaussian random vector, there is a probability greater than zero that $\mathbf{r}$ could have been produced by some signal other than $\mathbf{s}_2$. For example, if the transmitter had sent $\mathbf{s}_2$, then $\mathbf{r}$ would be the sum of signal plus noise, $\mathbf{s}_2 + \mathbf{n}_a$, and the decision to choose $\mathbf{s}_2$ is correct; however, if the transmitter had actually sent $\mathbf{s}_3$, then $\mathbf{r}$ would be the sum of signal plus noise, $\mathbf{s}_3 + \mathbf{n}_b$, and the decision to select $\mathbf{s}_2$ is an error. The error performance of coherently detected FSK systems is treated in Section 4.7.3.

Example 4.2 Received Phase as a Function of Propagation Delay

(a) In Figure 4.8, the schematics fail to indicate where the correlator reference signals come from. One might think that they are known for all time and stored in memory until needed. Under some controlled circumstances, it is conceivable that the receiver might predict some expected value of the arriving signal's amplitude or its frequency within tolerable limits. But the one parameter that cannot be estimated without special help is the received signal phase. The most popular way to achieve phase estimation is through the use of circuitry called a *phase-locked loop* (PLL). This carrier-recovery method locks on to the arriving carrier wave (or recreates it) and estimates its phase. To show how futile it would be to predict the phase without a PLL, consider the mobile radio link shown in Figure 4.15. In the figure, a mobile user is positioned at point A, at a distance d from the base station, such that the propagation delay is T_d. Using complex notation, the transmitted waveform emanating from the transmitter can be described as $s(t) = \exp(j2\pi f_0 t)$. Let us take f_0 equal to 1 GHz. Neglecting noise, the waveform received at the base station can be denoted as $r(t) = \exp[j2\pi f_0(t + T_d)]$. If the mobile user moves in-line away from the base station to point B, or in-line toward the base station to point C, what is the minimum distance of movement that will cause a 2π phase rotation of the received waveform?

(b) Do we really care about a 2π phase rotation? Of course not, because the received phasor would then be located at exactly the point at which it was initially postulated with the user located at point A. But let us ask, What is the minimum distance that will cause a $\pi/2$ phase rotation (say a lag of $\pi/2$)? Had the receiver predicted that the received phasor would correspond to $r(t)$, as denoted in part (a),

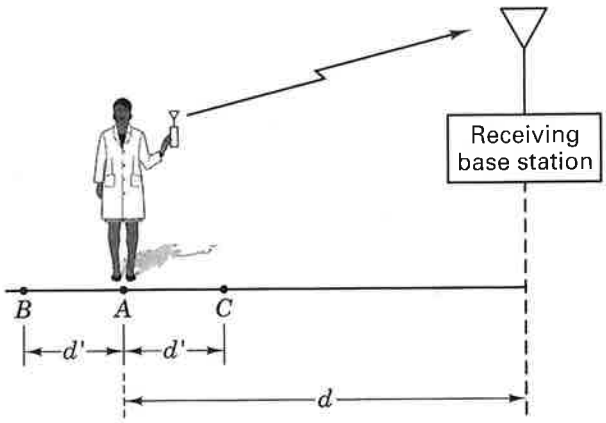


Figure 4.15 Mobile radio link.

but instead the lag caused the received phasor to appear as exp $[j2\pi f_0(t + T_d) - \pi/2]$, then the detecting correlator would yield a zero output. This is so because

$$\int_0^T \cos \omega_0 t \cos\left(\omega_0 t - \frac{\pi}{2}\right) dt = \int_0^T \cos \omega_0 t \sin \omega_0 t \, dt = 0$$

Find the user's minimum distance movement that will cause a $\pi/2$ phase rotation.

Solution

(a) Initially, let $t = 0$, so that when the mobile user is located at point A, the received phasor at the base station can be expressed as $r(t) = \exp(j2\pi f_0 T_d)$. Then, after the user's movement to point B, the received (further delayed) phasor $r_d(t = T_d + T_d')$, can be written as $r_d(t) = \exp[j2\pi f_0(T_d + T_d')]$. The minimum delay time $T_{d'}$ corresponding to a 2π (one wavelength) phasor rotation is $T_d' = 1/f_0 = 10^{-9}$ second. Therefore, the minimum distance for such a rotation (assuming ideal electromagnetic propagation at the speed of light) is

$$d' = \frac{c}{f_0} = 3 \times 10^8 \text{ m/s} \times 10^{-9} \text{ s} = 0.3 \text{ m}$$

(b) Thus, for a $\pi/2$ phasor rotation, the minimum distance is

$$d'' = \frac{d'}{4} = \frac{0.3 \text{ m}}{4} = 7.5 \text{ cm}$$

It should be clear that even if a transmitter and receiver are located on fixed towers, a small amount of wind movement can bring about complete uncertainty regarding phase. If we scale our example from a frequency of 1 GHz to that of 10 GHz, the minimum distance scales from 7.5 cm to 0.75 cm. Very often we might want to avoid building receivers with PLLs for carrier recovery. The results of this example might then motivate us to ask, How will the error performance suffer if phase information is not used in the detection process? In other words, how will the system fare if the detection is performed noncoherently? We address this question in the sections that follow.

4.5 NONCOHERENT DETECTION

4.5.1 Detection of Differential PSK

The name *differential* PSK (DPSK) sometimes needs clarification because two separate aspects of the modulation/demodulation format are being referred to: the encoding procedure and the detection procedure. The term *differential encoding* refers to the procedure of encoding the data differentially; that is, the presence of a binary one or zero is manifested by the symbol's similarity or difference when compared with the preceding symbol. The term *differentially coherent detection* of differentially encoded PSK, the usual meaning of DPSK, refers to a detection scheme often classified as noncoherent because it does not require a reference in phase with the received carrier. Occasionally, differentially encoded PSK is *coherently* detected. This will be discussed in Section 4.7.2.

With noncoherent systems, no attempt is made to determine the actual value of the phase of the incoming signal. Therefore, if the transmitted waveform is

$$s_i(t) = \sqrt{\frac{2E}{T}} \cos [\omega_0 t + \theta_i(t)] \qquad \begin{matrix} 0 \le t \le T \\ i = 1, \ldots, M \end{matrix}$$

the received signal can be characterized by

$$r(t) = \sqrt{\frac{2E}{T}} \cos [\omega_0 t + \theta_i(t) + \alpha] + n(t) \qquad \begin{matrix} 0 \le t \le T \\ i = 1, \ldots, M \end{matrix} \tag{4.41}$$

where α is an arbitrary constant and is typically assumed to be a random variable uniformly distributed between zero and 2π, and $n(t)$ is an AWGN process.

For coherent detection, matched filters (or their equivalents) are used; for noncoherent detection, this is not possible because the matched filter output is a function of the unknown angle α. However, if we assume that α varies slowly relative to two period times ($2T$), the phase difference between two successive waveforms $\theta_j(T_1)$ and $\theta_k(T_2)$ is independent of α; that is,

$$[\theta_k(T_2) + \alpha] - [\theta_j(T_1) + \alpha] = \theta_k(T_2) - \theta_j(T_1) = \phi_i(T_2) \tag{4.42}$$

The basis for *differentially coherent detection* of differentially encoded PSK (DPSK) is as follows. The carrier phase of the previous signaling interval can be used as a phase reference for demodulation. Its use requires *differential encoding* of the message sequence at the transmitter since the information is carried by the difference in phase between two successive waveforms. To send the ith message ($i = 1, 2, \ldots, M$), the present signal waveform must have its phase advanced by $\phi_i = 2\pi i/M$ radians over the previous waveform. The detector, in general, calculates the coordinates of the incoming signal by correlating it with locally generated waveforms, such as $\sqrt{2/T} \cos \omega_0 t$ and $\sqrt{2/T} \sin \omega_0 t$. the detector then measures the angle between the currently received signal vector and the previously received signal vector, as illustrated in Figure 4.16.

In general, DPSK signaling performs less efficiently than PSK, because the errors in DPSK tend to propagate (to adjacent symbol times) due to the correlation between signaling waveforms. One way of viewing the difference between PSK and

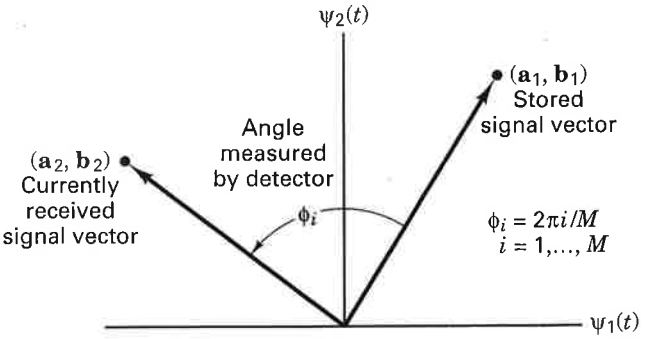


Figure 4.16 Signal space for DPSK.

DPSK is that the former compares the received signal with a clean reference; in the latter, however, two noisy signals are compared with each other. We might say that there is twice as much noise associated with DPSK signaling compared to PSK signaling. Consequently, as a first guess, we might estimate that the error probability for DPSK is approximately two times (3 dB) worse than PSK; this degradation decreases rapidly with increasing signal-to-noise ratio. The trade-off for this performance loss is reduced system complexity. The error performance for the detection of DPSK is treated in Section 4.7.5.

4.5.2 Binary Differential PSK Example

The essence of differentially coherent detection in DPSK is that the identity of the data is inferred from the changes in phase from symbol to symbol. Therefore, because the data are detected by differentially examining the waveform, the transmitted waveform must first be encoded in a differential fashion. Figure 4.17a illustrates a differential encoding of a binary message data stream $m(k)$, where k is the sample time index. The differential encoding starts (third row in the figure) with the first bit of the code-bit sequence $c(k = 0)$, chosen arbitrarily (here taken to be a one). Then the sequence of encoded bits $c(k)$ can, in general, be encoded in one of two ways:

$$c(k) = c(k - 1) \oplus m(k) \tag{4.43}$$

or

$$c(k) = \overline{c(k - 1) \oplus m(k)} \tag{4.44}$$

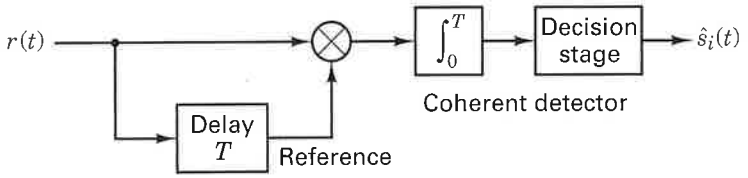
where the symbol $\oplus$ represents modulo-2 addition (defined in Section 2.9.3) and the overbar denotes complement. In Figure 4.17a the differentially encoded message was obtained by using Equation (4.44). In other words, the present code bit $c(k)$ is a one if the message bit $m(k)$ and the prior coded bit $c(k - 1)$ are the same, otherwise, $c(k)$ is a zero. The fourth row translates the coded bit sequence $c(k)$ into the phase shift sequence $\theta(k)$, where a one is characterized by a 180° phase shift, and a zero is characterized by a 0° phase shift.

Figure 4.17b illustrates the binary DPSK detection scheme in block diagram form. Notice that the basic product integrator of Figure 4.7 is the essence of the demodulator; as with coherent PSK, we are still attempting to correlate a received signal with a reference. The interesting difference here is that the reference signal is simply a delayed version of the received signal. In other words, during each symbol time, we are matching a received symbol with the prior symbol and looking for a correlation or an anticorrelation (180° out of phase).

Consider the received signal with phase shift sequence $\theta(k)$ entering the correlator of Figure 4.17b, in the absence of noise. The phase $\theta(k = 1)$ is matched with $\theta(k = 0)$; they have the same value, π; hence the first bit of the detected output is $\hat{m}(k = 1) = 1$. Then $\theta(k = 2)$ is matched with $\theta(k = 1)$; again they have the same value, and $\hat{m}(k = 2) = 1$. Then $\theta(k = 3)$ is matched with $\theta(k = 2)$; they are different, so that $\hat{m}(k = 3) = 0$, and so on.

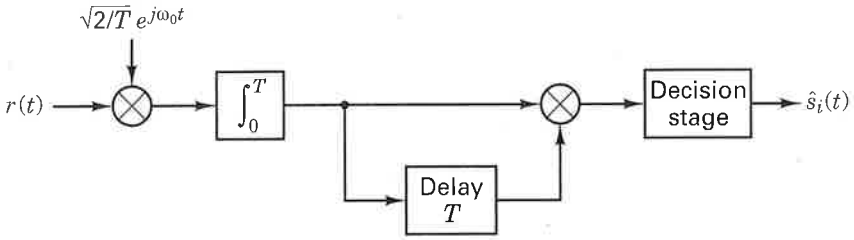
Sample index, k	0	1	2	3	4	5	6	7	8	9	10
Information message, $m(k)$		1	1	0	1	0	1	1	0	0	1
Differentially encoded message (first bit arbitrary), $c(k)$	1	1	1	0	0	1	1	1	0	1	1
Corresponding phase shift, $\theta(k)$	π	π	π	0	0	π	π	π	0	π	π

(a)



Detected message, $\hat{m}(k)$ 1 1 0 1 0 1 1 0 0 1

(b)



(c)

Figure 4.17 Differential PSK (DPSK). (a) Differential encoding. (b) Differentially coherent detection. (c) Optimum differentially coherent detection.

It must be pointed out that the detector in Figure 4.17b is suboptimum [3] in the sense of error performance. The optimum differential detector for DPSK requires a reference carrier in frequency but not necessarily in phase with the received carrier. Hence the optimum differential detector is shown in Figure 4.17c [4]. Its performance is treated in Section 4.7.5. In Figure 4.17c, the reference signal is shown in complex form ($\sqrt{2/T}\ e^{j\omega_0 t}$) to indicate that a quadrature implementation using both I and Q references are required (see Section 4.6.1).

4.5.3 Noncoherent Detection of FSK

A detector for the *noncoherent* detection of FSK waveforms described by Equation (4.8) can be implemented with correlators similar to those shown in Figure 4.7. However, the hardware must be configured as an *energy detector,* without exploiting phase measurements. For this reason, the noncoherent detector typically requires twice as many channel branches as the coherent detector. Figure 4.18 illustrates the in-phase (I) and quadrature (Q) channels used to detect a binary FSK (BFSK) signal set noncoherently. Notice that the upper two branches are configured to detect the signal with frequency ω_1; the reference signals are $\sqrt{2/T}\cos\omega_1 t$ for the I branch and $\sqrt{2/T}\sin\omega_1 t$ for the Q branch. Similarly, the lower two branches are configured to detect the signal with frequency ω_2; the reference signals are $\sqrt{2/T}\cos\omega_2 t$ for the I branch and $\sqrt{2/T}\sin\omega_2 t$ for the Q branch. Imagine that the received signal $r(t)$, by chance alone, is exactly of the form $\cos\omega_1 t + n(t)$; that is, the phase is exactly zero, and thus the signal component of the received signal exactly matches the top-branch reference signal with regard to frequency and phase. In that event, the product integrator of the top branch should yield the maximum output. The second branch should yield a near-zero output (integrated zero-mean noise), since its reference signal $\sqrt{2/T}\sin\omega_1 t$ is orthogonal to the sig-

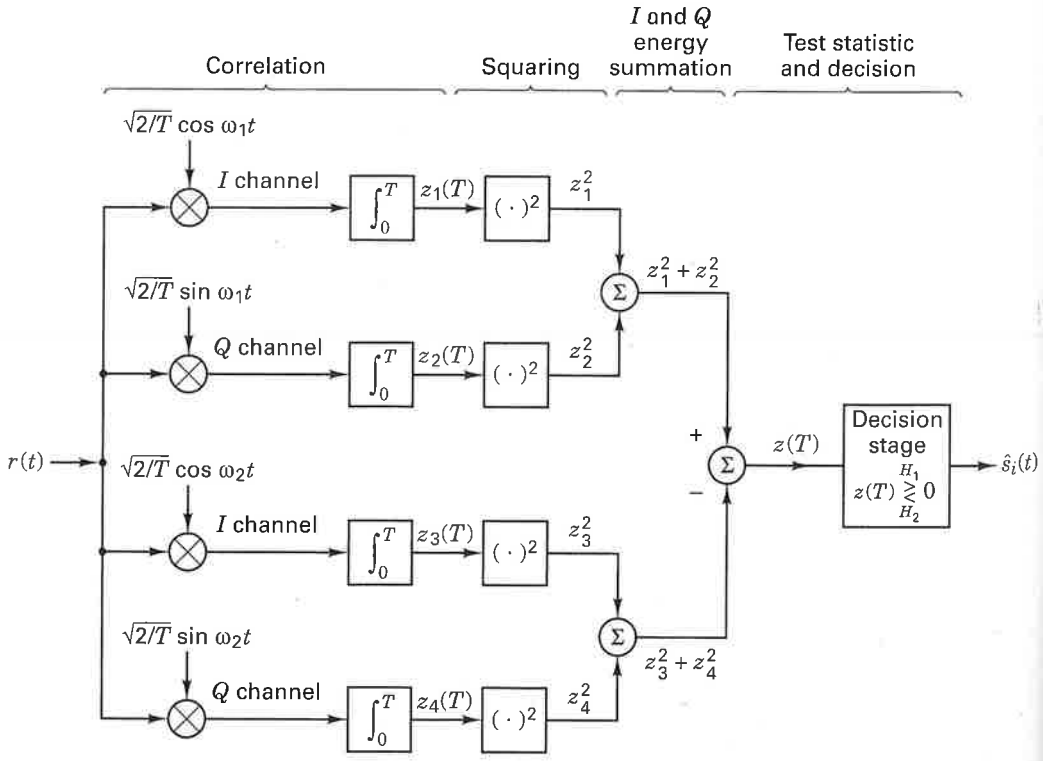


Figure 4.18 Quadrature receiver.

nal component of $r(t)$. For orthogonal signaling (see Section 4.5.4), the third and fourth branches should also yield near-zero outputs, since their ω_2 reference signals are also orthogonal to the signal component of $r(t)$.

Now imagine a different scenario: Suppose that by chance alone, the received signal $r(t)$ is of the form $\sin \omega_1 t + n(t)$. In that event, the second branch in Figure 4.18 should yield the maximum output, while the others should yield near-zero outputs. In actual practice, the most likely scenario is that $r(t)$ is of the form $\cos(\omega_1 t + \phi) + n(t)$; that is, the incoming signal will *partially* correlate with the $\cos \omega_1 t$ reference and *partially* correlate with the $\sin \omega_1 t$ reference. Hence a noncoherent quadrature receiver for orthogonal signals requires an I and Q branch for each candidate signal in the signaling set. In Figure 4.18, the blocks following the product integrators perform a squaring operation to prevent the appearance of any negative values. Then for each signal class in the set (two in this binary example), z_1^2 from the I channel and z_2^2 from the Q channel are added. The final stage forms the test statistic $z(T)$ and chooses the signal with frequency ω_1 or the signal with frequency ω_2, depending upon which pair of energy detectors yields the maximum output.

Another possible implementation for noncoherent FSK detection uses bandpass filters—centered at $f_i = \omega_i/2\pi$ with bandwidth $W_f = 1/T$—followed by *envelope detectors,* as shown in Figure 4.19. An envelope detector consists of a rectifier and a low-pass filter. The detectors are matched to the *signal envelopes* and not to the signals themselves. The phase of the carrier is of no importance in defining the envelope; hence no phase information is used. In the case of binary FSK, the decision as to whether a one or a zero was transmitted is made on the basis of which of two envelope detectors has the largest amplitude at the moment of measurement. Similarly, for a multiple frequency shift-keying (MFSK) system, the decision as to which of M signals was transmitted is made on the basis of which of the M envelope detectors has the maximum output.

Even though the envelope detector block diagram of Figure 4.19 looks functionally more simple than the quadrature receiver shown in Figure 4.18, the use of

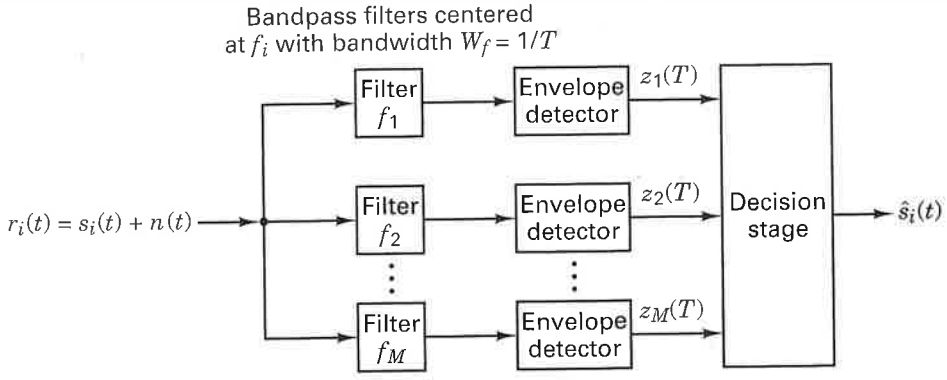


Figure 4.19 Noncoherent detection of FSK using envelope detectors.

(analog) filters usually results in the envelope detector design having greater weight and cost than the quadrature receiver. Quadrature receivers can be implemented digitally; thus, with the advent of large-scale integrated (LSI) circuits, they are often the preferred choice for noncoherent detectors. The detector shown in Figure 4.19 can also be implemented digitally by performing discrete Fourier transformations instead of using analog filters, but such a design is usually more complex than a digital implementation of the quadrature receiver.

4.5.4 Required Tone Spacing for Noncoherent Orthogonal FSK Signaling

Frequency shift keying (FSK) is usually implemented as orthogonal signaling, but not all FSK signaling is orthogonal. How can we tell if the tones in a signaling set form an orthogonal set? For example, suppose we have two tones $f_1 = 10{,}000$ Hz and $f_2 = 11{,}000$ Hz. Are they orthogonal to each other? In other words, do they fulfill the criterion (as set forth in Equation 3.69) that they are uncorrelated over a symbol time T? We do not have enough information to answer that question yet. Tones f_1 and f_2 manifest orthogonality if, for a transmitted tone at f_1, the sampled envelope of the receiver output filter tuned to f_2 is zero (i.e., no crosstalk). A property that insures such orthogonality between tones in an FSK signaling set states that any pair of tones in the set must have a frequency separation that is a multiple of $1/T$ hertz. (This will be proven below, in Example 4.3.) A tone with frequency f_i that is switched on for a symbol duration of T seconds and then switched off, such as the FSK tone described in Equation (4.8), can be analytically described by

$$s_i(t) = (\cos 2\pi f_i t) \text{ rect } (t/T)$$

$$\text{where rect } (t/T) = \begin{cases} 1 & \text{for } -T/2 \leq t \leq T/2 \\ 0 & \text{for } |t| > T/2 \end{cases}$$

From Table A.1, the Fourier transform of $s_i(t)$ is

$$\mathcal{F}\{s_i(t)\} = T \operatorname{sinc}(f - f_i)T$$

where the sinc function is as defined in Equation (1.39). The spectra of two such adjacent tones—tone 1 with frequency f_1 and tone 2 with frequency f_2—are plotted in Figure 4.20.

4.5.4.1 Minimum Tone Spacing and Bandwidth

In order for a noncoherently detected tone to manifest a maximum output signal at its associated receiver filter and a zero-output signal at any neighboring filter (as implemented in Figure 4.19), the peak of the spectrum of tone 1 must coincide with one of the zero crossings of the spectrum of tone 2, and similarly, the peak of the spectrum of tone 2 must coincide with one of the zero crossings of the spectrum of tone 1. The frequency difference between the center of the spectral main lobe and the first zero crossing represents the *minimum required spacing*. With noncoherent detection, this corresponds to a minimum tone separation of $1/T$ hertz, as can be seen in Figure 4.20. Even though FSK signaling entails the trans-

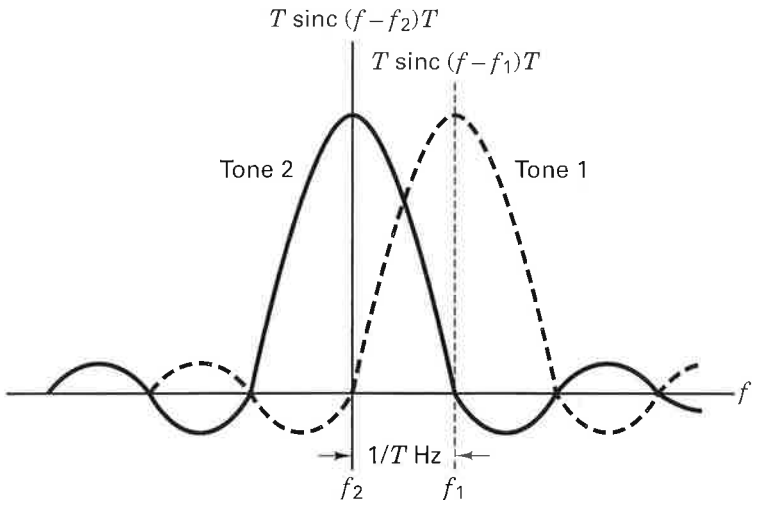


Figure 4.20 Minimum tone spacing for noncoherently detected orthogonal FSK signaling.

mission of just one single-sideband tone during each symbol time, when we speak of the signal bandwidth, we mean the amount of spectrum that needs to be available for the entire range of all the tones in the M-ary set. Thus for FSK, the bandwidth requirements are related to the spectral spacing between the tones. From a group of adjacent tones, the spectrum associated with just one tone can be considered to make up one-half of the tone spacing on each side of the tone. Thus, for the binary FSK case shown in Figure 4.20, the signaling bandwidth is equal to the spectrum that separates the tone centers plus one-half the tone spacing on each side of the set. This amounts to two times the tone spacing. Extrapolating this approach to the M-ary case, the bandwidth of noncoherently detected orthogonal MFSK is equal to M/T.

Thus far, we have only addressed noncoherently detected orthogonal FSK. Is the minimum tone-spacing criterion (and hence the bandwidth) different if the signaling is detected coherently? Indeed, it is. As we will see below, in Example 4.3, the minimum tone spacing is reduced to $1/2T$ when coherent detection is used.

4.5.4.2 Dual Relationships

The engineering concept of duality can be defined as follows: Two processes (functions, elements, or systems) are *dual* to each other if their mathematical relationships are identical, even though they are described with different variables (e.g., time and frequency). Consider the case of FSK signaling, where rectangular-shaped keying corresponds to sinc (fT) shaped tones, as seen in Figure 4.20. A given tone duration gives rise to the minimum frequency spacing between tones that would be needed to achieve orthogonality. This frequency-domain relationship has as its dual counterpart a time-domain relationship—the case of pusle sig-

naling, where a rectangular-shaped bandwidth corresponds to sinc (t/T) shaped pulses, as seen in Figure 3.16b. A given bandwidth gives rise to the minimum time spacing between pulses that would be needed to achieve zero ISI.

Example 4.3 Minimum Tone Spacing for Orthogonal FSK

Consider two waveforms $\cos(2\pi f_1 t + \phi)$ and $\cos 2\pi f_2 t$ to be used for *noncoherent* FSK-signaling, where $f_1 > f_2$. The symbol rate is equal to $1/T$ symbols/s, where T is the symbol duration and ϕ is a constant arbitrary angle from 0 to 2π.

(a) Prove that the minimum tone spacing for *noncoherently detected* orthogonal FSK-signaling is $1/T$.

(b) What is the minimum tone spacing for *coherently detected* orthogonal FSK signaling?

Solution

(a) For the two waveforms to be orthogonal, they must fulfill the orthogonality constraint of Equation (3.69):

$$\int_0^T \cos(2\pi f_1 t + \phi) \cos 2\pi f_2 t \, dt = 0 \tag{4.45}$$

Using the basic trigonometric identities shown in Equations (D.6) and (D.1) to (D.3), we can write Equation (4.45) as

$$\cos\phi \int_0^t \cos 2\pi f_1 t \cos 2\pi f_2 t \, dt - \sin\phi \int_0^T \sin 2\pi f_1 t \cos 2\pi f_2 t \, dt = 0 \tag{4.46}$$

so that

$$\cos\phi \int_0^T [\cos 2\pi(f_1 + f_2)t + \cos 2\pi(f_1 - f_2)t] \, dt - \sin\phi \int_0^T [\sin 2\pi(f_1 + f_2)t + \sin 2\pi(f_1 - f_2)t] \, dt = 0 \tag{4.47}$$

which, in turn, yields

$$\cos\phi \left[\frac{\sin 2\pi(f_1 + f_2)t}{2\pi(f_1 + f_2)} + \frac{\sin 2\pi(f_1 - f_2)t}{2\pi(f_1 - f_2)} \right]_0^T + \sin\phi \left[\frac{\cos 2\pi(f_1 + f_2)t}{2\pi(f_1 + f_2)} + \frac{\cos 2\pi(f_1 - f_2)t}{2\pi(f_1 - f_2)} \right]_0^T = 0 \tag{4.48}$$

or

$$\cos\phi \left[\frac{\sin 2\pi(f_1 + f_2)T}{2\pi(f_1 + f_2)} + \frac{\sin 2\pi(f_1 - f_2)T}{2\pi(f_1 - f_2)} \right] + \sin\phi \left[\frac{\cos 2\pi(f_1 + f_2)T - 1}{2\pi(f_1 + f_2)} + \frac{\cos 2\pi(f_1 - f_2)T - 1}{2\pi(f_1 - f_2)} \right] = 0 \tag{4.49}$$

We can assume that $f_1 + f_2 \gg 1$ and can thus make the following approximation:

$$\frac{\sin 2\pi(f_1 + f_2)T}{2\pi(f_1 + f_2)} \approx \frac{\cos 2\pi(f_1 + f_2)T}{2\pi(f_1 + f_2)} \approx 0 \tag{4.50}$$

Then, combining Equations (4.49) and (4.50), we can write

$$\cos\phi \sin 2\pi(f_1 - f_2)T + \sin\phi\,[\cos 2\pi(f_1 - f_2)T - 1] \approx 0 \tag{4.51}$$

Note that for arbitrary ϕ, the terms in Equation (4.51) can sum to zero only when $\sin 2\pi(f_1 - f_2)T = 0$, and simultaneously $\cos 2\pi(f_1 - f_2)T = 1$.

Since

$$\sin x = 0 \qquad \text{for } x = n\pi$$

and

$$\cos x = 1 \qquad \text{for } x = 2k\pi$$

where n and k are integers, then both $\sin x = 0$ and $\cos x = 1$ occur simultaneously when $n = 2k$. From Equation (4.51), for arbitrary ϕ, we can therefore write

$$2\pi(f_1 - f_2)T = 2k\pi$$

or

$$f_1 - f_2 = \frac{k}{T} \tag{4.52}$$

Thus the minimum tone spacing for *noncoherent* FSK signaling occurs for $k = 1$, in which case we write

$$f_1 - f_2 = \frac{1}{T} \tag{4.53}$$

Recall the question posed earlier. Having two tones f_1 = 10,000 hertz and f_2 = 11,000 hertz, we asked, Are they orthogonal? Now we have sufficient information to answer that question. The answer depends on the speed of the FSK signaling. If the tones are being keyed (switched) at the rate of 1,000 symbols/s and noncoherently detected, then they are orthogonal. If they are being keyed faster—say, at the rate of 10,000 symbols/s—they are not orthogonal.

(b) For the noncoherent detection in part (a) of this example, the tone spacings that rendered the signals orthogonal was found by satisfying Equation (4.45) for any arbitrary phase. In this case, however, for coherent detection, the tone spacings needed for orthogonality are found by setting $\phi = 0$. This is because we know the phase of the received signal (from our phase-locked loop estimate). This received signal will be correlated with each reference signal using the same phase estimate for the reference signals. We can now rewrite Equation (4.51) with $\phi = 0$, which gives

$$\sin 2\pi(f_1 - f_2)T = 0 \tag{4.54}$$

or

$$f_1 - f_2 = \frac{n}{2T} \tag{4.55}$$

Thus the minimum tone spacing for *coherent* FSK signaling occurs for $n = 1$ as follows:

$$f_1 - f_2 = \frac{1}{2T} \tag{4.56}$$

Therefore, for the same symbol rate, coherently detected FSK can occupy less bandwidth than noncoherently detected FSK and still retain orthogonal signaling. We can say that coherent FSK is more *bandwidth efficient.* (The subject of bandwidth efficiency is addressed in greater detail in Chapter 9.)

The required tone spacings are now closer than in part (a) because when we align two periodic waveforms so that their starting phases are the same, we achieve orthogonality by virtue of an even-versus-odd symmetry in the respective waveforms over one symbol time. This is unlike the way orthogonality was achieved in part (a), where we paid no attention to phase. In the coherent case here, the phase alignment in the correlator stages means that we can bring the tones closer together in frequency and still maintain orthogonality among the set of FSK tones. Prove it to yourself by plotting two sine waves (or cosine waves or square waves). Start them off at the same phase (0 radians is convenient). Using quadrille paper, choose a simple time scale to represent one symbol time T. Then, plot a tone at one cycle per T. Below that, with the same starting phase, plot another tone at one-and-a-half cycles per T. Perform product-summation over period T, and verify that these waveforms (spaced $1/2T$ hertz apart) are indeed orthogonal.

4.6 COMPLEX ENVELOPE

The description of real-world modulators and demodulators is facilitated by the use of complex notation which began in Section 4.2.1 and continues here. Any real bandpass waveform $s(t)$ can be represented using complex notation as

$$s(t) = \text{Re}\{g(t)e^{j\omega_0 t}\} \tag{4.57}$$

where $g(t)$ is known as the *complex envelope,* expressed as

$$g(t) = x(t) + jy(t) = |g(t)|e^{j\theta(t)} = R(t)e^{j\theta(t)} \tag{4.58}$$

The magnitude of the complex envelope is then

$$R(t) = |g(t)| = \sqrt{x^2(t) + y^2(t)} \tag{4.59}$$

and its phase is

$$\theta(t) = \tan^{-1}\frac{y(t)}{x(t)} \tag{4.60}$$

With respect to Equation (4.57), we can call $g(t)$ the baseband message or data in complex form, and $e^{j\omega_0 t}$ the carrier wave in complex form. The product of these two represents modulation, and $s(t)$, the real part of this product, is the transmitted waveform. Therefore, using Equations (4.4), (4.57), and (4.58), we can express $s(t)$ as follows:

$$s(t) = \text{Re}\{[x(t) + jy(t)][\cos \omega_0 t + j \sin \omega_0 t]\}$$
$$= x(t) \cos \omega_0 t - y(t) \sin \omega_0 t \tag{4.61}$$

Note that the modulation of signals expressed in the general form of $(a + jb)$ times $(c + jd)$ yields a waveform with a sign inversion (at the quadrature term of the carrier wave) of the form *ac-bd.*

4.6.1 Quadrature Implementation of a Modulator

Consider an example of a baseband waveform $g(t)$, described by a sequence of ideal pulses $x(t)$ and $y(t)$ appearing at discrete times $k = 1, 2, \ldots$. Thus, $g(t)$, $x(t)$, and $y(t)$ in Equation (4.58) can be written as g_k, x_k, and y_k, respectively. Let the pulse-amplitude values be $x_k = y_k = 0.707A$. The complex envelope can then be expressed in discrete form as

$$g_k = x_k + jy_k = 0.707A + j0.707A \tag{4.62}$$

We know from complex algebra that $j = \sqrt{-1}$, but from a practical point of view, the j term can be regarded as a "flag," reminding us that we may not use ordinary addition in combing the terms in Equation (4.62). In preparation for the next step, inphase and quadrature modulation, x_k and y_k are treated as an ordered pair. A quadrature-type modulator is shown in Figure 4.21, where we see that the x_k pulse is multiplied by $\cos \omega_0 t$ (inphase component of the carrier wave), and the y_k pulse is multiplied by $\sin \omega_0 t$ (quadrature component of the carrier wave). The modulation process can be described succinctly as multiplying the complex envelope by $e^{j\omega_0 t}$, and then transmitting the real part of the product. Hence, we write

$$\begin{aligned} s(t) &= \text{Re}\{g_k e^{j\omega_0 t}\} \\ &= \text{Re}\{(x_k + jy_k)(\cos \omega_0 t + j \sin \omega_0 t)\} \\ &= x_k \cos \omega_0 t - y_k \sin \omega_0 t \\ &= 0.707A \cos \omega_0 t - 0.707A \sin \omega_0 t \\ &= A \cos\left(\omega_0 t + \frac{\pi}{4}\right) \end{aligned} \tag{4.63}$$

Again, note that the quadrature term of the carrier wave has undergone a sign inversion in the modulation process. If we use $0.707A \cos \omega_0 t$ as a reference, then the transmitted waveform $s(t)$ in Equation (4.63) leads the reference waveform by $\pi/4$.

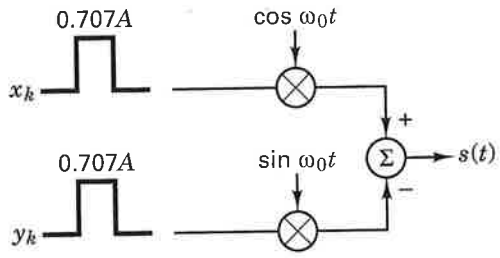


Figure 4.21 Quadrature type modulator.

Similarly, if we consider $-0.707A \sin \omega_0 t$ as the reference, then $s(t)$ in Equation (4.63) lags the reference waveform by $\pi/4$. These relationships are illustrated in Figure 4.22.

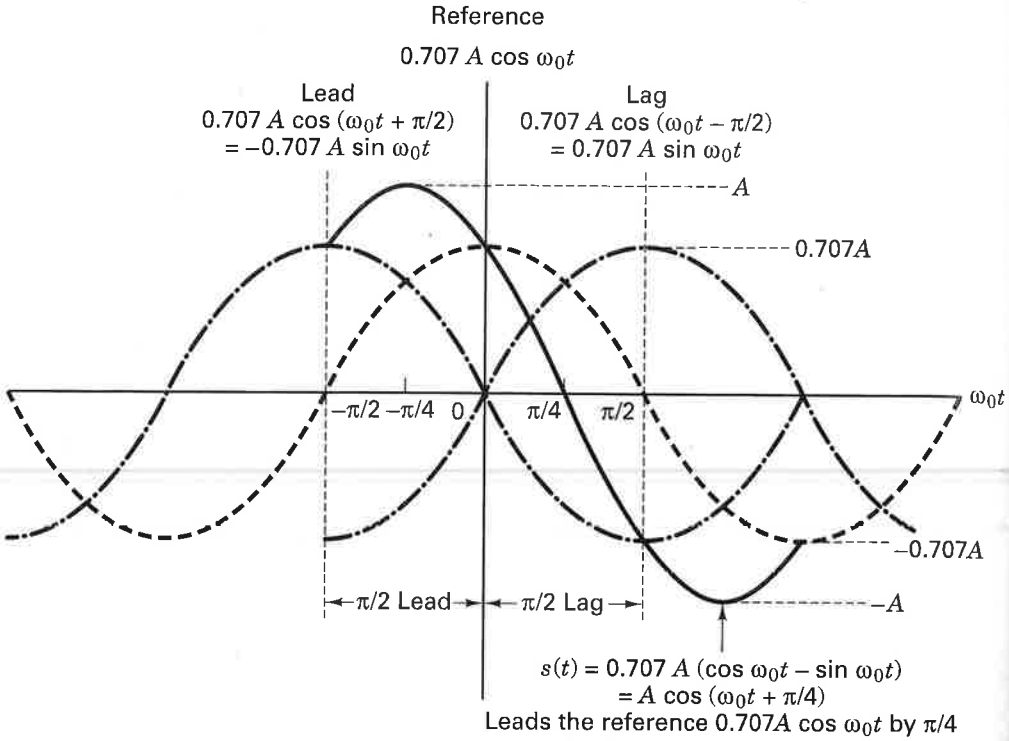


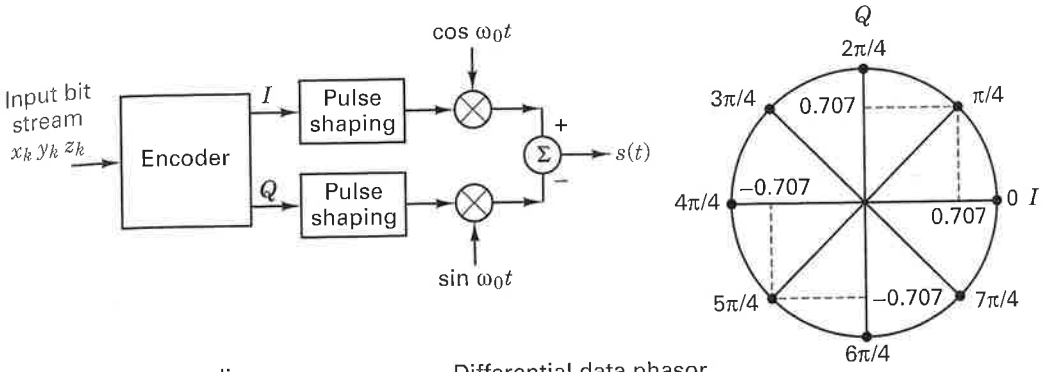
Figure 4.22 Lead/Lag relationships of sinusoids.

4.6.2 D8PSK Modulator Example

Figure 4.23 depicts a quadrature implementation of a differential 8-PSK (D8PSK) modulator. Because the modulation is 8-ary, we assign a 3-bit message (x_k, y_k, z_k) to each phase $\Delta\phi_k$. Because the modulation is differential, at each kth transmission time we send a data phasor ϕ_k, which can be expressed as

$$\phi_k = \Delta\phi_k + \phi_{k-1} \tag{4.64}$$

The process of adding the current message-to-phase assignment $\Delta\phi_k$ to the prior data phase ϕ_{k-1} provides for the differential encoding of the message. A sequence of phasors created by following Equation (4.64) yields a similar differential encoding as the procedures described in Section 4.5.2. You might notice in Figure 4.23 that the assignment of 3-bit message sequences to $\Delta\phi_k$ does not proceed along the natural binary progression from 000 to 111. There is a special code being used here called a *Gray code.* (The benefits that such a binary to M-ary assignment provides is explained in Section 4.9.4.)



Data encoding

x_k	y_k	z_k	$\Delta\phi_k$
0	0	0	0
0	0	1	π/4
0	1	1	2π/4
0	1	0	3π/4
1	1	0	4π/4
1	1	1	5π/4
1	0	1	6π/4
1	0	0	7π/4

Differential data phasor

$\phi_k = \phi_{k-1} + \Delta\phi_k$

Let $\phi_0 = 0$	$k = 1$	$k = 2$	$k = 3$	$k = 4$
$x_k\, y_k\, z_k$	110	001	110	010
$\Delta\phi_k$:	π	π/4	π	3π/4
ϕ_k:	π	5π/4	π/4	π
I:	−1	−0.707	0.707	−1
Q:	0	−0.707	0.707	0

Figure 4.23 Quadrature implementation of a D8PSK modulator.

For the modulator in Figure 4.23, let the input data sequence at times $k = 1, 2, 3, 4$, be equal to 110, 001, 110, 010, respectively. Next, we use the date-encoding table shown in Figure 4.23 and Equation (4.64), with the starting phase at time $k = 0$ to be $\phi_0 = 0$. At time $k = 1$, the differential data phase corresponding to $x_1\, y_1\, z_1 = 110$ is $\phi_1 = 4\pi/4 = \pi$. Taking the magnitude of the rotating phasor to be unity, the inphase (I) and the quadrature (Q) baseband pulses are –1 and 0, respectively. As indicated in Figure 4.23, these pulses are generally shaped with a filter (such as a root-raised-cosine).

For time $k = 2$, the table in Figure 4.23 shows that the message 001 is assigned $\Delta\phi_2 = \pi/4$. Therefore, following Equation (4.64), the second differential data phase is $\phi_2 = \pi + \pi/4 = 5\pi/4$, and for time $k = 2$, the I and Q baseband pulses are $x_k = -0.707$ and $y_k = -0.707$, respectively. The transmitted waveform follows the form of Equation (4.61), rewritten as

$$\begin{aligned} s(t) &= \text{Re}\{(x_k + jy_k)(\cos\omega_0 t + j\sin\omega_0 t)\} \\ &= x_k \cos\omega_0 t - y_k \sin\omega_0 t \end{aligned} \tag{4.65}$$

For a signaling set that can be represented on a phase-amplitude plane, such as MPSK or MQAM, Equation (4.65) provides an interesting observation. That is, quadrature implementation of the transmitter transforms all such signaling types

to simple amplitude modulation. Any phasor on the plane is transmitted by amplitude-modulating its inphase and quadrature projections onto the cosine and sine wave components of the carrier, respectively. For ease of notation, we neglect the pulse shaping; that is, we assume that the data pulses have ideal rectangular shapes. Then, using Equation (4.65), for time $k = 2$, where $x_k = -0.707$ and $y_k = -0.707$, the transmitted $s(t)$ can be written as follows:

$$\begin{aligned} s(t) &= -0.707 \cos \omega_0 t + 0.707 \sin \omega_0 t \\ &= \sin\left(\omega_0 t - \frac{\pi}{4}\right) \end{aligned} \tag{4.66}$$

4.6.3 D8PSK Demodulator Example

In the previous section, the quadrature implementation of a modulator began with multiplying the complex envelope (baseband message) by $e^{j\omega_0 t}$, and transmitting the real part of the product $s(t)$, as described in Equation (4.63). Using a similar quadrature implementation, demodulation consists of reversing the process—that is, multiplying the received bandpass waveform by $e^{-j\omega_0 t}$ in order to recover the baseband waveform. The left side of Figure 4.24 shows the modulator of Figure 4.23 in simplified form, and the waveform $s(t) = \sin(\omega_0 t - \pi/4)$ that was launched at time $k = 2$ for that example. We continue this same example in this section, and also show a quadrature implementation of the demodulator on the right side of Figure 4.24.

Notice the subtle difference between the $-\sin \omega_0 t$ term at the modulator and the $-\sin \omega_0 t$ term at the demodulator. At the modulator, the minus sign stems from taking the real part of the complex waveform (product of the complex envelope and complex carrier wave). At the demodulator, $-\sin \omega_0 t$ stems from multiplying the bandpass waveform by the conjugate $e^{-j\omega_0 t}$ of the modulator carrier wave; demodulation is coherent if phase is recovered. To simplify writing the basic relationships of the process, the noise is neglected. After the inphase multiplication by $\cos \omega_0 t$ in the demodulator, we get the signal at point A:

$$\begin{aligned} A &= (-0.707 \cos \omega_0 t + 0.707 \sin \omega_0 t) \cos \omega_0 t \\ &= -0.707 \cos^2 \omega_0 t + 0.707 \sin \omega_0 t \cos \omega_0 t \end{aligned} \tag{4.67}$$

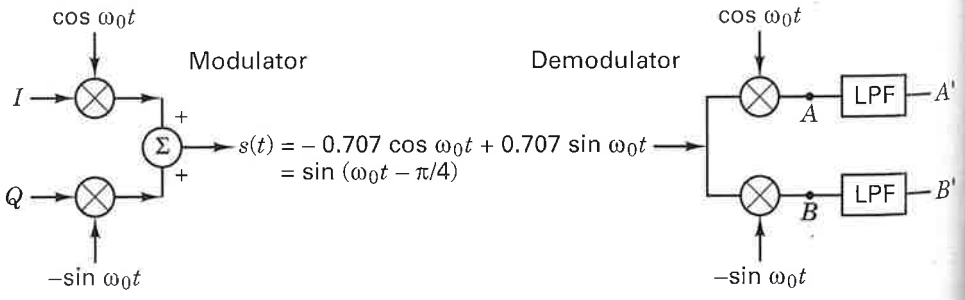


Figure 4.24 Modulator/demodulator example.

Using the trigonometric identities shown in Equation (D.7) and (D.9), we get

$$A = \frac{-0.707}{2}(1 + \cos 2\omega_0 t) + \frac{0.707}{2}\sin 2\omega_0 t \tag{4.68}$$

After filtering with a low-pass filter (LPF) we recover at point A' an ideal negative pulse, as follows:

$$A' = -0.707 \text{ (times a scale factor)} \tag{4.69}$$

Similarly, after the quadrature multiplication by $-\sin \omega_0 t$ in the demodulator, we get the signal at point B, as follows:

$$\begin{aligned} B &= (-0.707 \cos \omega_0 t + 0.707 \sin \omega_0 t)(-\sin \omega_0 t) \\ &= \frac{0.707}{2}\sin 2\omega_0 t - \frac{0.707}{2}(1 - \cos 2\omega_0 t) \end{aligned} \tag{4.70}$$

After the LPF we recover at point B' an ideal negative pulse, as follows:

$$B' = -0.707 \text{ (times a scale factor)} \tag{4.71}$$

Thus, we see from the demodulated and filtered points A' and B' that the differential (ideal) data pulses for the I and Q channels are each equal to -0.707. Since the modulator/demodulator is differential, then for this $k = 2$ example,

$$\Delta\phi_{k=2} = \phi_{k=2} - \phi_{k-1} \tag{4.72}$$

We presume that the demodulator at the earlier time $k = 1$ had properly recovered the signal phase to be π. Then, from Equation (4.72), we can write

$$\Delta\phi_{k=2} = \frac{5\pi}{4} - \pi = \frac{\pi}{4} \tag{4.73}$$

Referring back to the data-encoding table in Figure 4.23, we see that the detected data sequence is $x_2\, y_2\, z_2 = 001$, which corresponds to the data that was sent at time $k = 2$.

4.7 ERROR PERFORMANCE FOR BINARY SYSTEMS

4.7.1 Probability of Bit Error for Coherently Detected BPSK

An important measure of performance used for comparing digital modulation schemes is the probability of error, P_E. For the correlator or matched filter, the calculations for obtaining P_E can be viewed geometrically. (See Figure 4.6.) They involve finding the probability that given a particular transmitted signal vector, say $\mathbf{s}_1$, the noise vector $\mathbf{n}$ will give rise to a received signal falling outside region 1. The probability of the detector making an incorrect decision is termed the *probability of symbol error* (P_E). It is often convenient to specify system performance by the probability of bit error (P_B), even when decisions are made on the basis of symbols

for which $M > 2$. The relationship between P_B and P_E is treated in Section 4.9.3 for orthogonal signaling and in Section 4.9.4 for multiple phase signaling.

For convenience, this section is restricted to the coherent detection of BPSK modulation. For this case the symbol error probability is the bit error probability. Assume that the signals are equally likely. Also assume that when signal $s_i(t)$ $(i = 1, 2)$ is transmitted, the received signal $r(t)$ is equal to $s_i(t) + n(t)$, where $n(t)$ is an AWGN process, and any degradation effects due to channel-induced ISI or circuit-induced ISI have been neglected. The antipodal signals $s_1(t)$ and $s_2(t)$ can be characterized in a one-dimensional signal space as described in Section 4.4.1, where

and

$$\left.\begin{aligned} s_1(t) &= \sqrt{E}\,\psi_1(t) \\ s_2(t) &= -\sqrt{E}\,\psi_1(t) \end{aligned}\right\} 0 \le t \le T \tag{4.74}$$

The detector will choose the $s_i(t)$ with the largest correlator output $z_i(T)$; or, in this case of equal-energy antipodal signals, the detector, using the decision rule in Equation (4.20), decides on the basis of

and

$$\begin{aligned} &s_1(t) \quad \text{if } z(T) > \gamma_0 = 0 \\ &s_2(t) \quad \text{otherwise} \end{aligned} \tag{4.75}$$

Two types of errors can be made, as shown in Figure 4.9: The first type of error takes place if signal $s_1(t)$ is transmitted but the noise is such that the detector measures a negative value for $z(T)$ and chooses hypothesis H_2, the hypothesis that signal $s_2(t)$ was sent. The second type of error takes place if signal $s_2(t)$ is transmitted but the detector measures a positive value for $z(T)$ and chooses hypothesis H_1, the hypothesis that signal $s_1(t)$ was sent.

In Section 3.2.1.1, an expression was developed in Equation (3.42) for the probability of a bit error P_B, for this binary *minimum error* detector. We rewrite this relationship as

$$P_B = \int_{(a_1 - a_2)/2\sigma_0}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) du = Q\left(\frac{a_1 - a_2}{2\sigma_0}\right) \tag{4.76}$$

where σ_0 is the standard deviation of the noise out of the correlator. The function $Q(x)$, called the *complementary error function* or *co-error function.* is defined as

$$Q(X) = \frac{1}{\sqrt{2\pi}} \int_x^{\infty} \exp\left(-\frac{u^2}{2}\right) du \tag{4.77}$$

and is described in greater detail in Sections 3.2 and B.3.2.

For equal-energy antipodal signaling, such as the BPSK format in Equation (4.74), the receiver output signal components are $a_1 = \sqrt{E_b}$ when $s_1(t)$ is sent and $a_2 = -\sqrt{E_b}$ when $s_2(t)$ is sent, where E_b is the signal energy per binary symbol. For AWGN we can replace the noise variance σ_0^2 out of the correlator with $N_0/2$ (see Appendix C), so that we can rewrite Equation (4.76) as follows:

$$P_B = \int_{\sqrt{2E_b/N_0}}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) du \tag{4.78}$$

$$= Q\left(\sqrt{\frac{2E_b}{N_0}}\right) \tag{4.79}$$

This result for bandpass antipodal BPSK signaling is the same as the results that were developed earlier in Equation (3.70) for the matched-filter detection of antipodal signaling in general, and in Equation (3.76) for the matched-filter detection of baseband antipodal signaling in particular. This is an example of an *equivalence theorem,* described earlier. For linear systems the equivalence theorem establishes that the mathematics of detection is unaffected by a shift in frequency. Hence in this chapter, the use of matched filters or correlators in the detection of bandpass signals yields the same relationships as those developed for comparable signals at baseband.

Example 4.4 Bit Error Probability for BPSK Signaling

Find the bit error probability for a BPSK system with a bit rate of 1 Mbit/s. The received waveforms $s_1(t) = A \cos \omega_0 t$ and $s_2(t) = -A \cos \omega_0 t$, are coherently detected with a matched filter. The value of A is 10 mV. Assume that the single-sided noise power spectral density is $N_0 = 10^{-11}$ W/Hz and that signal power and energy per bit are normalized relative to a 1-Ω load.

Solution

$$A = \sqrt{\frac{2E_b}{T}} = 10^{-2} \text{ V} \qquad T = \frac{1}{R} = 10^{-6} \text{ s}$$

Thus,

$$E_b = \frac{A^2}{2} T = 5 \times 10^{-11} \text{ J} \qquad \text{and} \qquad \sqrt{\frac{2E_b}{N_0}} = 3.16$$

Also,

$$P_B = Q\left(\sqrt{\frac{2E_b}{N_0}}\right) = Q(3.16)$$

Using Table B.1 or Equation (3.44), we obtain

$$P_B = 8 \times 10^{-4}$$

4.7.2 Probability of Bit Error for Coherently Detected, Differentially Encoded Binary PSK

Channel waveforms sometimes experience inversion; for example, when using a coherent reference generated by a phase-locked loop, one may have phase ambiguity. If the carrier phase were reversed in a DPSK modulation application, what would be the effect on the message? The only effect would be an error in the bit during

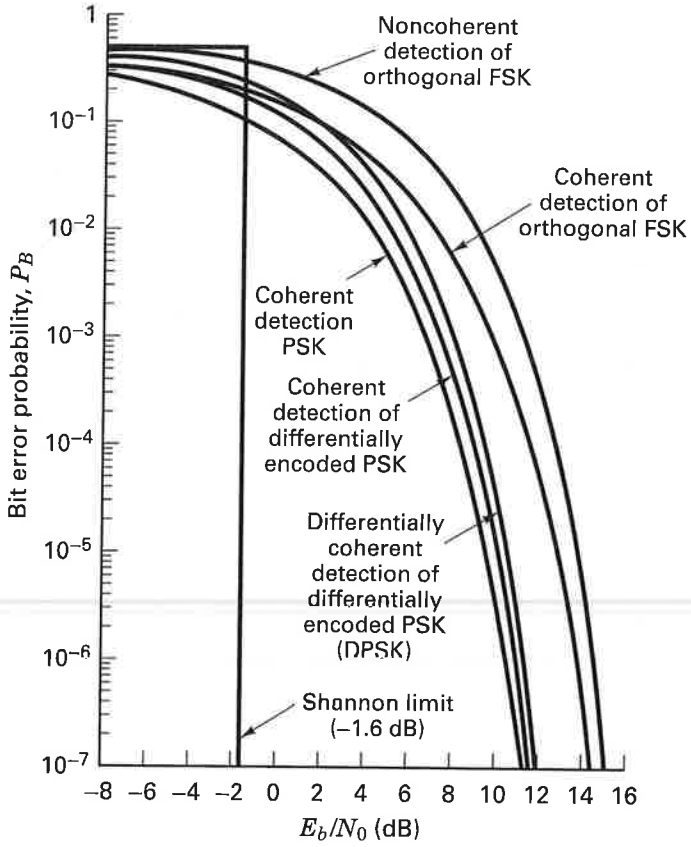


Figure 4.25 Bit error probability for several types of binary systems.

which inversion occurred or the bit just after inversion, since the message information is encoded in the similarity or difference between adjacent symbols. The similarity or difference quality remains unchanged if the carrier is inverted. Sometimes messages (and their assigned waveforms) are *differentially encoded* and *coherently detected* simply to avoid these phase ambiguities.

The probability of bit error for coherently detected, differentially encoded PSK is given by [5]

$$P_B = 2Q\left(\sqrt{\frac{2E_b}{N_0}}\right)\left[1 - Q\left(\sqrt{\frac{2E_b}{N_0}}\right)\right] \tag{4.80}$$

This relationship is plotted in Figure 4.25. Notice that there is a small degradation of error performance compared with the coherent detection of PSK. This is due to the differential encoding, since any single detection error will usually result in two decision errors. Error performance for the more popular differentially coherent detection (DPSK) is covered in Section 4.7.5.

4.7.3 Probability of Bit Error for Coherently Detected Binary Orthogonal FSK

Equations (4.78) and (4.79) describe the probability of bit error for coherent antipodal signals. A more general treatment for binary coherent signals (not limited to antipodal signals) yields the following equation for P_B [6]:

$$P_B = \frac{1}{\sqrt{2\pi}} \int_{\sqrt{(1-\rho)E_b/N_0}}^{\infty} \exp\left(-\frac{u^2}{2}\right) du \tag{4.81}$$

From Equation (3.64b), $\rho = \cos\theta$ is the time cross-correlation coefficient between signal $s_1(t)$ and $s_2(t)$, where θ is the angle between signal vectors $\mathbf{s}_1$ and $\mathbf{s}_2$ (see Figure 4.6). For antipodal signals such as BPSK, $\theta = \pi$, thus $\rho = -1$.

For orthogonal signals such as binary FSK (BFSK), $\theta = \pi/2$, since the $\mathbf{s}_1$ and $\mathbf{s}_2$ vectors are perpendicular to each other; thus $\rho = 0$, as can be verified with Equation (3.64a), and Equation (4.81) can then be written as

$$P_B = \frac{1}{\sqrt{2\pi}} \int_{\sqrt{E_b/N_0}}^{\infty} \exp\left(-\frac{u^2}{2}\right) du = Q\left(\sqrt{\frac{E_b}{N_0}}\right) \tag{4.82}$$

where the *co-error function* $Q(x)$ is described in greater detail in Sections 3.2 and B.3.2. The result in Equation (4.82) for the coherent detection of orthogonal BFSK plotted in Figure 4.25 is the same as the results that were developed earlier in Equation (3.71) for the matched-filter detection of orthogonal signaling, in general, and in Equation (3.73) for the matched-filter detection of baseband orthogonal signaling (unipolar pulses), in particular. The details of *on-off keying* (OOK) are not treated in this book. However, it is worth noting that on-off keying is an orthogonal signaling set (unipolar pulse signaling is the baseband equivalent of OOK). Thus the relationship in Equation (4.82) applies to the matched-filter detection of OOK, as it does to any coherent detection of orthogonal signaling.

The relationship in Equation (4.82) can also be confirmed by noting that the energy difference between the orthogonal signal vectors $\mathbf{s}_1$ and $\mathbf{s}_2$, with amplitudes of $\sqrt{E_b}$, as shown in Figure 3.10b, can be computed as the square of the distance between the heads of the orthogonal vectors, which will be $E_d = 2E_b$. Using this result in Equation (3.63) also yields Equation (4.82). If we compare Equation (4.82) with Equation (4.79), we can see that 3-dB more E_b/N_0 is required for BFSK to provide the same performance as BPSK. It should not be surprising that the performance of BFSK signaling is 3-dB worse than BPSK signaling, since for a given signal power, the distance-squared between orthogonal vectors is a factor of two less than the distance squared between antipodal vectors.

4.7.4 Probability of Bit Error for Noncoherently Detected Binary Orthogonal FSK

Consider the equally likely binary orthogonal FSK signal set $\{s_1(t)\}$, defined in Equation (4.8) as follows:

$$s_i(t) = \sqrt{\frac{2E}{T}} \cos(\omega_i t + \phi) \qquad 0 \le t \le T, \qquad i = 1, 2$$

The phase term ϕ is unknown and assumed constant. The detector is characterized by $M = 2$ channels of bandpass filters and envelope detectors, as shown in Figure 4.19. The input to the detector consists of the received signal $r(t) = s_i(t) + n(t)$, where $n(t)$ is a white Gaussian noise process with two-sided power spectral density $N_0/2$. Assume that $s_1(t)$ and $s_2(t)$ are separated in frequency sufficiently that they have negligible overlap. For $s_1(t)$ and $s_2(t)$ being equally likely, we start the bit-error probability P_B computation with Equation (3.38) as we did for baseband signaling:

$$\begin{aligned} P_B &= \tfrac{1}{2}P(H_2|s_1) + \tfrac{1}{2}P(H_1|s_2) \\ &= \frac{1}{2}\int_{-\infty}^{0} p(z|s_1)\,dz + \frac{1}{2}\int_{0}^{\infty} p(z|s_2)\,dz \end{aligned} \tag{4.83}$$

For the binary case, the *test statistic* $z(T)$ is defined by $z_1(T) - z_2(T)$. Assume that the bandwidth of the filter W_f is $1/T$, so that the envelope of the FSK signal is (approximately) preserved at the filter output. If there was no noise at the receiver, the value of $z(T) = \sqrt{2E/T}$ when $s_1(t)$ is sent, and $z(T) = -\sqrt{2E/T}$ when $s_2(t)$ is sent. Because of this symmetry, the optimum threshold is $\gamma_0 = 0$. The pdf $p(z|s_1)$ is similar to $p(z|s_2)$; that is,

$$p(z|s_1) = p(-z|s_2) \tag{4.84}$$

Therefore, we can write

$$P_B = \int_{0}^{\infty} p(z|s_2)\,dz \tag{4.85}$$

or

$$P_B = P(z_1 > z_2|s_2) \tag{4.86}$$

where z_1 and z_2 denote the outputs $z_1(T)$ and $z_2(T)$ from the envelope detectors shown in Figure 4.19. For the case in which the tone $s_2(t) = \cos \omega_2 t$ is sent, such that $r(t) = s_2(t) + n(t)$, the output $z_1(T)$ is a *Gaussian noise random variable only;* it has no signal component. A Gaussian distribution into the *nonlinear envelope detector* yields a Rayleigh distribution at the output [6], so that

$$p(z_1|s_2) = \begin{cases} \dfrac{z_1}{\sigma_0^2} \exp\left(-\dfrac{z_1^2}{2\sigma_0^2}\right) & z_1 \ge 0 \\ 0 & z_1 < 0 \end{cases} \tag{4.87}$$

where σ_0^2 is the noise at the filter output. On the other hand, $z_2(T)$ has a Rician distribution, since the input to the lower envelope detector is a sinusoid plus noise [6]. The pdf $p(z_2|s_2)$ is written as

$$p(z_2|s_2) = \begin{cases} \dfrac{z_2}{\sigma_0^2} \exp\left[-\dfrac{(z_2^2 + A^2)}{2\sigma_0^2}\right] I_0\left(\dfrac{z_2 A}{\sigma_0^2}\right) & z_2 \geq 0 \\ 0 & z_2 < 0 \end{cases} \tag{4.88}$$

where $A = \sqrt{2E/T}$, and as before, σ_0^2 is the noise at the filter output. The function $I_0(x)$, known as the modified zero-order Bessel function of the first kind [7], is defined as

$$I_0(x) = \frac{1}{2\pi} \int_0^{2\pi} \exp(x \cos\theta)\, d\theta \tag{4.89}$$

When $s_2(t)$ is transmitted, the receiver makes an error whenever the envelope sample $z_1(T)$ obtained from the upper channel (due to noise alone) exceeds the envelope sample $z_2(T)$ obtained from the lower channel (due to signal plus noise). Thus the probability of this error can be obtained by integrating $p(z_1|s_2)$ with respect to z_1 from z_2 to infinity, and then averaging over all possible values of z_2. That is,

$$\begin{aligned} P_B &= P(z_1 > z_2|s_2) \\ &= \int_0^\infty p(z_2|s_2) \left[\int_{z_2}^\infty p(z_1|s_2)\, dz_1\right] dz_2 \qquad (4.90) \\ &= \int_0^\infty \frac{z_2}{\sigma_0^2} \exp\left[-\frac{(z_2^2 + A^2)}{2\sigma_0^2}\right] I_0\left(\frac{z_2 A}{\sigma_0^2}\right)\left[\int_{z_2}^\infty \frac{z_1}{\sigma_0^2} \exp\left(-\frac{z_1^2}{2\sigma_0^2}\right) dz_1\right] dz_2 \qquad (4.91) \end{aligned}$$

where $A = \sqrt{2E/T}$ and where the inner integral is the conditioned probability of an error for a fixed value of z_2, given that $s_2(t)$ was sent, and the outer integral averages this conditional probability over all possible values of z_2. This integral can be evaluated [8], yielding

$$P_B = \frac{1}{2} \exp\left(-\frac{A^2}{4\sigma_0^2}\right) \tag{4.92}$$

Using Equation (1.19), we can express the filter output noise as

$$\sigma_0^2 = 2\left(\frac{N_0}{2}\right) W_f \tag{4.93}$$

where $G_n(f) = N_0/2$ and W_f is the filter bandwidth. Thus Equation (4.92) becomes

$$P_B = \frac{1}{2} \exp\left(-\frac{A^2}{4N_0 W_f}\right) \tag{4.94}$$

Equation (4.94) indicates that the error performance depends on the bandpass filter bandwidth, and that P_B becomes smaller as W_f is decreased. The result is valid only when the intersymbol interference (ISI) is negligible. The minimum W_f allowed (i.e., for no ISI) is obtained from Equation (3.81) with the filter roll-off factor $r = 0$. Thus $W_f = R$ bits/s $= 1/T$, and we can write Equation (4.94) as

$$P_B = \frac{1}{2}\exp\left(-\frac{A^2T}{4N_0}\right) \tag{4.95}$$

$$= \frac{1}{2}\exp\left(-\frac{E_b}{2N_0}\right) \tag{4.96}$$

where $E_b = (1/2)A^2T$ is the energy per bit. When comparing the error performance of noncoherent FSK with coherent FSK (see Figure 4.25), it is seen that for the same P_B, noncoherent FSK requires approximately 1 dB more E_b/N_0 than that for coherent FSK (for $P_B \leq 10^{-4}$). The noncoherent receiver is easier to implement, because coherent reference signals need not be generated. Therefore, almost all FSK receivers use noncoherent detection. It can be seen in the following section that when comparing noncoherent orthogonal FSK to noncoherent DPSK, the same 3-dB difference occurs as for the comparison between coherent orthogonal FSK and coherent PSK.

As mentioned earlier, the details of on-off keying (OOK) are not treated in this book. However, it is worth noting that the bit error probability P_B described in Equation (4.96) is identical to the P_B for the noncoherent detection of OOK signaling.

4.7.5 Probability of Bit Error for Binary DPSK

Let us define a BPSK signal set as follows:

$$\begin{aligned} x_1(t) &= \sqrt{\frac{2E}{T}}\cos(\omega_0 t + \phi) \qquad 0 \leq t \leq T \\ x_2(t) &= \sqrt{\frac{2E}{T}}\cos(\omega_0 t + \phi \pm \pi) \qquad 0 \leq t \leq T \end{aligned} \tag{4.97}$$

A characteristic of DPSK is that there are no fixed decision regions in the signal space. Instead, the decision is based on the phase difference between successively received signals. Then for DPSK signaling we are really transmitting each bit with the binary signal pair

and

$$\begin{aligned} s_1(t) &= (x_1, x_1) \quad \text{or} \quad (x_2, x_2) \qquad 0 \leq t \leq 2T \\ s_2(t) &= (x_1, x_2) \quad \text{or} \quad (x_2, x_1) \qquad 0 \leq t \leq 2T \end{aligned} \tag{4.98}$$

where (x_i, x_j) $(i, j = 1, 2)$ denotes $x_i(t)$ followed by $x_j(t)$ defined in Equation (4.97). The first T seconds of each waveform are actually the last T seconds of the previous waveform. Note that $s_1(t)$ and $s_2(t)$ can each have either of two possible forms and that $x_1(t)$ and $x_2(t)$ are antipodal signals. Thus the correlation between $s_1(t)$ and $s_2(t)$ for *any combination* of forms can be written as

$$z(2T) = \int_0^{2T} s_1(t)s_2(t)\,dt = \int_0^T [x_1(t)]^2\,dt - \int_0^T [x_1(t)]^2\,dt = 0 \tag{4.99}$$

Therefore, pairs of DPSK signals can be represented as orthogonal signals $2T$ seconds long. Detection could correspond to noncoherent envelope detection with four channels matched to each of the possible envelope outputs, as shown in Figure 4.26a. Since the two envelope detectors representing each symbol are negatives of each other, the envelope sample of each will be the same. Hence we can implement the detector as a single channel for $s_1(t)$ matched to either (x_1, x_1) or (x_2, x_2), and a single channel for $s_2(t)$ matched to either (x_1, x_2) or (x_2, x_1), as shown in Figure 4.26b. The DPSK detector is therefore reduced to a standard two-channel noncoherent detector. In reality, the filter can be matched to the difference signal so that only one channel is necessary. In Figure 4.26, the filters are matched to the signal envelopes (over two symbol times). What does this mean in light of the fact that

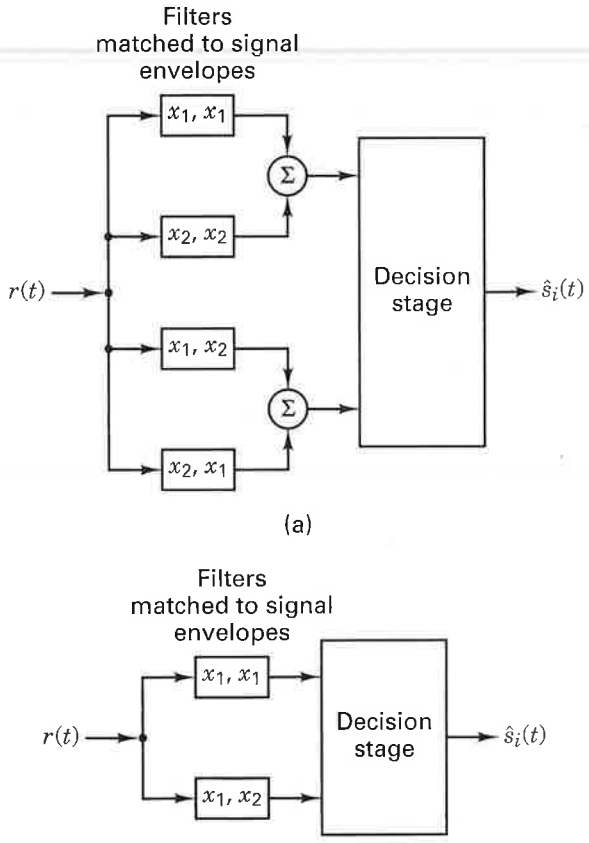


Figure 4.26 DPSK detection. (a) Four-channel differentially coherent detection of binary DPSK. (b) Equivalent two-channel detector for binary DPSK.

DPSK is a constant envelope signaling scheme? It means that we need to implement an energy detector, similar to the quadrature receiver in Figure 4.18, with each of the I and Q reference signals occuring in pairs during the time interval $(0 \le t \le 2T)$, as follows:

$$\begin{aligned} s_1(t)\ I \text{ reference:} &\quad \sqrt{2/T}\cos\omega_0 t,\ \sqrt{2/T}\cos\omega_0 t \\ s_1(t)\ Q \text{ reference:} &\quad \sqrt{2/T}\sin\omega_0 t,\ \sqrt{2/T}\sin\omega_0 t \end{aligned}$$

$$\begin{aligned} s_2(t)\ I \text{ reference:} &\quad \sqrt{2/T}\cos\omega_0 t,\ -\sqrt{2/T}\cos\omega_0 t \\ s_2(t)\ Q \text{ reference:} &\quad \sqrt{2/T}\sin\omega_0 t,\ -\sqrt{2/T}\sin\omega_0 t \end{aligned}$$

Since pairs of DPSK signals are orthogonal, such noncoherent detection operates with the bit-error probability given by Equation (4.96). However, since the DPSK signals have a bit interval of $2T$, the $s_i(t)$ signals defined in Equation (4.98) have twice the energy of a signal defined over a single-symbol duration. Thus, we may write P_B as

$$P_B = \frac{1}{2}\exp\left(-\frac{E_b}{N_0}\right) \tag{4.100}$$

Equation (4.100) is plotted in Figure 4.25, designated as differentially coherent detection of differentially encoded PSK, or simply DPSK. This expression is valid for the optimum DPSK detector shown in Figure 4.17c. For the detector shown in Figure (4.17b), the error probability will be slightly inferior to that given in Equation (4.100) [3]. When comparing the error performance of Equation (4.100) with that of coherent PSK (see Figure 4.25), it is seen that for the same P_B, DPSK requires approximately 1 dB more E_b/N_0 than does BPSK (for $P_B \le 10^{-4}$). It is easier to implement a DPSK system than a PSK system, since the DPSK receiver does not need phase synchronization. For this reason, DPSK, although less efficient than PSK, is sometimes the preferred choice between the two.

4.7.6 Comparison of Bit Error Performance for Various Modulation Types

The P_B expressions for the best known of the binary modulation schemes discussed above are listed in Table 4.1 and are illustrated in Figure 4.25. For $P_B = 10^{-4}$, it can be seen that there is a difference of approximately 4-dB between the best (coherent PSK) and the worst (noncoherent orthogonal FSK) that were discussed here. In some cases, 4 dB is a small price to pay for the implementation simplicity gained in going from coherent PSK to noncoherent FSK; however, for other cases, even a 1-dB saving is worthwhile. There are other considerations besides P_B and system complexity; for example, in some cases (such as a randomly fading channel), a noncoherent system is more desirable because there may be difficulty in establishing and maintaining a coherent reference. Signals that can withstand significant degradation before their ability to be detected is affected are clearly desirable in military and space applications.

TABLE 4.1 Probability of Error for Selected Binary Modulation Schemes

Modulation	P_B
PSK (coherent)	$Q\left(\sqrt{\frac{2E_b}{N_0}}\right)$
DPSK (differentially coherent)	$\frac{1}{2}\exp\left(-\frac{E_b}{N_0}\right)$
Orthogonal FSK (coherent)	$Q\left(\sqrt{\frac{E_b}{N_0}}\right)$
Orthogonal FSK (noncoherent)	$\frac{1}{2}\exp\left(-\frac{1}{2}\frac{E_b}{N_0}\right)$

4.8 *M*-ARY SIGNALING AND PERFORMANCE

4.8.1 Ideal Probability of Bit Error Performance

The typical probability of error versus E_b/N_0 curve was shown to have a waterfall-like shape in Figure 3.6. The probability of bit error (P_B) characteristics of various binary modulation schemes in AWGN also display this shape, as shown in Figure 4.25. What should an *ideal* P_B versus E_b/N_0 curve look like? Figure 4.27 displays the ideal characteristic as the *Shannon limit*. The limit represents the threshold E_b/N_0 below which reliable communication cannot be maintained. Shannon's work is described in greater detail in Chapter 9.

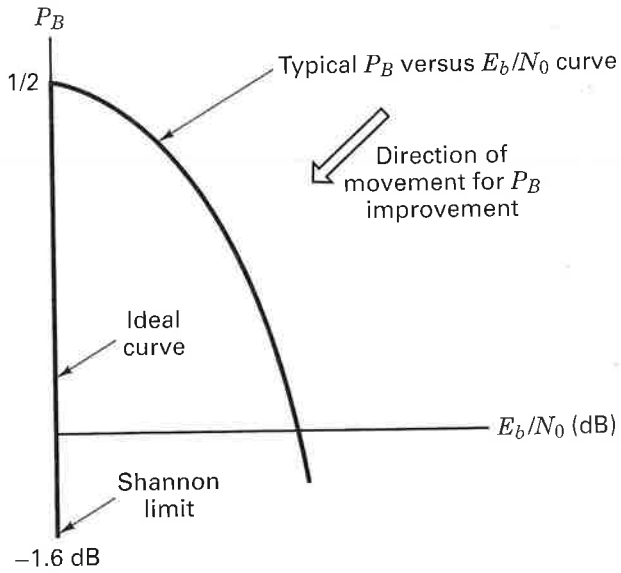


Figure 4.27 P_B versus E_b/N_0 curves: ideal and typical.

We can describe the ideal curve in Figure 4.27 as follows. For all values of E_b/N_0 above the Shannon limit of −1.6 dB, P_B is zero. Once E_b/N_0 is reduced below the Shannon limit, P_B degrades to the worst-case value of $\frac{1}{2}$. (Note that $P_B = 1$ is not the worst case for binary signaling, since that value is just as good as $P_B = 0$; if the probability of making a bit error is 100%, the bit stream could simply be inverted to retrieve the correct data.) It should be clear, by comparing the typical P_B curve with the ideal one in Figure 4.27 that the large arrow in the figure describes the desired direction of movement to achieve improved P_B performance.

4.8.2 *M*-ary Signaling

Let us review M-ary signaling. The processor considers k bits at a time. It instructs the modulator to produce one of $M = 2^k$ waveforms; binary signaling is the special case where $k = 1$. Does M-ary signaling improve or degrade error performance? (Be careful with your answer—the question is a loaded one.) Figure 4.28 illustrates the probability of bit error $P_B(M)$ versus E_b/N_0 for coherently detected *orthogonal* M-ary signaling over a Gaussian channel. Figure 4.29 similarly illustrates $P_B(M)$ versus E_b/N_0 for coherently detected *multiple phase* M-ary signaling over a Gaussian channel. In which direction do the curves move as the value of k (or M) increases? From Figure 4.27 we know the directions of curve movement for improved and degraded error performance. In Figure 4.28, as k increases, the curves move in the direction of improved error performance. In Figure 4.29, as k increases, the curve move in the direction of degraded error performance. Such movement tells us that M-ary signaling produces improved error performance with orthogonal signaling and degraded error performance with multiple phase signaling. Can that be true? Why would anyone ever use multiple phase PSK signaling if it provides degraded error performance compared to binary PSK signaling? It *is* true, and many systems do use multiple phase signaling. The question, as stated, is loaded because it implies that error probability versus E_b/N_0 is the *only* performance criterion; there are many others (e.g., bandwidth, throughput, complexity, cost), but in Figures 4.28 and 4.29 error performance is the characteristic that stands out explicitly.

A performance characteristic that is not explicitly seen in Figures 4.28 and 4.29 is the required system bandwidth. For the curves characterizing M-ary orthogonal signals in Figure 4.28, as k increases, the required bandwidth also increases. For the M-ary multiple phase curves in Figure 4.29, as k increases, a larger bit rate can be transmitted within the same bandwidth. In other words, for a fixed data rate, the required bandwidth is decreased. Therefore, *both* the orthogonal and multiple phase error performance curves tell us that M-ary signaling represents a vehicle for performing a system trade-off. In the case of orthogonal signaling, error performance improvement can be achieved at the expense of bandwidth. In the case of multiple phase signaling, bandwidth performance can be achieved at the expense of error performance. Error performance versus bandwidth performance, a fundamental communications trade-off, is treated in greater detail in Chapter 9.

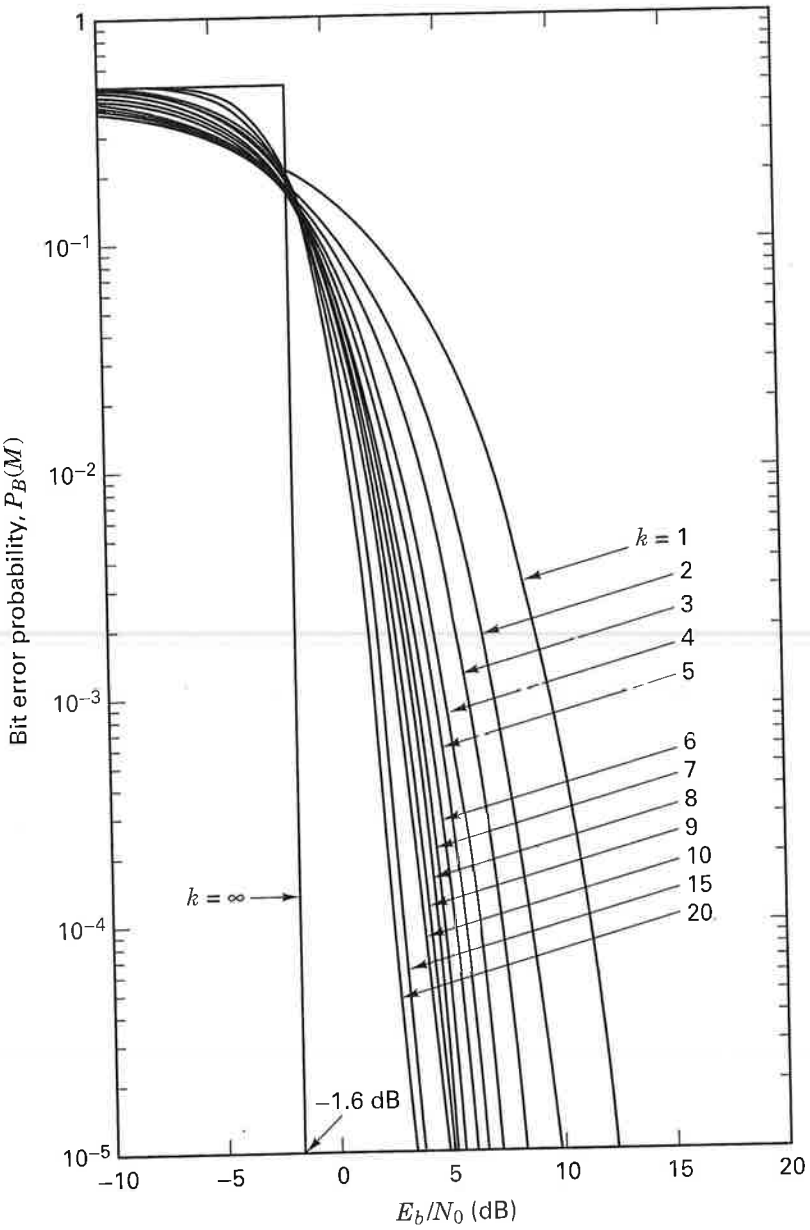


Figure 4.28 Bit error probability for coherently detected *M*-ary orthogonal signaling. (Reprinted from W. C. Lindsey and M. K. Simon, *Telecommunication Systems Engineering,* Prentice Hall, Inc., Englewood Cliffs, N.J., 1973, courtesy of W. C. Lindsey and Marvin K. Simon).

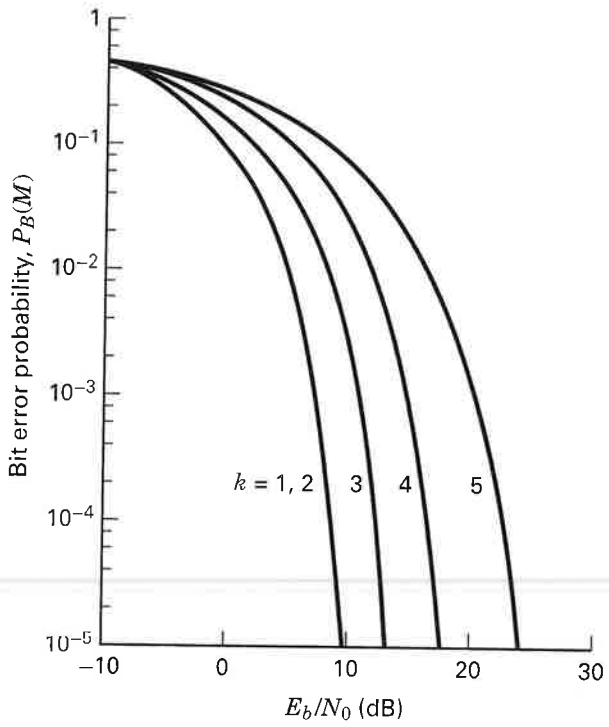


Figure 4.29 Bit error probability for coherently detected multiple phase signaling.

4.8.3 Vectorial View of MPSK Signaling

Figure 4.30 illustrates MPSK signal sets for $M = 2, 4, 8$, and 16. In Figure 4.30a we see the binary ($k = 1$, $M = 2$) antipodal vectors $\mathbf{s}_1$ and $\mathbf{s}_2$ positioned 180° apart. The decision boundary is drawn so as to partition the signal space into two regions. On the figure is also shown a noise vector $\mathbf{n}$ equal in magnitude to $\mathbf{s}_1$. The figure establishes the magnitude and orientation of the minimum energy noise vector that would cause the detector to make a symbol error.

In Figure 4.30b we see the 4-ary ($k = 2$, $M = 4$) vectors positioned 90° apart. The decision boundaries (only one line is drawn) divide the signal space into four regions. Again a noise vector $\mathbf{n}$ is drawn (from the head of a signal vector, normal to the closest decision boundary) to illustrate the minimum energy noise vector that would cause the detector to make a symbol error. Notice that the minimum energy noise vector of Figure 4.30b is smaller than that of Figure 4.30a, illustrating that the 4-ary system is more vulnerable to noise than the 2-ary system (signal energy being equal for each case). As we move on to Figure 4.30c for the 8-ary case and Figure 4.30d for the 16-ary case, it should be clear that for multiple phase signaling, as M increases, we are crowding more signal vectors into the signal plane. As the vectors are moved closer together, a smaller amount of noise energy is required to cause an error.

Figure 4.30 adds some insight as to why the curves of Figure 4.29 behave as they do as k is increased. Figure 4.30 also provides some insight into a basic trade-

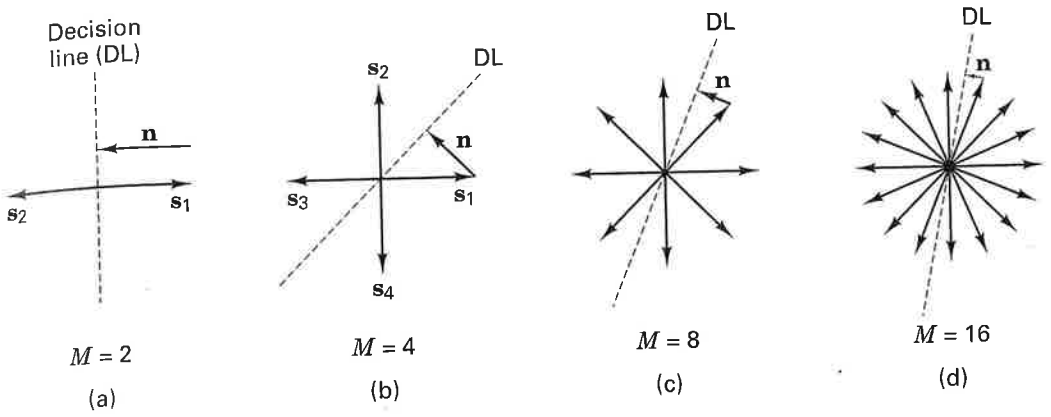


Figure 4.30 MPSK signal sets for $M = 2, 4, 8, 16$.

off in multiple phase signaling. Crowding more signal vectors into the signal space is tantamount to increasing the data rate without increasing the system bandwidth (the vectors are all confined to the same plane). In other words, we have increased the bandwidth utilization at the expense of error performance. Look at Figure 4.30d, where the error performance is worse than any of the other examples in Figure 4.30. How might we "buy back" the degraded error performance? In other words, what can we trade-off so that the distance between neighboring signal vectors in Figure 4.30d is increased to that in Figure 4.30a? We can increase the signal strength (make the signal vectors larger) until the minimum distance from the head of a signal vector to a decision line equals the length of the noise vector in Figure 4.30a. Therefore, in a multiple phase system, as M is increased, we can either achieve improved bandwidth performance at the expense of error performance, or if we increase the E_b/N_0 so that the error probability is not degraded, we can achieve improved bandwidth performance at the expense of increasing E_b/N_0.

Note that Figure 4.30 has been sketched so that all phasors have the same length for any of the M-ary cases. This is tantamount to saying that the comparisons are being considered for a fixed E_s/N_0, where E_s is symbol energy. The figure can also be drawn for a fixed E_b/N_0, in which case the phasor magnitudes would increase with increasing M. The phasors for $M = 4$, 8, and 16 would then have lengths greater than the $M = 2$ case by the factors $\sqrt{2}$, $\sqrt{3}$, and 2 respectively. We would still see crowding and increased vulnerability to noise, with increasing M, but the appearance would not be as pronounced as it is in Figure 4.30.

4.8.4 BPSK and QPSK Have the Same Bit Error Probability

In Equation (3.30) we stated the general relationship between E_b/N_0 and S/N which is rewritten

$$\frac{E_b}{N_0} = \frac{S}{N}\left(\frac{W}{R}\right) \tag{4.101}$$

where S is the average signal power and R is the bit rate. A BPSK signal with the available E_b/N_0 found from Equation (4.101) will perform with a P_B that can be read from the $k = 1$ curve in Figure 4.29. QPSK can be characterized as two orthogonal BPSK channels. The QPSK bit stream is usually partitioned into an even and odd (I and Q) stream; each new stream modulates an orthogonal component of the carrier at half the bit rate of the original stream. The I stream modulates the cos $\omega_0 t$ term and the Q stream modulates the sin $\omega_0 t$ term. If the magnitude of the original QPSK vector has the value A, the magnitude of the I and Q component vectors each has a value of $A/\sqrt{2}$, as shown in Figure 4.31. Thus, each of the quadrature BPSK signals has half of the average power of the original QPSK signal. Hence if the original QPSK waveform has a bit rate of R bits/s and an average power of S watts, the quadrature partitioning results in each of the BPSK waveforms having a bit rate of $R/2$ bits/s and an average power of $S/2$ watts.

Therefore, the E_b/N_0 characterizing each of the orthogonal BPSK channels, making up the QPSK signal, is equivalent to the E_b/N_0 in Equation (4.101), since it can be written as

$$\frac{E_b}{N_0} = \frac{S/2}{N_0}\left(\frac{W}{R/2}\right) = \frac{S}{N_0}\left(\frac{W}{R}\right) \tag{4.102}$$

Thus each of the orthogonal BPSK channels, and hence the composite QPSK signal, is characterized by the same E_b/N_0 and hence the same P_B performance as a BPSK signal. The natural orthogonality of the 90° phase shifts between adjacent QPSK symbols results in the *bit error probabilities* being equal for both BPSK and QPSK signaling. It is important to note that the symbol *error probabilities* are *not* equal for BPSK and QPSK signaling. The relationship between bit error probability and symbol error probability is treated in Sections 4.9.3 and 4.9.4. We see that, in effect, QPSK is the equivalent of two BPSK channels in quadrature. This same idea can be extended to any symmetrical M-ary amplitude/phase signaling, such as quadrature amplitude modulation (QAM) described in Section 9.8.3.

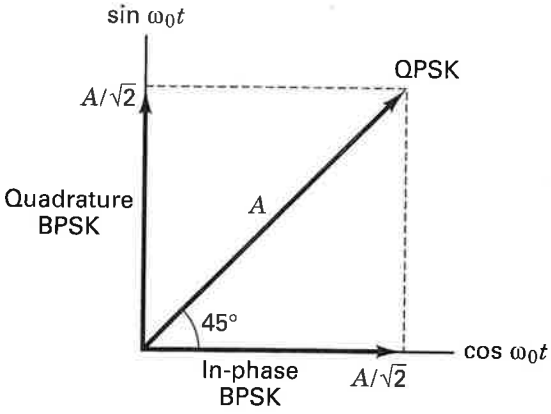


Figure 4.31 In-phase and quadrature BPSK components of QPSK signaling.

4.8.5 Vectorial View of MFSK Signaling

In Section 4.8.3, Figure 4.30 provides some insight as to why the error performance of MPSK signaling degrades as k (or M) increases. It would be useful to have a similar vectorial illustration for the error performance of orthogonal MFSK signaling as seen in the curves of Figure (4.28). Since the MFSK signal space is characterized by M mutually perpendicular axes, we can only conveniently illustrate the cases $M = 2$ and $M = 3$. In Figure 4.32a we see the binary orthogonal vectors $\mathbf{s}_1$ and $\mathbf{s}_2$ positioned 90° apart. The decision boundary is drawn so as to partition the signal space into two regions. On the figure is also shown a noise vector $\mathbf{n}$, which represents the minimum noise vector that would cause the detector to make an error.

In Figure 4.32b we see a 3-ary signal space with axes positioned 90° apart. Here decision planes partition the signal space into three regions. Noise vectors $\mathbf{n}$ are shown added to each of the prototype signal vectors $\mathbf{s}_1$, $\mathbf{s}_2$, and $\mathbf{s}_3$; each noise vector illustrates an example of the minimum noise vector that would cause the detector to make a symbol error. The minimum noise vectors in Figure 4.32b are the same length as the noise vector in Figure 4.32a. In Section 4.4.4 we stated that for a given level of received energy, the distance between any two prototype signal vectors $\mathbf{s}_i$ and $\mathbf{s}_j$ in an M-ary orthogonal space is constant. It follows that the minimum distance between a prototype signal vector and any of the decision boundaries remains fixed as M increases. Unlike the case of MPSK signaling, where adding new signals to the signal set makes the signals vulnerable to smaller noise vectors, here, in the case of MFSK signaling, adding new signals to the signal set does *not* make the signals vulnerable to smaller noise vectors.

It would be convenient to illustrate the point by drawing higher dimensional orthogonal spaces, but of course this is not possible. We can only use our "mind's eye" to understand that increasing the signal set M—by adding additional axes, where each new axis is mutually perpendicular to all the others—does not crowd

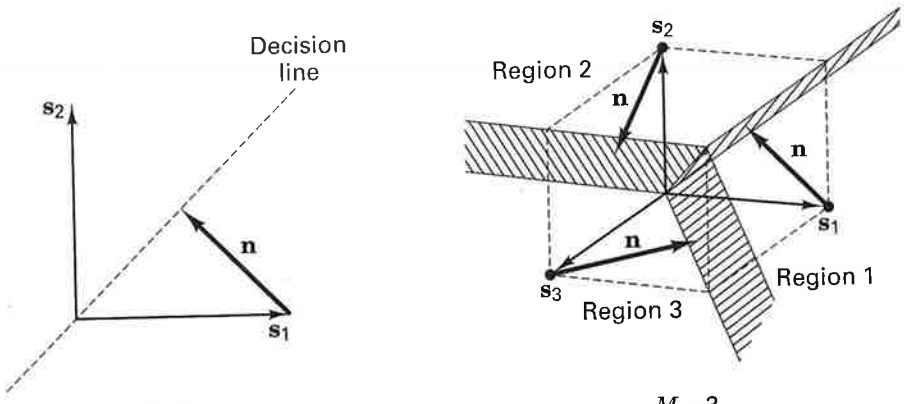


Figure 4.32 MFSK signal sets for $M = 2, 3$.

the signal set more closely together. Thus, a transmitted signal from an orthogonal set is *not* more vulnerable to a noise vector when the set is increased in size. In fact, we see from Figure 4.28 that as k increases, the bit error performance improves.

Understanding the error performance improvement of orthogonal signaling, as illustrated in Figure 4.28, is facilitated by comparing the probability of symbol error (P_E) versus unnormalized SNR, with P_E versus E_b/N_0. Figure 4.33 represents a set of P_E performance curves plotted against unnormalized SNR for coherent FSK signaling. Here we see that P_E degrades as M is increased. Didn't we say that a signal from an orthogonal set is *not* made more vulnerable to a given noise vector, as the orthogonal set is increased in size? It is correct that for orthogonal signaling, with a given SNR it takes a fixed size noise vector to perturb a transmitted signal into an error region; the signals do not become vulnerable to smaller noise vectors as M increases. However, as M increases, more neighboring decision regions are introduced; thus the number of ways in which a symbol error can be made increases. Figure 4.33 reflects the degradation in P_E versus unnormalized SNR as M is increased; there are $(M-1)$ ways to make an error. Examining performance under the condition of a fixed SNR (as M increases) is not very useful for digital communications. A fixed SNR means a fixed amount of energy per symbol; thus as M increases, there is a fixed amount of energy to be apportioned over a larger number of bits, or there is less energy per bit. The most useful way of comparing one digital system with another is on the basis of *bit-normalized SNR* or E_b/N_0. The error performance improvement with increasing M (see Figure 4.28) manifests itself only when error probability is plotted against E_b/N_0. For this case, as M increases, the required E_b/N_0 (to meet a given error probability) is reduced for a fixed SNR; therefore, we need to map the plot shown in Figure 4.33 into a new plot, similar to that shown in Figure 4.28, where the abscissa represents E_b/N_0 instead of SNR. Figure 4.34 illustrates such a mapping; it demonstrates that curves manifesting degraded P_E with increasing M (such as Figure 4.33) are transformed into curves manifesting improved P_E with increasing M. The basic mapping relationship is expressed in Equation (4.101), repeated here as

$$\frac{E_b}{N_0} = \frac{S}{N}\left(\frac{W}{R}\right)$$

where W is the detection bandwidth. Since

$$R = \frac{\log_2 M}{T} = \frac{k}{T}$$

where T is the symbol duration, we can then write

$$\frac{E_b}{N_0} = \frac{S}{N}\left(\frac{WT}{\log_2 M}\right) = \frac{S}{N}\left(\frac{WT}{k}\right) \tag{4.103}$$

For FSK signaling, the detection bandwidth W (in hertz) is typically equal in value to the symbol rate $1/T$; in other words, $WT \approx 1$. Therefore,

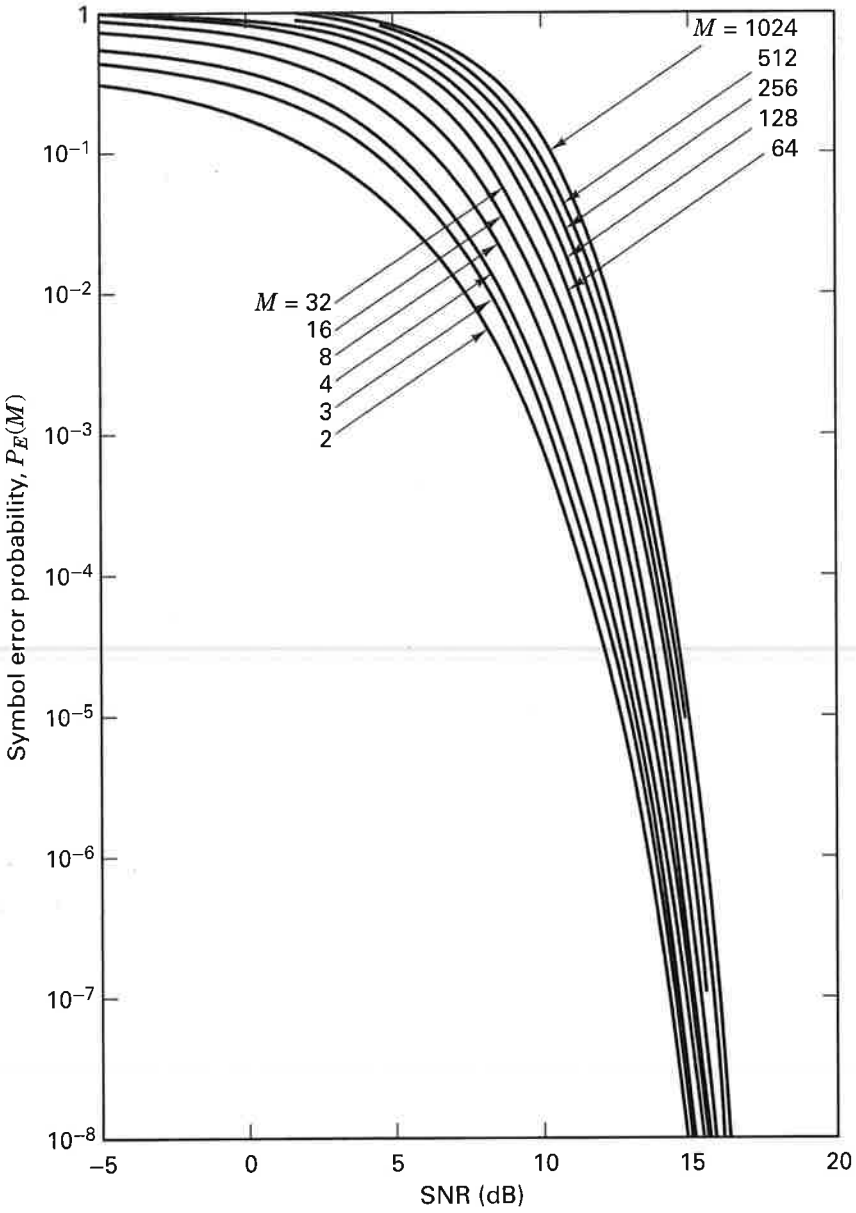


Figure 4.33 Symbol error probability versus SNR for coherent FSK signaling. (From Bureau of Standards, *Technical Note 167*, March 1963.) (Reprinted from *Central Radio Propagation Laboratory Technical Note 167*, March 25, 1963, Fig. 1, p. 5, courtesy of National Bureau of Standards.)

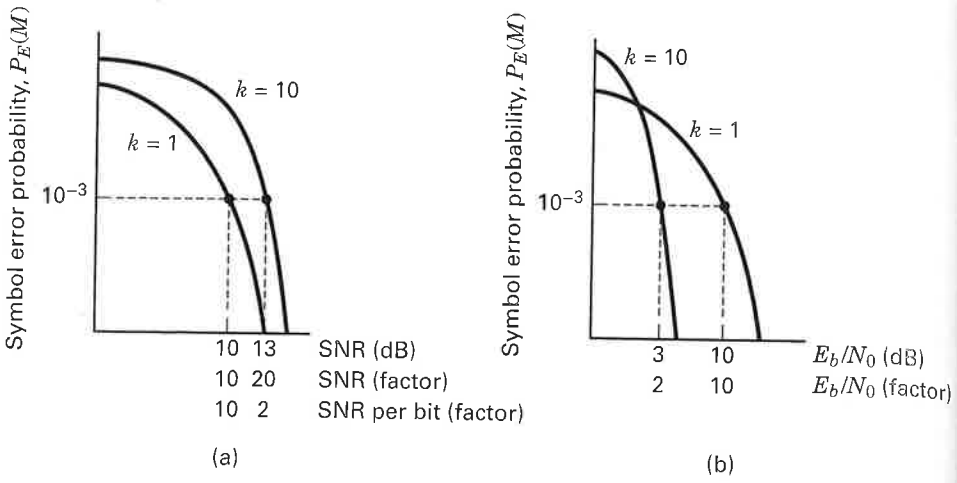


Figure 4.34 Mapping P_E versus SNR into P_E versus E_b/N_0 for orthogonal signaling. (a) Unnormalized. (b) Normalized.

$$\frac{E_b}{N_0} \approx \frac{S}{N}\left(\frac{1}{k}\right) \tag{4.104}$$

Figure 4.34 illustrates the mapping from P_E versus SNR to P_E versus E_b/N_0 for coherently detected M-ary orthogonal signaling, with "ballpark" numbers on the axes. In Figure 4.34a, on the $k = 1$ curve is shown an operating point corresponding to $P_E = 10^{-3}$ and SNR = 10 dB. On the $k = 10$ curve is shown an operating point at the same $P_E = 10^{-3}$ but with SNR = 13 dB (approximate values taken from Figure 4.33). Here we clearly see the degradation in error performance as k increases. To appreciate where the performance improvement comes from, let us convert the abscissa from the nonlinear scale of SNR in decibels to a linear one—SNR expressed as a factor. This is shown in Figure 4.34a as the factors 10 and 20 for the $k = 1$ and $k = 10$ cases, respectively. Next, we further convert the abscissa scale to SNR per bit (expressed as a factor). This is shown in Figure 4.34a as the factors 10 and 2 for the $k = 1$ and $k = 10$ cases, respectively. It is convenient to think of the 1024-ary symbol or waveform ($k = 10$ case) as being interchangeable with its 10-bit meaning. Thus, if the symbol requires 20 units of SNR then the 10 bits belonging to that symbol require that same 20 units; or, in other words, each bit requires 2 units.

Rather than performing such computations, we can simply map these same $k = 1$ and $k = 10$ cases onto the Figure 4.34b plane, representing P_E versus E_b/N_0. The $k = 1$ case looks exactly the same as it does in Figure 4.34a. But for the $k = 10$ case, there is a dramatic change. We can immediately see that signaling with the $k = 10$-bit symbol requires only 2 units (3 dB) of E_b/N_0 compared with 10 units (10 dB) for the binary symbol. The mapping that gives rise to the required E_b/N_0 for the $k = 10$ case is obtained from Equation (4.104) as follows: $E_b/N_0 = 20\ (1/10) = 2$ (or 3-dB), which shows the error performance improvement as k is increased. In digital communication systems, error performance is almost always considered in terms of

E_b/N_0, since such a measurement makes for a meaningful comparison between one system's performance and another. Therefore, the curves shown in Figures 4.33 and 4.34a are hardly ever seen.

Although Figure 4.33 is not often seen, we can still use it for gaining insight into why orthogonal signaling provides improved error performance as M or k increases. Let us consider the analogy of purchasing a commodity—say, grade A cottage cheese. The choice of the grade corresponds to some point on the P_E axis of Figure 4.33—say, 10^{-3}. From this point, construct a horizontal line through all of the curves (from $M = 2$ through $M = 1024$). At the grocery store we buy the very smallest container of cottage cheese, containing 2 ounces and costing \$1. On Figure 4.33 we can say that this purchase corresponds to our horizontal construct intercepting the $M = 2$ curve. We look down at the corresponding SNR and call the intercept on this axis our cost of \$1. The next time we purchase cottage cheese, we remember that the first purchase seemed expensive at 50 cents an ounce. So, we decide to buy a larger carton, containing 8 ounces and costing \$2. On Figure 4.33, we can say that this purchase corresponds to the point at which our horizontal construct intercepts the $M = 8$ curve. We look down at the corresponding SNR, and call this intercept our cost of \$2. Notice that we bought a larger container so the price went up, but because we bought a greater quantity, the price per ounce went down (the unit cost is now only 25 cents per ounce). We can continue this analogy by purchasing larger and larger containers so that the price of the container (SNR) keeps going up, but the price per ounce keeps going down. This is the age-old story called the *economy of scale.* Buying larger quantities at a time is commensurate with purchasing at the wholesale level; it makes for a lower unit price. Similarly, when we use orthogonal signaling with symbols that contain more bits, we need more power (more SNR), but the requirement per bit (E_b/N_0) is reduced.

4.9 SYMBOL ERROR PERFORMANCE FOR *M*-ARY SYSTEMS ($M > 2$)

4.9.1 Probability of Symbol Error for MPSK

For large energy-to-noise ratios, the symbol error performance $P_E(M)$, for equally likely, coherently detected M-ary PSK signaling, can be expressed [7] as

$$P_E(M) \approx 2Q\left(\sqrt{\frac{2E_s}{N_0}}\sin\frac{\pi}{M}\right) \tag{4.105}$$

where $P_E(M)$ is the probability of symbol error, $E_s = E_b(\log_2 M)$ is the energy per symbol, and $M = 2^k$ is the size of the symbol set. The $P_E(M)$ performance curves for coherently detected MPSK signaling are plotted versus E_b/N_0 in Figure 4.35.

The symbol error performance for differentially coherent detection of M-ary DPSK (for large E_s/N_0) is similarly expressed [7] as

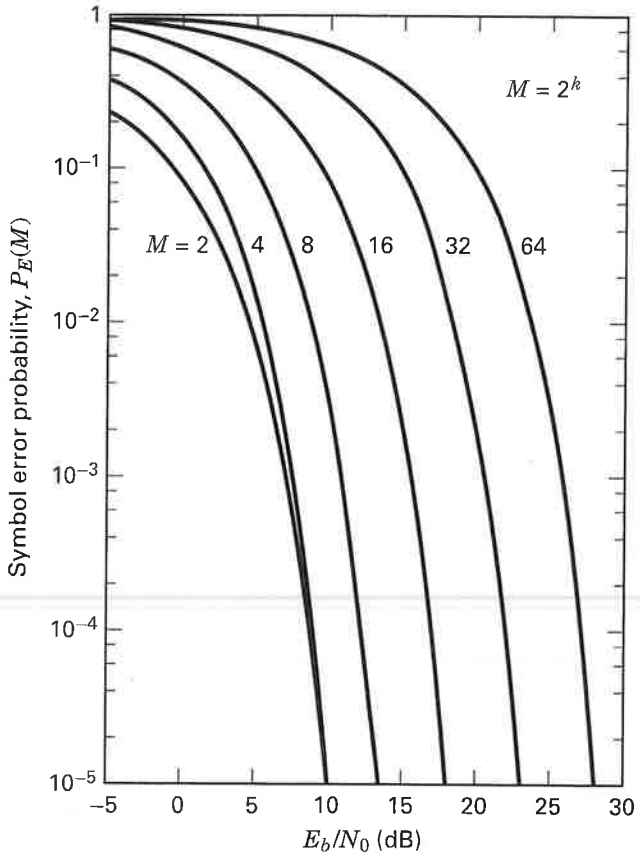


Figure 4.35 Symbol error probability for coherently detected multiple phase signaling. (Reprinted from W. C. Lindsey and M. K. Simon, *Telecommunication Systems Engineering,* Prentice-Hall, Inc. Englewood Cliffs, N.J., 1973, courtesy of W. C. Lindsey and Marvin K. Simon.)

$$P_E(M) \approx 2Q\left(\sqrt{\frac{2E_s}{N_0}}\sin\frac{\pi}{\sqrt{2}M}\right) \tag{4.106}$$

4.9.2 Probability of Symbol Error for MFSK

The symbol error performance $P_E(M)$, for equally likely, *coherently* detected M-ary orthogonal signaling can be upper bounded [5] as

$$P_E(M) \leq (M-1)\,Q\left(\sqrt{\frac{E_s}{N_0}}\right) \tag{4.107}$$

where $E_s = E_b(\log_2 M)$ is the energy per symbol and M is the size of the symbol set. The $P_E(M)$ performance curves for coherently detected M-ary orthogonal signaling are plotted versus E_b/N_0 in Figure 4.36.

The symbol error performance for equally likely, *noncoherently* detected M-ary orthogonal signaling is [9]

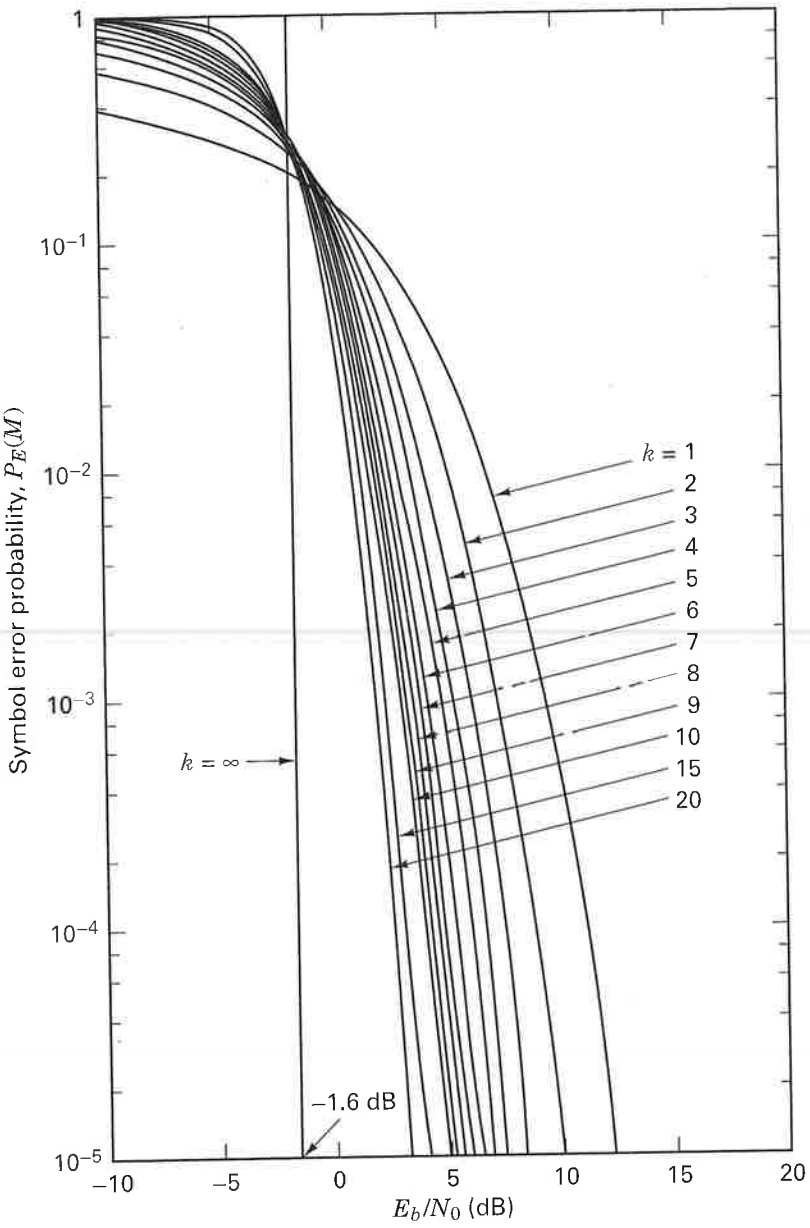


Figure 4.36 Symbol error probability for coherently detected M-ary orthogonal signaling. (Reprinted from W. C. Lindsey and M. K. Simon, *Telecommunication Systems Engineering*, Prentice-Hall, Inc., Englewod Cliffs, N.J., 1973, courtesy of W. C. Lindsey and Marvin K. Simon.)

$$P_E(M) = \frac{1}{M}\exp\left(-\frac{E_s}{N_0}\right)\sum_{j=2}^{M}(-1)^j\binom{M}{j}\exp\left(\frac{E_s}{jN_0}\right) \tag{4.108}$$

where

$$\binom{M}{j} = \frac{M!}{j!\,(M-j)!} \tag{4.109}$$

is the standard binomial coefficient yielding the number of ways in which j symbols out of M may be in error. Note that for binary case, Equation (4.108) reduces to

$$P_B = \frac{1}{2}\exp\left(-\frac{E_b}{2N_0}\right) \tag{4.110}$$

which is the same result as that described by Equation (4.96). The $P_E(M)$ performance curves for noncoherently detected M-ary orthogonal signaling are plotted versus E_b/N_0 in Figure 4.37. If we compare this noncoherent orthogonal $P_E(M)$ performance with the corresponding $P_E(M)$ results for the coherent detection of orthogonal signals in Figure 4.36, it can be seen that for $k > 7$, there is a negligible difference. An upper bound for coherent as well as noncoherent reception of orthogonal signals is [9]

$$P_E(M) < \frac{M-1}{2}\exp\left(\frac{E_s}{2N_0}\right) \tag{4.111}$$

where E_s is the energy per symbol and M is the size of the symbol set.

4.9.3 Bit Error Probability versus Symbol Error Probability for Orthogonal Signals

It can be shown [9] that the relationship between probability of bit error (P_B) and probability of symbol error (P_E) for an M-ary orthogonal signal set is

$$\frac{P_B}{P_E} = \frac{2^{k-1}}{2^k - 1} = \frac{M/2}{M-1} \tag{4.112}$$

In the limit as k increases, we get

$$\lim_{k\to\infty}\frac{P_B}{P_E} = \frac{1}{2}$$

A simple example will make Equation (4.112) intuitively acceptable. Figure 4.38 describes an octal message set. The message symbols (assumed equally likely) are to be transmitted on orthogonal waveforms such as FSK. With orthogonal signaling, a decision error will transform the correct signal into any one of the $(M-1)$ incorrect signals with equal probability. The example in Figure 4.38 indicates that the symbol comprising bits 0 1 1 was transmitted. An error might occur in any one of the other $2^k - 1 = 7$ symbols, with equal probability. Notice that just because a symbol error is made does not mean that all the bits within the symbol will be in error. In Figure 4.38, if the receiver decides that the transmitted symbol is the bot-

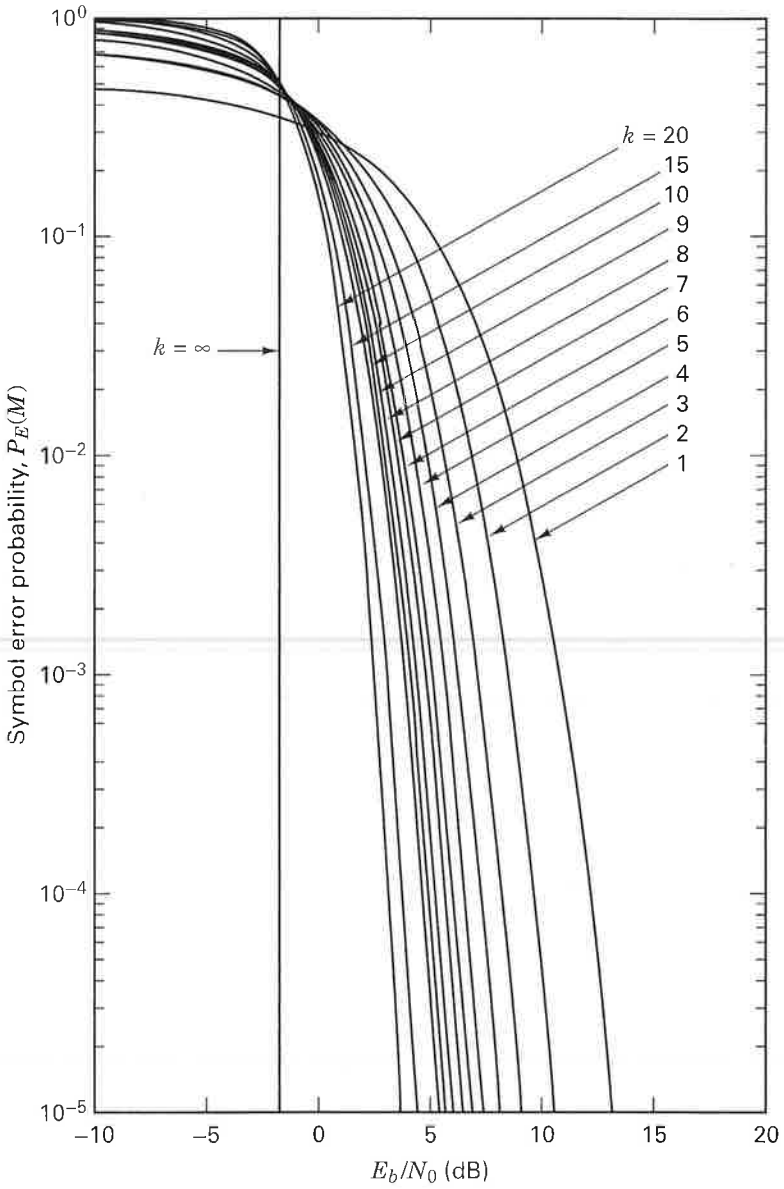


Figure 4.37 Symbol error probability for noncoherently detected *M*-ary orthogonal signaling. (Reprinted from W. C. Lindsey and M. K. Simon, *Telecommunication Systems Engineering,* Prentice-Hall, Inc., Englewood Cliffs, N.J., 1973, courtesy of W. C. Lindsey and Marvin K. Simon.)

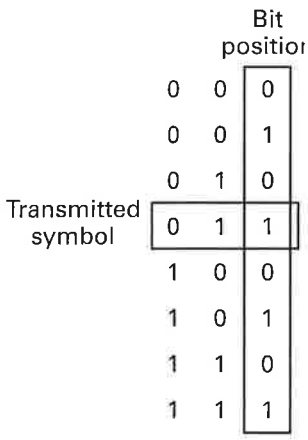


Figure 4.38 Example of P_B versus P_E.

tom one listed, comprising bits 1 1 1, two of the three transmitted symbol bits will be correct; only one bit will be in error. It should be apparent that for nonbinary signaling, P_B will always be less than P_E (keep in mind that P_B and P_E reflect the frequency of making errors on the *average*.)

Consider any of the bit-position columns in Figure 4.38. For each bit position, the digit occupancy consists of 50% ones and 50% zeros. In the context of the first bit position (rightmost column) and the transmitted symbol, how many ways are there to cause an error to the binary one? There are $2^{k-1} = 4$ ways (four places where zeros appear in the column) that a bit error can be made; it is the same for each of the columns. The final relationship P_B/P_E, for orthogonal signaling, in Equation (4.112), is obtained by forming the following ratio: the number of ways that a bit error can be made (2^{k-1}) divided by the number of ways that a symbol error can be made ($2^k - 1$). For the Figure 4.38 example, $P_B/P_E = 4/7$.

4.9.4 Bit Error Probability Versus Symbol Error Probability for Multiple Phase Signaling

For the case of MPSK signaling, P_B is less than or equal to P_E, just as in the case of MFSK signaling. However, there is an important difference. For orthogonal signaling, selecting any one of the $(M-1)$ erroneous symbols is equally likely. In the case of MPSK signaling, each signal vector is not equidistant from all of the others. Figure 4.39a illustrates an 8-ary decision space with the pie-shaped decision regions denoted by the 8-ary symbols in binary notation. If symbol (0 1 1) is transmitted, it is clear that should an error occur, the transmitted signal will most likely be mistaken for one of its closest neighbors, (0 1 0) or (1 0 0). The likelihood that (0 1 1) would get mistaken for (1 1 1) is relatively remote. If the assignment of bits to symbols follows the binary sequence shown in the symbol decision regions of Figure 4.39a, some symbol errors will usually result in two or more bit errors, even with a large signal-to-noise ratio.

For nonorthogonal schemes, such as MPSK signaling, one often uses a binary-to-M-ary code such that binary sequences corresponding to adjacent sym-

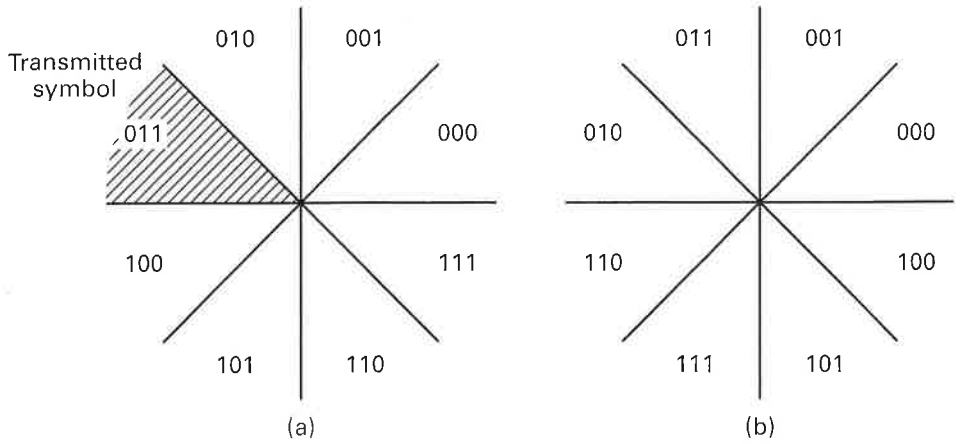


Figure 4.39 Binary-coded versus Gray-coded decision regions in an MPSK signal space. (a) Binary coded. (b) Gray coded.

bols (phase shifts) differ in only one bit position; thus when an M-ary symbol error occurs, it is more likely that only one of the k input bits will be in error. A code that provides this desirable feature is the Gray code [7]; Figure 4.39b illustrates the bit-to-symbol assignment using a Gray code for 8-ary PSK. Here it can be seen that neighboring symbols differ from one another in only one bit position. Therefore, the occurrence of a multibit error, for a given symbol error, is much reduced compared to the uncoded binary assignment seen in Figure 4.39a. Implementing such a Gray code, represents one of the few cases in digital communications where a benefit can be achieved without incurring any cost. The Gray code is simply an assignment that requires no special or additional circuitry. Utilizing the Gray code assignment, it can be shown [5] that

$$P_B \approx \frac{P_E}{\log_2 M} \qquad (\text{for } P_E \ll 1) \tag{4.113}$$

Recall from Section 4.8.4 that BPSK and QPSK signaling have the same bit error probability. Here, in Equation (4.113), we verify that they do not have the same symbol error probability. For BPSK, $P_E = P_B$. However, for QPSK, $P_E \approx 2P_B$.

An exact closed-form expression for the bit-error probability P_B of 8-ary PSK, together with tight upper and lower bounds on P_B for M-ary PSK with larger M, may be found in Lee [10].

4.9.5 Effects of Intersymbol Interference

In the previous sections and in Chapter 3 we have treated the detection of signals in the presence of AWGN under the assumption that there is no intersymbol interference (ISI). Thus the analysis has been straightforward, since the zero-mean AWGN process is characterized by its variance alone. In practice we find that ISI is

often a second source of interference which must be accounted for. As explained in Section 3.3, ISI can be generated by the use of bandlimiting filters at the transmitter output, in the channel, or at the receiver input. The result of this additional interference is to degrade the error probabilities for coherent as well as for noncoherent reception. Calculating error performance in the presence of ISI in addition to AWGN is much more complicated since it involves the impulse response of the channel. The subject will not be treated here; however, for those readers interested in the details of the analysis, References [11–16] should prove interesting.

4.10 CONCLUSION

We have catalogued some basic bandpass digital modulation formats, particularly phase shift keying (PSK) and frequency shift keying (FSK). We have considered a geometric view of signal vectors and noise vectors, particularly antipodal and orthogonal signal sets. This geometric view allows us to consider the detection problem in the light of an orthogonal signal space and signal regions. This view of the space, and the effect of noise vectors causing transmitted signals to be received in the incorrect region, facilitates the understanding of the detection problem and the performance of various modulation and demodulation techniques. In Chapter 9 we reconsider the subjects of modulation and demodulation, and we investigate some bandwidth-efficient modulation techniques.

REFERENCES

1. Schwartz, M., *Information, Transmission, Modulation, and Noise,* McGraw-Hill Book Company, New York, 1970.
2. Van Trees, H. L., *Detection, Estimation, and Modulation Theory,* Part 1, John Wiley & Sons, Inc., New York, 1968.
3. Park, J. H., Jr., "On Binary DPSK Detection," *IEEE Trans. Commun.,* vol. COM26, no. 4, Apr. 1978, pp. 484–486.
4. Ziemer, R. E., and Peterson, R. L., *Digital Communications and Spread Spectrum Systems,* Macmillan Publishing Company, Inc., New York, 1985.
5. Lindsey, W. C., and Simon, M. K., *Telecommunication Systems Engineering,* Prentice-Hall, Inc., Englewood Cliffs, N.J., 1973.
6. Whalen, A. D., *Detection of Signals in Noise,* Academic Press, Inc., New York, 1971.
7. Korn, I., *Digital Communications,* Van Nostrand Reinhold Company, Inc., New York, 1985.
8. Couch, L. W. II, *Digital and Analog Communication Systems,* Macmillan Publishing Company, New York, 1983.
9. Viterbi, A. J., *Principles of Coherent Communications,* McGraw-Hill Book Company, New York, 1966.
10. Lee, P. J., "Computation of the Bit Error Rate of Coherent M-*ary* PSK with Gray Code Bit Mapping," *IEEE Trans. Commun.,* vol. COM34, no. 5, May 1986, pp. 488–491.

11. Hoo, E. Y., and Yeh, Y. S., "A New Approach for Evaluating the Error Probability in the Presence of the Intersymbol Interference and Additive Gaussian Noise," *Bell Syst. Tech. J.,* vol. 49, Nov. 1970, pp. 2249–2266.
12. Shimbo, O., Fang, R. J., and Celebiler, M., "Performance of *M*-ary PSK Systems in Gaussian Noise and Intersymbol Interference," *IEEE Trans. Inf. Theory,* vol. IT19, Jan. 1973, pp. 44–58.
13. Prabhu, V. K., "Error Probability Performance of *M*-ary CPSK Systems with Intersymbol Interference," *IEEE Trans. Commun.,* vol. COM21, Feb. 1973, pp. 97–109.
14. Yao, K., and Tobin, R. M., "Moment Space Upper and Lower Error Bounds for Digital Systems with Intersymbol Interference," *IEEE Trans. Inf. Theory,* vol. IT22, Jan. 1976, pp. 65–74.
15. King, M. A., Jr., "Three Dimensional Geometric Moment Bounding Techniques," *J. Franklin Inst.,* vol. 309, no. 4, Apr. 1980, pp. 195–213.
16. Prabhu, V. K., and Salz, J., "On the Performance of Phase-Shift Keying Systems," *Bell Syst. Tech. J.,* vol. 60, Dec. 1981, pp. 2307–2343.

PROBLEMS

4.1. Find the expected number of bit errors made in one day by the following continuously operating coherent BPSK receiver. The data rate is 5000 bits/s. The input digital waveforms are $s_1(t) = A \cos \omega_0 t$ and $s_2(t) = -A \cos \omega_0 t$ where $A = 1$ mV and the single-sided noise power spectral density is $N_0 = 10^{-11}$ W/Hz. Assume that signal power and energy per bit are normalized relative to a 1-Ω resistive load.

4.2. A continuously operating coherent BPSK system makes errors at the average rate of 100 errors per day. The data rate is 1000 bits/s. The single-sided noise power spectral density is $N_0 = 10^{-10}$ W/Hz.

(a) If the system is ergodic, what is the average bit error probability?

(b) If the value of received average signal power is adjusted to be 10^{-6} W, will this received power be adequate to maintain the error probability found in part (a)?

4.3. If a system's main performance criterion is bit error probability, which of the following two modulation schemes would be selected for an AWGN channel? Show computations.

Binary noncoherent orthogonal FSK with $E_b/N_0 = 13$ dB

Binary coherent PSK with $E_b/N_0 = 8$ dB

4.4. The bit stream

1 0 1 0 1 0 1 1 1 1 0 1 0 1 0 1 0 0 0 0 1 1 1 1

is to be transmitted using DPSK modulation. Show four different differentially encoded sequences that can represent the data sequence above, and explain the algorithm that generated each.

4.5. (a) Calculate the minimum required bandwidth for a noncoherently detected orthogonal binary FSK system. The higher-frequency signaling tone is 1 MHz and the symbol duration is 1 ms.

(b) What is the minimum required bandwidth for a noncoherent MFSK system having the same symbol duration?

4.6. Consider a BPSK system with equally likely waveforms $s_1(t) = \cos \omega_0 t$ and $s_2(t) = -\cos \omega_0 t$. Assume that the received $E_b/N_0 = 9.6$ dB, giving rise to a bit-error probability of 10^{-5}, when the synchronization is perfect. Consider that carrier recovery with the PLL suffers some fixed error ϕ associated with the phase estimate, so that the reference signals are expressed as $\cos(\omega_0 t + \phi)$ and $-\cos(\omega_0 t + \phi)$. Note that the error-degradation effect of a fixed known bias can be computed by using the closed-end relationships presented in this chapter. However, if the phase error were to consist of a random jitter, computing its effect would then require a stochastic treatment. (See Chapter 10.)

(a) How badly does the bit-error probability degrade when $\phi = 25°$?

(b) How large a phase error would cause the bit-error probability to degrade to 10^{-3}?

4.7. Find the probability of bit error, P_B, for the coherent matched filter detection of the equally likely binary FSK signals

$$s_1(t) = 0.5 \cos 2000\pi t$$

and

$$s_2(t) = 0.5 \cos 2020\pi t$$

where the two-sided AWGN power spectral density is $N_0/2 = 0.0001$. Assume that the symbol duration is $T = 0.01$ s.

4.8. Find the optimum (minimum probability of error) threshold γ_0, for detecting the equally likely signals $s_1(t) = \sqrt{2E/T} \cos \omega_0 t$ and $s_2(t) = \sqrt{\frac{1}{2}E/T} \cos(\omega_0 t + \pi)$ in AWGN, using a correlator receiver as shown in Figure 4.7b. Assume a reference signal of $\psi_1(t) = \sqrt{2/T} \cos \omega_0 t$.

4.9. A system using matched filter detection of equally likely BPSK signals, $s_1(t) = \sqrt{2E/T} \cos \omega_0 t$ and $s_2(t) = \sqrt{2E/T} \cos(\omega_0 t + \pi)$, operates in AWGN with a received E_b/N_0 of 6.8 dB. Assume that $\mathbf{E}\{z(T)\} = \pm \sqrt{E}$.

(a) Find the minimum probability of bit error, P_B, for this signal set and E_b/N_0.

(b) If the decision threshold is $\gamma = 0.1\sqrt{E}$, find P_B.

(c) The threshold of $\gamma = 0.1\sqrt{E}$ is optimum for a particular set of a priori probabilities, $P(s_1)$ and $P(s_2)$. Find the values of these probabilities (refer to Section B.2).

4.10. (a) Describe the impulse response of a matched filter for detecting the discrete signal shown in Figure P4.1. With this signal at the input to the filter, show the output as a function of time. Neglect the effects of noise. What is the maximum output value?

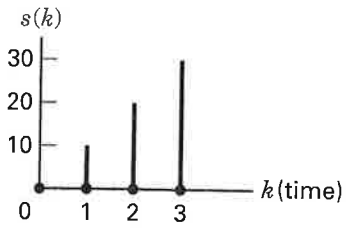


Figure P4.1

(b) In a matched filter (MF), a signal is convolved with a time-reversed function of the signal (impulse response of the MF). Convolution reverses the function again; thus, the MF yields the correlation of a signal and its "look alike" copy (even though the MF operation is designated as convolution). In implementing a

MF, suppose that you accidently connect the circuitry so as to yield the correlation of a signal and its time-reversed copy; the output would then yield a signal convolved with itself. Show the output as a function of time. What is the maximum output value? Note that the maximum output value for part (a) will occur at a different time index compared with that of part (b).

(c) By examining the filter's output values from the flawed circuit of part (b), compared with the correct values in part (a), can you find a clue that can help you predict when such an output sequence appears to be a valid matched filter output, and when it does not?

(d) If noise were added to the signal, compare the output SNR for the correlator versus the convolver. Also, if the input consists of noise only, compare the output of the correlator versus the convolver.

4.11. A binary source with equally likely symbols controls the switch position in a transmitter operating over an AWGN channel, as shown in Figure P4.2. The noise has two-sided spectral density $N_0/2$. Assume antipodal signals of time duration T seconds and energy E joules. The system clock produces a clock pulse every T seconds, and the binary source rate is $1/T$ bits/s. Under *normal* operation, the switch is up when the source produces a binary zero, and it is down when the source produces a binary one. However, the switch is *faulty*. With probability p, it will be thrown in the wrong direction during a given T-second interval. The presence of a switch error during any interval is independent of the presence of a switch error at any other time. Assume that $\mathbf{E}\{z(T)\} = \pm\sqrt{E}$.

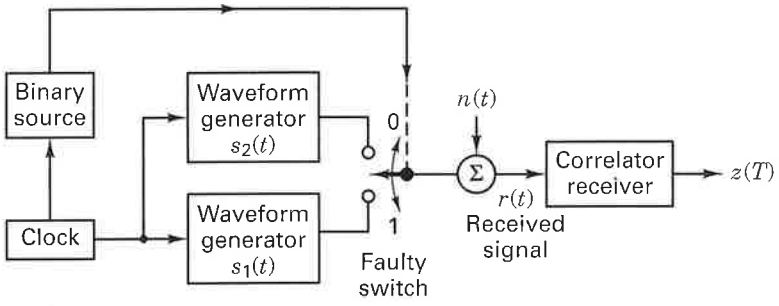


Figure P4.2

(a) Sketch the conditional probability functions, $p(z|s_1)$ and $p(z|s_2)$.

(b) The correlator receiver observes $r(t)$ in the interval $(0, T)$. Sketch the block diagram of an optimum receiver for minimizing the bit error probability when it is known that the switch is faulty with probability, p.

(c) Which one of the following two systems would you prefer to have?

$$p = 0.1 \text{ and } \frac{E_b}{N_0} = \infty$$

or

$$p = 0 \text{ and } \frac{E_b}{N_0} = 7 \text{ dB}$$

4.12. **(a)** Consider a 16-ary PSK system with symbol error probability $P_E = 10^{-5}$. A Gray code is used for the symbol to bit assignment. What is the approximate bit error probability?

(b) Repeat part (a) for a 16-ary orthogonal FSK system.

4.13. Consider a coherent orthogonal MFSK system with $M = 8$ having the equally likely waveforms $s_i(t) = A \cos 2\pi f_i t$, $i = 1, \ldots, M$, $0 \le t \le T$, where $T = 0.2$ ms. The received carrier amplitude, A, is 1 mV, and the two-sided AWGN spectral density, $N_0/2$, is 10^{-11} W/Hz. Calculate the probability of bit error, P_B.

4.14. A bit error probability of $P_B = 10^{-3}$ is required for a system with a data rate of 100 kbits/s to be transmitted over an AWGN channel using coherently detected MPSK modulation. The system bandwidth is 50 kHz. Assume that the system frequency transfer function is a raised cosine with a roll-off characteristic of $r = 1$ and that a Gray code is used for the symbol to bit assignment.

(a) What E_s/N_0 is required for the specified P_B?

(b) What E_b/N_0 is required?

4.15. A differentially coherent MPSK system operates over an AWGN channel with an E_b/N_0 of 10 dB. What is the symbol error probability for $M = 8$ and equally likely symbols?

4.16. If a system's main performance criterion is bit-error probability, which of the following two modulation schemes would be selected for transmission over an AWGN channel?

$$\text{coherent 8-ary orthogonal FSK with } \frac{E_b}{N_0} = 8 \text{ dB}$$

or

$$\text{coherent 8-ary PSK with } \frac{E_b}{N_0} = 13 \text{ dB}$$

(Assume that a Gray code is used for the MPSK symbol-to-bit assignment, and show computations.)

4.17. Consider that a BPSK demodulator/detector has a synchronization error consisting of a time bias pT, where p is a fraction ($0 \le p \le 1$) of the symbol time T. In other words, the detection of a symbol starts early (late) and concludes early (late) by an amount pT. Assume equally likely signaling and perfect frequency and phase synchronization. Note that the error-degradation effect of a fixed known bias can be computed by using the closed-end relationships presented in this chapter. However, if the timing error were to consist of a random jitter, computing its effect would then require a stochastic treatment. (See Chapter 10.)

(a) Find the general expression for bit-error probability P_b as a function of p.

(b) If the received $E_b/N_0 = 9.6$ dB and $p = 0.2$, compute the value of degraded P_b due to the timing bias.

(c) If one did not compensate for the timing bias in this example, how much additional E_b/N_0 in dB must be provided in order to restore the P_b that exists when $p = 0$?

4.18. Using all of the stated specifications, repeat problem 4.17 for the case of coherently detected, binary frequency-shift-keying (BFSK) modulation.

4.19. Consider that a BPSK demodulator/detector has a synchronization error consisting of a time bias pT, where p is a fraction ($0 \le p \le 1$) of the symbol time T. Consider that

there is also a constant phase-estimation error ϕ. Assume equally-likely signaling and perfect frequency synchronization.

(a) Find the general expression for bit-error probability P_b as a function of p and ϕ.

(b) If received $E_b/N_0 = 9.6$ dB, $p = 0.2$, and $\phi = 25°$, compute the value of degraded P_b due to the combined effects of timing and phase bias.

(c) If one did not compensate for the biases in this example, how much additional E_b/N_0 in dB must be provided in order to restore the P_b that exists when $p = 0$ and $\phi = 0°$?

4.20. Correlating to a known Barker sequence is an often used synchronization technique, since the Barker sequence yields a prominent correlation peak when properly synchronized, and a small correlation output when not synchronized. Using the short Barker sequence 1 0 1 1 1, where the leftmost bit is the earliest bit, devise a discrete matched filter similar to the one in Figure 4.10 that is matched to this sequence. Verify its usefulness for synchronization by plotting the output versus input as a function of time, when the input is the 1 0 1 1 1 sequence.

QUESTIONS

4.1. At what location in the system is E_b/N_0 defined? (See Section 4.3.2.)

4.2. Amplitude- or phase-shift keying is visualized as a constellation of points or phasors on a plane. Why can't we use a similarly simple visualization for orthogonal signaling such as FSK? (See Section 4.4.4.)

4.3. In the case of MFSK signaling, what is the minimum tone spacing that insures signal *orthogonality?* (See Section 4.5.4.)

4.4. What benefits are there in using *complex notation* for representing sinusoids? (See Sections 4.2.1 and 4.6.)

4.5. Digital modulation schemes fall into one of the two classes with opposite behavior characteristics: *orthogonal* signaling, and *phase/amplitude* signaling. Describe the behavior of each class. (See Sections 4.8.2.)

4.6. Why do binary phase shift keying (BPSK) and quaternary phase shift keying (QPSK) manifest the same bit-error-probability relationship? (See Section 4.8.4.)

4.7. In the case of multiple-phase shift keying (MPSK), why does *bandwidth efficiency* improve with higher dimensional signaling? (See Sections 4.8.2 and 4.8.3.)

4.8. In the case of orthogonal signaling such as MFSK, why does *error-performance* improve with higher dimensional signaling? (See Section 4.8.5.)

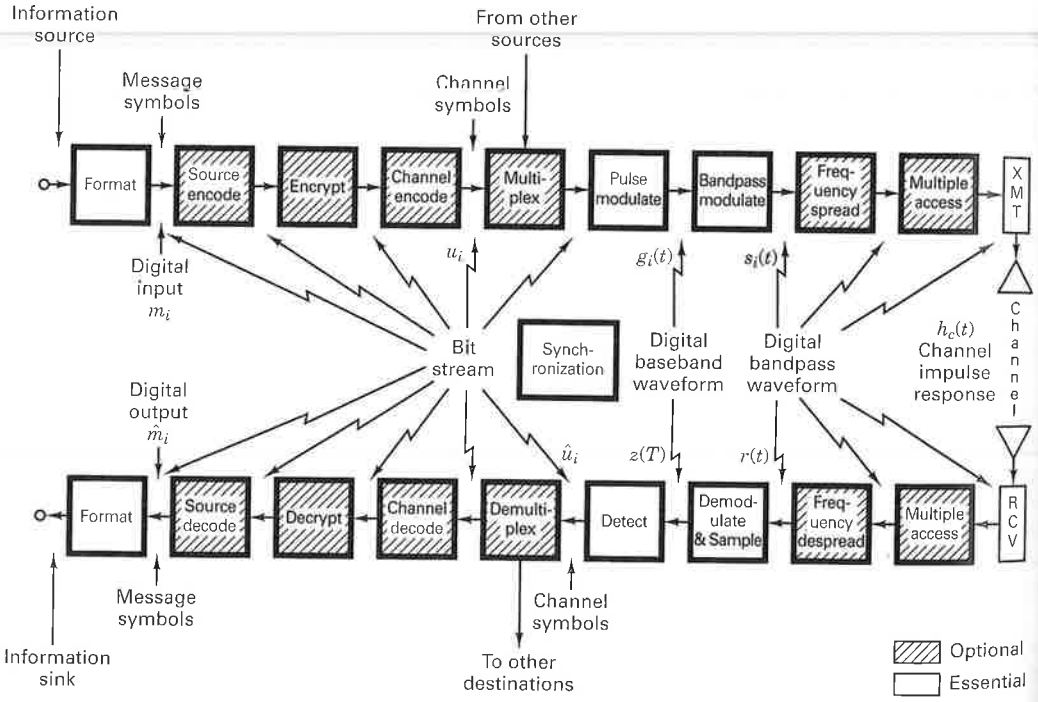
4.9. The use of a Gray code for assigning bits to symbols, represents one of the few cases in digital communications where a benefit can be achieved *free-of-charge.* Explain why there is no cost. (See Section 4.9.4.)

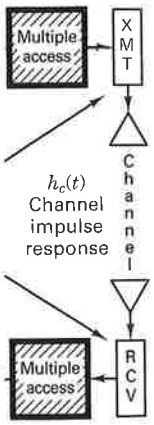
EXERCISES

Using the Companion CD, run the exercises associated with Chapter 4.

CHAPTER 5

Communications Link Analysis





5.1 WHAT THE SYSTEM LINK BUDGET TELLS THE SYSTEM ENGINEER

When we talk about a communications *link,* to what part of the system are we referring? Is it simply the channel or region between the transmitter and receiver? No, it is far more than that. The link encompasses the entire communications path, from the information source, through all the encoding and modulation steps, through the transmitter and the channel, up to and including the receiver with all its signal processing steps, and terminating at the information sink.

What is a link analysis, and what purpose does it serve in the development of a communication system? The link analysis, and its output, the *link budget,* consist of the calculations and tabulation of the useful signal power and the interfering noise power available at the receiver. The link budget is a balance sheet of gains and losses; it outlines the detailed apportionment of transmission and reception resources, noise sources, signal attenuators, and effects of processes throughout the link. Some of the budget parameters are statistical (e.g., allowances for the fading of signals as described in Chapter 15). The budget is an *estimation* technique for evaluating communication system error performance. In Chapters 3 and 4 we examined probability of error versus E_b/N_0 curves having a "waterfall-like" shape, such as the one shown in Figure 3.6. We thereby related error probability to E_b/N_0 for various modulation types in Gaussian noise. Once a modulation scheme has been chosen, the requirement to meet a particular error probability dictates a particular operating point on the curve; in other words, the required error perfor-

mance dictates the value of E_b/N_0 that must be made available at the receiver in order to meet that performance. The primary purpose of a link analysis is to determine the *actual* system operating point in Figure 3.6 and to establish that the error probability associated with that point is less than or equal to the system requirement. Of the many specifications, analyses, and tabulations that are used in the development of a communication system, the link budget stands out as a basic tool for providing the system engineer with overall system insight.

By examining the link budget, one can learn many things about overall system design and performance. For example, from the link margin, one learns whether the system will meet many of its requirements comfortably, marginally, or not at all. The link budget may reveal if there are any hardware constraints, and whether such constraints can be compensated for in other parts of the link. The link budget is often used as a "score sheet" in considering system trade-offs and configuration changes, and in understanding subsystem nuances and interdependencies. From a quick examination of the link budget and its supporting documentation, one can judge whether the analysis was done precisely or if it represents a rough estimate. Together with other modeling techniques, the link budget can help predict equipment weight, size, prime power requirements, technical risk, and cost. The link budget is one of the system manager's most useful documents; it represents the "bottom-line" tally in the search for optimimum system performance.

5.2 THE CHANNEL

The propagating medium or electromagnetic path connecting the transmitter and receiver is called the *channel.* In general, a communications channel might consist of wires, coaxial cables, fiber optic cables, and in the case of radio-frequency (RF) links, waveguides, the atmosphere, or empty space. For most terrestrial communication links, the channel space is occupied by the atmosphere and partially bounded by the earth's surface. For satellite links, the channel is occupied mostly by empty space. Although some atmospheric effects occur at altitudes up to 100 km, the *bulk* of the atmosphere extends to an altitude of 20 km. Therefore, only a small part (0.05%) of the total synchronous altitude (35,800 km) path is occupied by significant amounts of atmosphere. Most of this chapter presents link analysis in the context of such a satellite communications link. In Chapter 15, the link budget issues are extended to terrestrial wireless links.

5.2.1 The Concept of Free Space

The concept of *free space* assumes a channel free of all hindrances to RF propagation, such as absorption, reflection, refraction, or diffraction. If there is any atmosphere in the channel, it must be perfectly uniform and meet all these conditions. Also, we assume that the earth is infinitely far away or that its reflection coefficient is negligible. The RF energy arriving at the receiver is assumed to be a function only of distance from the transmitter (following the inverse-square law as used in

optics). A free-space channel characterizes an ideal RF propagation path; in practice, propagation through the atmosphere and near the ground results in absorption, reflection, diffraction and scattering, which modify the free-space transmission. Atmospheric absorption is treated in later sections. Reflection, diffraction, and scattering, which play an important role in determining terrestrial communications performance, are treated in Chapter 15. Also, Panter [1] provides a comprehensive treatment of these mechanisms.

5.2.2 Error-Performance Degradation

In Chapter 3, it was established that there are two primary causes for degradation of error performance. The first is loss in signal-to-noise ratio. The second is signal distortion as might be caused by intersymbol interference (ISI). In Chapters 3 and 15, some of the equalization techniques to counter the degradation effect of ISI are treated. In this chapter, we are concerned with the "bookkeeping" of the gains and losses for signal power and interfering power. ISI will not be included in the link budget because the uniqueness of ISI is that an increase in signal power will not always mitigate the degradation that it causes. (See Section 3.3.2.)

For digital communications, error performance depends on the received E_b/N_0, which was defined in Equation (3.30) as

$$\frac{E_b}{N_0} = \frac{S}{N}\left(\frac{W}{R}\right)$$

In other words, E_b/N_0 is a measure of normalized signal-to-noise ratio (*S/N* or SNR). Unless otherwise stated, SNR refers to *average* signal power and *average* noise power. The signal can be an information signal, a baseband waveform, or a modulated carrier. The SNR can degrade in two ways: (1) through the decrease of the desired signal power, and (2) through the increase of noise power, or the increase of interfering signal power. Let us refer to these degradations as *loss* and *noise* (or *interference*), respectively. Losses occur when a portion of the signal is absorbed, diverted, scattered, or reflected along its route to the intended receiver; thus a portion of the transmitted energy does not arrive at the receiver. There are several sources of interfering electrical noise and interference produced by a variety of mechanisms and sources, such as thermal noise, galaxy noise, atmospheric noise, switching transients, intermodulation noise, and interfering signals from other sources. Industry usage of the terms *loss* and *noise* frequently confuses the underlying degradation mechanism; however, the net effect on the SNR is the same.

5.2.3 Sources of Signal Loss and Noise

Figure 5.1 is a block diagram of a satellite communications link, emphasizing the sources of signal loss and noise. In the figure a signal loss is distinguished from a noise source by a dot pattern or line pattern, respectively. The contributors of *both* signal loss *and* noise are identified by a crosshatched line pattern. The following list of 21

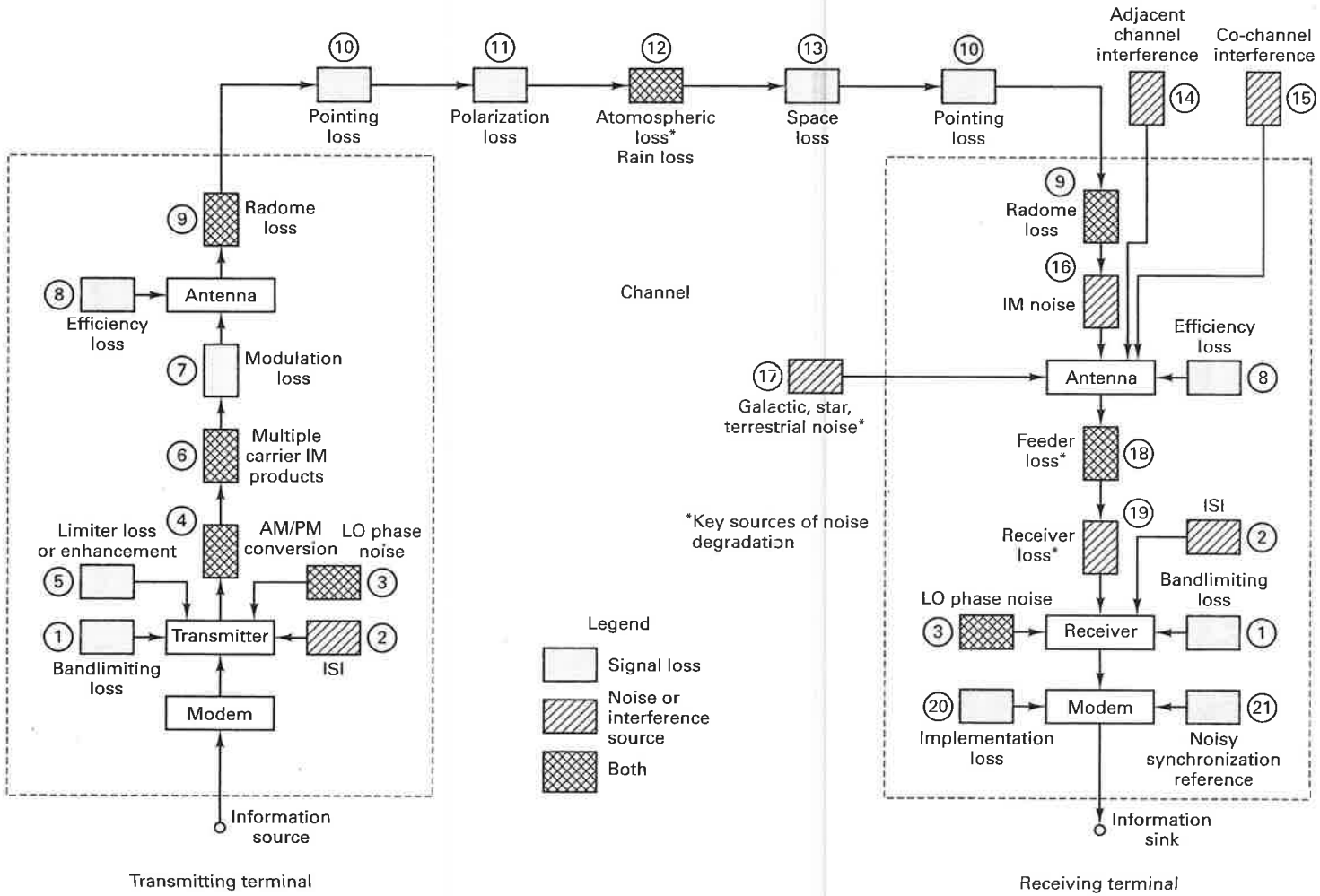


Figure 5.1 Satellite transmitter-to-receiver link with *typical loss and noise sources.*

sources of degradation represents a partial catalog of the major contributors to SNR degradation. The numbers correspond to the numbered circles in Figure 5.1.

1. *Bandlimiting loss.* All systems use filters in the transmitter to ensure that the transmitted energy is confined to the allocated or assigned bandwidth. This is to avoid interfering with other channels or users and to meet the requirements of regulatory agencies. Such filtering reduces the total amount of energy that would otherwise have been transmitted; the result is a *loss* in signal.
2. *Intersymbol interference (ISI).* As discussed in Chapter 3, filtering throughout the system—in the transmitter, in the receiver, and in the channel—can result in ISI. The received pulses overlap one another; the tail of one pulse "smears" into adjacent symbol intervals so as to *interfere* with the detection process. Even in the absence of thermal noise, imperfect filtering, system bandwidth constraints, and fading channels lead to ISI degradation.
3. *Local oscillator (LO) phase noise.* When an LO is used in signal mixing, phase fluctuations or jitter adds phase *noise* to the signal. When used as the reference signal in a receiver correlator, phase jitter can cause detector degradation and hence signal *loss.* At the transmitter, phase jitter can cause out-of-band signal spreading, which, in turn, will be filtered out and cause a *loss* in signal.
4. *AM/PM conversion.* AM-to-PM conversion is a phase *noise* phenomenon occurring in nonlinear devices such as traveling-wave tubes (TWT). Signal amplitude fluctuations (amplitude modulation) produce phase variations that contribute phase *noise* to signals that will be coherently detected. AM-to-PM conversion can also cause extraneous sidebands, resulting in signal *loss.*
5. *Limiter loss or enhancement.* A hard limiter can enhance the stronger of two signals, and suppress the weaker; this can result in either a signal *loss* or a signal *gain* [2].
6. *Multiple-carrier intermodulation (IM) products.* When several signals having different carrier frequencies are simultaneously present in a nonlinear device, such as a TWT, the result is a multiplicative interaction between the carrier frequencies which can produce signals at all combinations of sum and difference frequencies. The energy apportioned to these spurious signals (intermodulation or IM products) represents a *loss* in signal energy. In addition, if these IM products appear within the bandwidth region of these or other signals, the effect is that of added *noise* for those signals.
7. *Modulation loss.* The link budget is a calculation of received useful power (or energy). Only the power associated with information-bearing signals is useful. Error performance is a function of energy per transmitted symbol. Any power used for transmitting the carrier rather than the modulating signal (symbols) is a modulation *loss.* (However, energy in the carrier may be useful in aiding synchronization.)
8. *Antenna efficiency.* Antennas are transducers that convert electronic signals into electromagnetic fields, and vice versa. They are also used to focus the

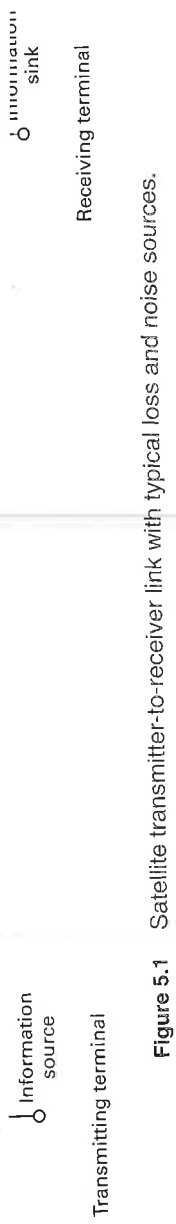


Figure 5.1 Satellite transmitter-to-receiver link with typical loss and noise sources.

electromagnetic energy in a desired direction. The larger the antenna aperture (area), the larger is the resulting signal power density in the desired direction. An antenna's efficiency is described by the ratio of its effective aperture to its physical aperture. Mechanisms contributing to a reduction in efficiency (*loss* in signal strength) are known as amplitude tapering, aperture blockage, scattering, re-radiation, spillover, edge diffraction, and dissipative loss [3]. Typical efficiencies due to the combined effects of these mechanisms range between 50 and 80%.

9. *Radome loss and noise.* A radome is a protective cover, used with some antennas, for shielding against weather effects. The radome, being in the path of the signal, will scatter and absorb some of the signal energy, thus resulting in a signal *loss.* A basic law of physics holds that a body capable of absorbing energy also radiates energy (at temperatures above 0 K). Some of this energy falls in the bandwidth of the receiver and constitutes injected *noise.*
10. *Pointing loss.* There is a *loss* of signal when either the transmitting antenna or the receiving antenna is imperfectly pointed.
11. *Polarization loss.* The polarization of an electromagnetic (EM) field is defined as the direction in space along which the field lines point, and the polarization of an antenna is described by the polarization of its radiated field. There is a *loss* of signal due to any polarization mismatch between the transmitting and receiving antennas.
12. *Atmospheric loss and noise.* The atmosphere is responsible for signal loss and is also a contributor of unwanted noise. The bulk of the atmosphere extends to an altitude of approximately 20 km; yet within that relatively short path, important loss and noise mechanisms are at work. Figure 5.2 is a plot of the theoretical one-way attenuation from a specified height to the top of the atmosphere. The calculations were made for several heights (0 km is sea level) and for a water vapor content of 7.5 g/m^3 at the earth's surface. The magnitude of signal *loss* due to oxygen (O_2) and water vapor absorption is plotted as a function of carrier frequency. Local maxima of attenuation occur in the vicinities of 22 GHz (water vapor), and 60 and 120 GHz (O_2). The atmosphere also contributes *noise* energy into the link. As in the case of the radome, molecules that absorb energy also radiate energy. The oxygen and water vapor molecules radiate noise throughout the RF spectrum. The portion of this noise that falls within the bandwidth of a given communication system will degrade its SNR. A primary atmospheric cause of signal *loss* and contributor of *noise* is rainfall. The more intense the rainfall, the more signal energy it will absorb. Also, on a day when rain passes through the antenna beam, there is a larger amount of atmospheric noise radiated into the system receiver than there is on a clear day. More will be said about atmospheric noise in later sections.
13. *Space loss.* There is a decrease in the electric field strength, and thus in signal strength (power density or flux density), as a function of distance. For a satellite communications link, the space loss is the largest single *loss* in the

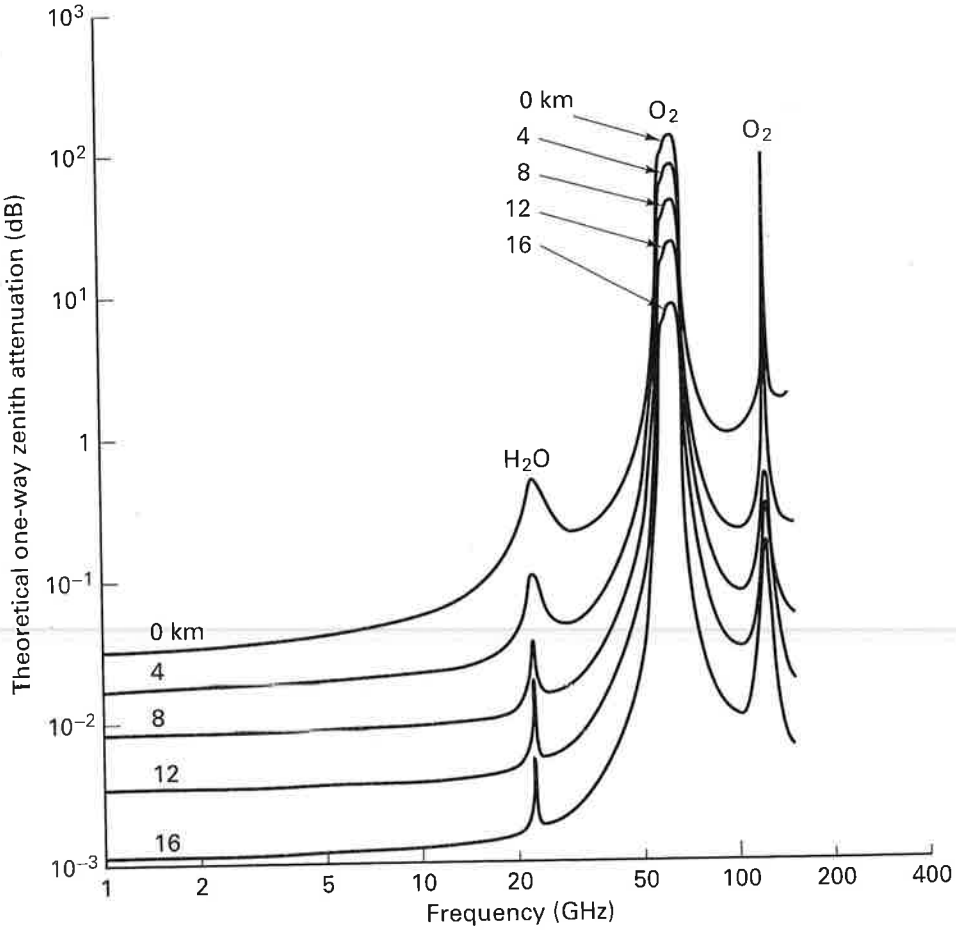


Figure 5.2 Theoretical vertical one-way attenuation from specified height to top of atmosphere for 7.5 g/m^3 of water vapor at the surface. (Does not include effect of rain or cloud attenuation.) (Reprinted from NASA Reference Publication 1082(03), "Propagation Effects Handbook for Satellite Systems Design," June 1983, Fig. 6.2-1, p. 218, courtesy of the National Aeronautics and Space Administration.)

system. It is a loss in the sense that all the radiated energy is not focused on the intended receiving antenna.

14. *Adjacent channel interference.* This *interference* is characterized by unwanted signals from other frequency channels "spilling over" or injecting energy into the channel of interest. The proximity with which channels can be located in frequency is determined by the modulation spectral roll-off and the width and shape of the main spectral lobe.
15. *Co-channel interference.* This *interference* refers to the degradation caused by an interfering waveform appearing within the signal bandwidth. It can be intro-

duced by a variety of ways, such as accidental transmissions, insufficient vertical and horizontal polarization discrimination, or by radiation spillover from an antenna sidelobe (low-energy beam surrounding the main antenna beam). It can be brought about by other authorized users of the same spectrum.

16. *Intermodulation (IM) noise.* The IM products described in item 6 result from multiple-carrier signals interacting in a nonlinear device. Such IM products are sometimes called *active intermods;* as described in item 6, they can either cause a loss in signal energy or be responsible for noise injected into a link. Here we consider *passive intermods;* these are caused by multiple-carrier transmission signals interacting with nonlinear components at the transmitter output. These nonlinearities generally occur at the junction of waveguide coupling joints, at corroded surfaces, and at surfaces having poor electrical contact. When large EM fields impinge on surfaces that have a diode-like transfer function (work potential), they cause multiplicative products, and hence *noise.* If such noise radiates into a closely located receiving antenna, it can seriously degrade the receiver performance.
17. *Galactic or cosmic, star, and terrestrial noise.* All the celestial bodies, such as the stars and the planets, radiate energy. Such *noise* energy in the field of view of the antenna will degrade the SNR.
18. *Feeder line loss.* The level of the received signal might be very small (e.g., 10^{-12} W), and thus will be particularly susceptible to noise degradation. The receiver front end, therefore, is a region where great care is taken to keep the noise as small as possible until the signal has been suitably amplified. The waveguide or cable (feeder line) between the receiving antenna and the receiver front end contributes both signal *attenuation* and thermal *noise;* the details are treated in Section 5.5.3.
19. *Receiver noise.* This is the thermal *noise* generated within the receiver; the details are treated in Sections 5.5.1 to 5.5.4.
20. *Implementation loss.* This *loss* in performance is the difference between theoretical detection performance and the actual performance due to imperfections such as timing errors, frequency offsets, finite rise and fall times of waveforms, and finite-value arithmetic.
21. *Imperfect synchronization reference.* When the carrier phase, the subcarrier phase, and the symbol timing references are all derived perfectly, the error probability is a well-defined function of E_b/N_0 discussed in Chapters 3 and 4. In general, they are not derived perfectly, resulting in a system *loss.*

5.3 RECEIVED SIGNAL POWER AND NOISE POWER

5.3.1 The Range Equation

The main purpose of the link budget is to verify that the communication system will operate according to plan—that is, the message quality (error performance) will meet the specifications. The link budget monitors the "ups" and "downs"

(gains and losses) of the transmitted signal, from its inception at the transmitter until it is ultimately received at the receiver. The compilation yields how much E_b/N_0 is received and how much safety margin exists beyond what is required. The process starts with the *range equation,* which relates power received to the distance between transmitter and receiver. We develop the range equation below.

In radio communication systems, the carrier wave is propagated from the transmitter by the use of a transmitting antenna. The transmitting antenna is a transducer that converts electronic signals into electromagnetic (EM) fields. At the receiver, a receiving antenna performs the reverse function; it converts EM fields into electronic signals. The development of the fundamental power relationship between the receiver and transmitter usually begins with the assumption of an omnidirectional RF source, transmitting uniformly over 4π steradians. Such an ideal source, called an *isotropic radiator,* is illustrated in Figure 5.3. The power density $p(d)$ on a hypothetical sphere at a distance d from the source is related to the transmitted power P_t by

$$p(d) = \frac{P_t}{4\pi d^2} \qquad \text{watts/m}^2 \tag{5.1}$$

since $4\pi d^2$ is the area of the sphere. For d much greater than the propagation wavelength (known as the *far field*), the power extracted at a receiving antenna is

$$P_r = p(d)A_{er} = \frac{P_t A_{er}}{4\pi d^2} \tag{5.2}$$

where the parameter A_{er} is the absorption cross section (effective area) of the receiving antenna, defined by

$$A_{er} = \frac{\text{total power extracted}}{\text{incident power flux density}} \tag{5.3}$$

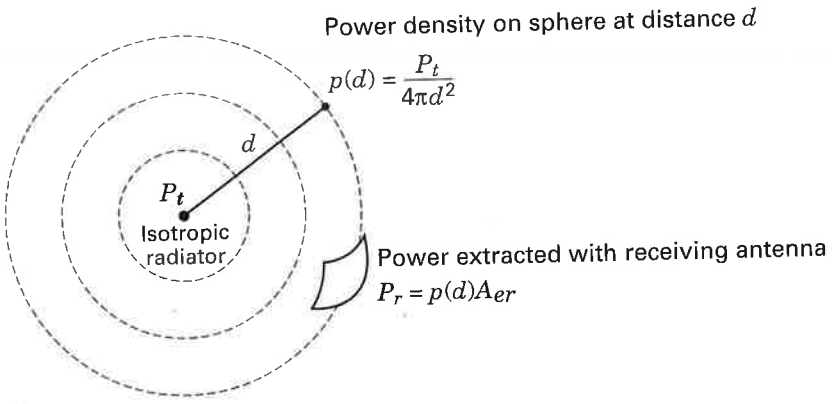


Figure 5.3 Range equation. Expresses received power in terms of distance.

If the antenna under consideration is a transmitting antenna, its effective area is designated by A_{et}. If the antenna in question is unspecified as to its receiving or transmitting function, its effective area is designated simply by A_e.

An antenna's effective area A_e and physical area A_p are related by an efficiency parameter η as

$$A_e = \eta A_p \tag{5.4}$$

which accounts for the fact that the total incident power is not extracted; it is lost through various mechanisms [3]. Nominal values for η are 0.55 for a dish (parabolic-shaped reflector) and 0.75 for a horn-shaped antenna.

The antenna parameter that relates the power output (or input) to that of an isotropic radiator as a purely geometric ratio is the antenna directivity or *directive gain*

$$G = \frac{\text{maximum power intensity}}{\text{average power intensity over } 4\pi \text{ steradians}} \tag{5.5}$$

In the absence of any dissipative loss or impedance mismatch loss, the antenna *gain* (in the direction of maximum intensity) is defined simply as the directive gain in Equation (5.5). However, in the event that there exists some dissipative or impedance mismatch loss, the antenna gain is then equal to the directive gain times a loss factor to account for these losses [4]. In this chapter we shall assume that the dissipative loss is zero and that the impedances are perfectly matched. Therefore, Equation (5.5) describes the *peak antenna gain;* it can be viewed as the result of concentrating the RF flux in some restricted region less than 4π steradians, as shown in Figure 5.4. Now we can define an *effective radiated power,* with respect to an isotropic radiator (EIRP), as the product of the transmitted power P_t and the gain of the transmitting antenna G_t, as follows:

$$\text{EIRP} = P_t G_t \tag{5.6}$$

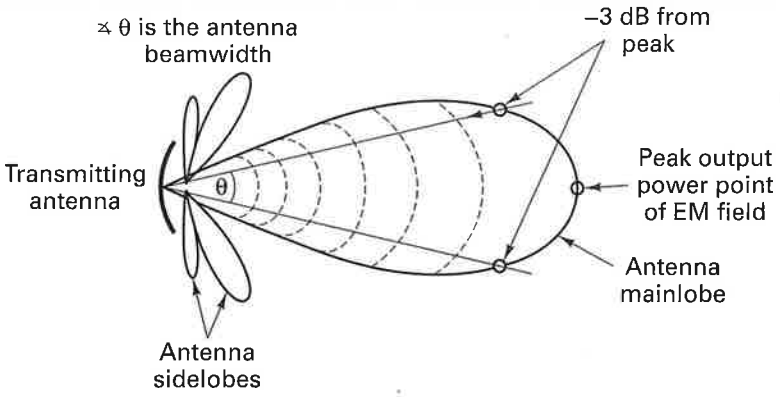


Figure 5.4 Antenna gain is the result of concentrating the isotropic RF flux.

Example 5.1 Effective Isotropic Radiated Power

Show that the same value of EIRP can be produced equally well by using a transmitter with $P_t = 100$ W or with $P_t = 0.1$ W, by employing the appropriate antenna in each case.

Solution

Figure 5.5a depicts a 100-W transmitter coupled to an isotropic antenna; the EIRP = $P_t G_t = 100 \times 1 = 100$ W. Figure 5.5b depicts a 0.1-W transmitter coupled to an antenna with gain $G_t = 1000$; the EIRP = $P_t G_t = 0.1 \times 1000 = 100$ W. If field-strength meters were positioned, as shown, to measure the effective power, the measurements could not distinguish between the two cases.

5.3.1.1 Back to the Range Equation

For the more general case in which the transmitter has some antenna gain relative to an isotropic antenna, we replace P_t with EIRP in Equation (5.2) to yield

$$P_r = \text{EIRP} \frac{A_{er}}{4\pi d^2} \tag{5.7}$$

The relationship between antenna gain G and antenna effective area A_e is [4]

$$G = \frac{4\pi A_e}{\lambda^2} \quad (\text{for } A_e \gg \lambda^2) \tag{5.8}$$

where λ is the wavelength of the carrier. Wavelength λ and frequency f are reciprocally related by $\lambda = c/f$, where c is the speed of light ($\simeq 3 \times 10^8$ m/s). Similar expressions apply for both the transmitting and receiving antennas. The *reciprocity*

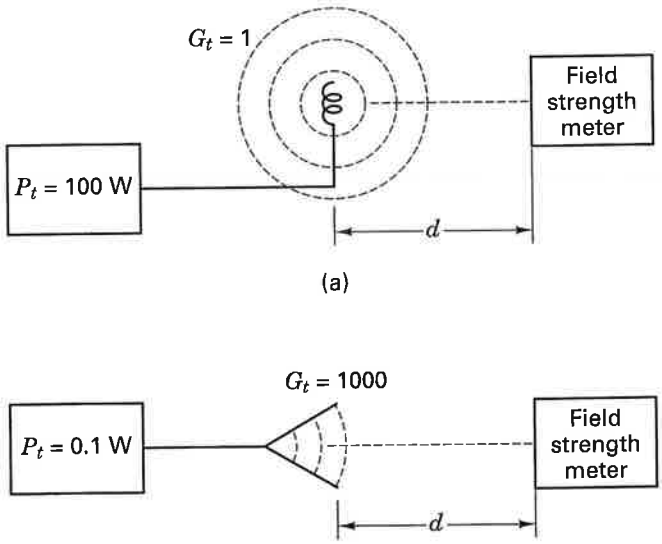


Figure 5.5 The same value of EIRP produced two different ways.

theorem states that for a given antenna and carrier wavelength, the transmitting and receiving gains are identical [4].

The antenna field of view is a measure of the solid angle into which most of the field power is concentrated. Field of view is a measure of the directional properties of the antenna; it is inversely related to antenna gain—high-gain antennas are commensurate with narrow fields of view. Instead of using the solid-angle field of view, we often deal with the planar angle *beamwidth* measured in radians or degrees. Figure 5.4 pictures a directive antenna pattern and illustrates the common definition of the antenna beamwidth. The beamwidth is the angle that subtends the points at which the peak field power is reduced by 3 dB. How does the antenna beamwidth vary with signal frequency? How does the beamwidth vary with antenna size? As can be seen from Equation (5.8), the antenna gain increases with decreased wavelength (increased frequency); antenna gain also increases with increased effective area. Increasing antenna gain is tantamount to focusing the flux density into a more restricted cone angle; hence, increasing either the signal frequency or the antenna size results in a *narrower beamwidth.*

We can calculate the effective area of an isotropic antenna by setting $G = 1$ in Equation (5.8) and solving for A_e as follows:

$$A_e = \frac{\lambda^2}{4\pi} \tag{5.9}$$

Then to find the power received, P_r, when the receiving antenna is isotropic, we substitute Equation (5.9) into Equation (5.7) to get

$$P_r = \frac{\text{EIRP}}{(4\pi d/\lambda)^2} = \frac{\text{EIRP}}{L_s} \tag{5.10}$$

where the collection of terms $(4\pi d/\lambda)^2$, called the *path loss* or *free-space loss,* is designated by L_s. Notice that Equation (5.10) states that the power received by an isotropic antenna is equal to the effective transmitted power, reduced only by the path loss. When the receiving antenna is not isotropic, replacing A_{er} in Equation (5.7) with $G_r\lambda^2/4\pi$ from Equation (5.8) yields the more general expression

$$P_r = \frac{\text{EIRP}\, G_r\lambda^2}{(4\pi d)^2} = \frac{\text{EIRP}\, G_r}{L_s} \tag{5.11}$$

where G_r is the receiving antenna gain. Equation (5.11) can be termed the *range equation.*

5.3.2 Received Signal Power as a Function of Frequency

Since the transmitting antenna and the receiving antenna can each be expressed as a gain or an area, P_r can be expressed four different ways:

$$P_r = \frac{P_t G_t A_{er}}{4\pi d^2} \tag{5.12}$$

$$P_r = \frac{P_t A_{et} A_{er}}{\lambda^2 d^2} \tag{5.13}$$

$$P_r = \frac{P_t A_{et} G_r}{4\pi d^2} \tag{5.14}$$

$$P_r = \frac{P_t G_t G_r \lambda^2}{(4\pi d)^2} \tag{5.15}$$

In these equations, A_{er} and A_{et} are the effective areas of the receiving and transmitting antennas, respectively.

In Equations (5.12) to (5.15) the dependent variable is received signal power, P_r, and the independent variables involve parameters such as transmitted power, antenna gain, antenna area, wavelength, and range. Suppose that we ask the question: How does received power vary as wavelength is decreased (or as frequency is increased), all other independent variables remaining constant? From Equations (5.12) and (5.14) it appears that P_r and wavelength are not related at all. From Equation (5.13), P_r appears to be inversely proportional to wavelength squared, and from Equation (5.15), P_r appears to be directly proportional to wavelength squared. Is there a paradox here? Of course there is not; Equations (5.12) to (5.15) seem to conflict only because antenna gain and antenna area are wavelength related, as stated in Equation (5.8). When should one use each of the Equations (5.12) to (5.15) for determining P_r as a function of wavelength? Consider a system that is already configured; that is, the antennas have already been built or their dimensions are fixed (A_{et} and A_{er} are fixed). Then Equation (5.13) is the appropriate choice for calculating the P_r performance. Equation (5.13) states that for fixed-size antennas, the received power increases as the wavelength is decreased.

Consider the use of Equation (5.12), where G_t and A_{er} are independent variables. We want G_t and A_{er} held fixed over the range of P_r versus wavelength calculations. What happens to the gain of a fixed-dimension transmitting antenna as the independent variable λ is decreased? G_t increases [see Equation 5.8)]. But we cannot have G_t increasing in Equation (5.12)—we want G_t held fixed. In other words, to ensure that G_t remains fixed, we would need to reduce the transmitting antenna size as wavelength decreases. It should be apparent that Equation (5.12) is the appropriate equation when starting with a *fixed transmitting antenna gain* (or beamwidth) requirement and the parameter A_{et} is not fixed. For similar reasons, Equation (5.14) is used when A_{et} and G_r are fixed, and Equation (5.15) is used when both the transmitting and receiving antenna gains (or beamwidths) are fixed.

Figure 5.6 illustrates a satellite application where the downlink antenna beam is required to provide earth coverage (a beamwidth of approximately 17° from synchronous altitude). Since the satellite antenna gain G_t must be fixed, the resulting P_r is independent of wavelength, as shown in Equation (5.12). If the transmission at some frequency f_1 ($= c/\lambda_1$) provides earth coverage, then a frequency change to f_2, where $f_2 > f_1$, will result in reduced coverage (since for a given antenna, G_t will increase); hence the antenna size must be reduced to maintain the required earth

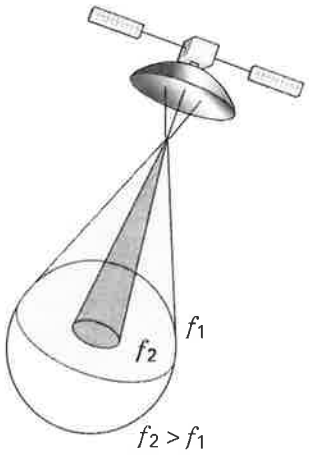


Figure 5.6 Received power as a function of frequency.

coverage or beamwidth. Thus earth coverage antennas become smaller as the carrier frequency is increased.

5.3.3 Path Loss Is Frequency Dependent

From Equation (5.10) it can be seen that path loss, L_s, is wavelength (frequency) dependent. The question is often asked: Why should path loss, which is just a geometric inverse-square loss, be a function of frequency? The answer is that path loss, as characterized in Equation (5.10), is a *definition* predicated on the use of an isotropic receiving antenna ($G_r = 1$). Hence path loss is a convenient tool; it represents a hypothetical received-power loss that *would occur if the receiving antenna were isotropic.* Figure 5.3 and Equation (5.1) have established that power density, $p(d)$, is a function of distance—a purely geometric consideration; $p(d)$ is *not* a function of frequency. However, since path loss is predicated on $G_r = 1$, when we attempt to collect some P_r with an *isotropic antenna,* the result is characterized by Equation (5.10). Again let us emphasize that L_s can be viewed as a convenient collection of terms that have been assigned the unfortunate name *path loss.* The name conjures up an image of a purely geometric effect and fails to emphasize the requirement that $G_r = 1$. A better choice of a name would have been *unity-gain propagation loss.* In a radio communication system, path loss accounts for the largest loss in signal power. In satellite systems, the path loss for a C-band (6-GHz) link to a synchronous altitude satellite is typically 200 dB.

Example 5.2 Antenna Design for Measuring Path Loss

Design a hypothetical experiment to measure path loss L_s, at frequencies $f_1 = 30$ MHz and $f_2 = 60$ MHz, when the distance between the transmitter and receiver is 100 km. Find the effective area of the receiving antenna, and calculate the path loss in decibels for each case.

Solution

Figure 5.7 illustrates the two links for measuring L_s at frequencies f_1 and f_2, respectively. The power density, $p(d)$, at each receiver is identical and equal to

$$p(d) = \frac{\text{EIRP}}{4\pi d^2}$$

This reduction in power density is due *only* to the inverse-square law. The actual power received at each receiver is found by multiplying the power density $p(d)$ at the receiver by the effective area, A_{er}, of the collecting antenna, as shown in Equation (5.7). Since path loss is predicated on $G_r = 1$, we compute the effective area A_{er1} at frequency f_1, and A_{er2} at frequency f_2, using Equation (5.9):

$$A_{er} = \frac{\lambda^2}{4\pi} = \frac{(c/f)^2}{4\pi}$$

$$A_{er1} = \frac{(3 \times 10^8/30 \times 10^6)^2}{4\pi} \approx 8 \text{ m}^2$$

$$A_{er2} = \frac{(3 \times 10^8/60 \times 10^6)^2}{4\pi} \approx 2 \text{ m}^2$$

The path loss for each case in decibels is

$$L_{s1} = 10 \times \log_{10}\left(\frac{4\pi d}{\lambda_1}\right)^2 = 10 \times \log_{10}\left(\frac{4\pi \times 10^5}{3 \times 10^8/30 \times 10^6}\right)^2$$
$$= 102 \text{ dB}$$

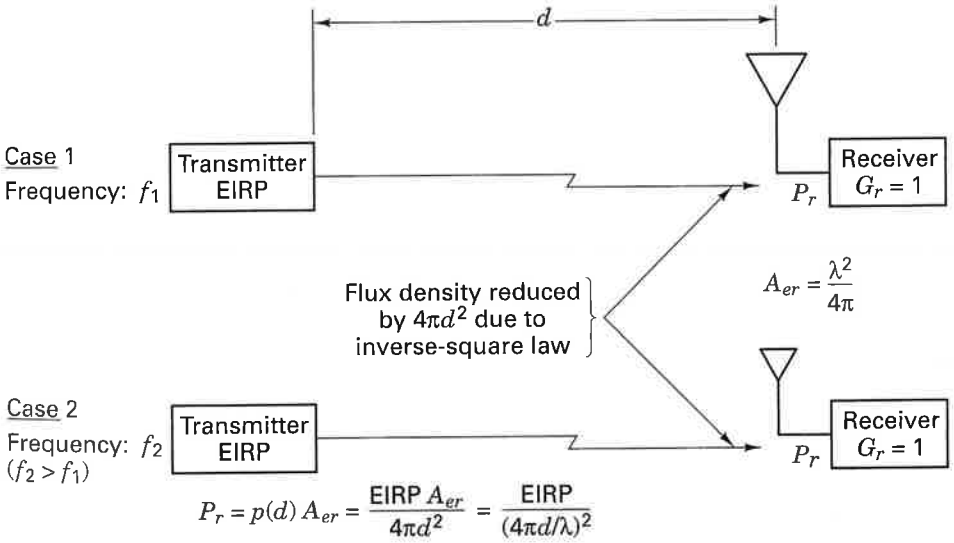


Figure 5.7 Path loss versus frequency. Hypothetical experiment to measure path loss at two different frequencies.

$$L_{s2} = 10 \times \log_{10}\left(\frac{4\pi d}{\lambda_2}\right)^2 = 10 \times \log_{10}\left(\frac{4\pi \times 10^5}{3 \times 10^8/60 \times 10^6}\right)^2$$
$$= 108 \text{ dB}$$

5.3.4 Thermal Noise Power

Thermal noise is caused by the thermal motion of electrons in all conductors. It is generated in the lossy coupling between an antenna and receiver and in the first stages of the receiver. The noise power spectral density is constant at all frequencies up to about 10^{12} Hz, giving rise to the name *white noise.* The thermal noise process in communication receivers is modeled as an additive white Gaussian noise (AWGN) process, as described in Section 1.5.5. The physical model [5, 6] for thermal or Johnson noise is a noise generator with an open-circuit mean-square voltage of $4\kappa T^\circ W\mathcal{R}$, where

κ = Boltzmann's constant = 1.38×10^{-23} J/K or W/K-Hz
= -228.6 dBW/K-Hz,

T° = temperature, kelvin,

W = bandwidth, hertz,

and

$\mathcal{R}$ = resistance, ohms.

The maximum thermal noise power N that could be coupled from the noise generator into the front end of an amplifier is

$$N = \kappa T^\circ W \quad \text{watts} \tag{5.16}$$

Thus, the maximum single-sided noise power spectral density N_0 (noise power in a 1-Hz bandwidth), available at the amplifier input is

$$N_0 = \frac{N}{W} = \kappa T^\circ \quad \text{watts/hertz} \tag{5.17}$$

It might seem that the noise power should depend on the magnitude of the resistance—but it does not. Consider an intuitive argument to verify this. Electrically connect a large resistance to a small one, such that they form a closed path and such that their physical temperatures are the same. If noise power were a function of resistance, there would be a net power flow from the large resistance to the small one; the large resistance would become cooler and the small one would become warmer. This violates our experience, not to mention the second law of thermodynamics. Therefore, the power delivered from the large resistance to the small one must be equal to the power it receives.

The available power from a thermal noise source is dependent on the ambient temperature of the source (the *noise temperature*), as is seen in Equation (5.16). This leads to the useful concept of an *effective noise temperature* for noise sources

that are not necessarily thermal in origin (e.g., galactic, atmospheric, interfering signals) that can be introduced into the receiving antenna. The effective noise temperature of such a noise source is defined as the temperature of a hypothetical thermal noise source that would give rise to an equivalent amount of interfering power. The subject of noise temperature is treated in greater detail in Section 5.5.

Example 5.3 Maximum Available Noise Power

Using a noise generator with mean-square voltage equal to $4\kappa T^\circ W\mathcal{R}$, demonstrate that the maximum amount of noise power that can be coupled from this source into an amplifier is $N_i = \kappa T^\circ W$.

Solution

A theorem from network theory states that maximum power is delivered to a load when the value of the load impedance is made equal to the complex conjugate of the generator impedance [7]. In this case the generator impedance is a pure resistance, $\mathcal{R}$; therefore, the condition for maximum power transfer is fulfilled when the input resistance of the amplifier equals $\mathcal{R}$. Figure 5.8 illustrates such a network. The input thermal noise source is represented by an electrically equivalent model consisting of a noiseless source resistor in series with an ideal voltage generator whose rms noise voltage is $\sqrt{4\kappa T^\circ W\mathcal{R}}$. The input resistance of the amplifier is made equal to $\mathcal{R}$. The noise voltage delivered to the amplifier input is just one-half the generator voltage, following basic circuit principles. The noise power delivered to the amplifier input can accordingly be expressed as

$$N_i = \frac{(\sqrt{4\kappa T^\circ W\mathcal{R}}/2)^2}{\mathcal{R}} = \frac{4\kappa T^\circ W\mathcal{R}}{4\mathcal{R}}$$
$$= \kappa T^\circ W$$

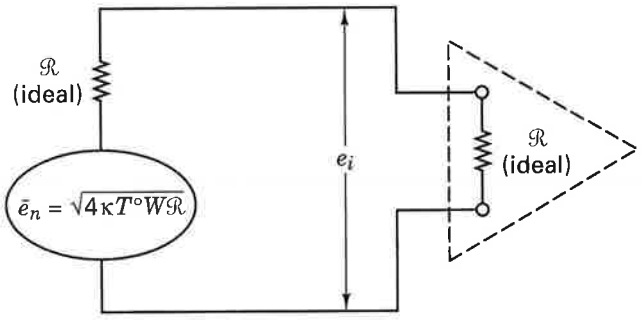


Figure 5.8 Electrical model of maximum available thermal noise power at amplifier input.

5.4 LINK BUDGET ANALYSIS

In link budget calculations, the quantity of greatest interest is the receiving-system signal-to-noise-ratio (SNR), often modeled at the output of the receiving antenna. This SNR is sometimes called *carrier-to-noise* ratio (C/N), where $N = kT_s^\circ W$, k is Boltzmann's constant, T_s° is system temperature in Kelvin (see Section 5.5.5), and W is bandwidth. C/N is often used in many satellite link-budget calculations. The

custom probably arose because satellite signals are usually suppressed-carrier signals, wherein the carrier can be viewed as being totally modulated (transformed into inforamtion-carrying sidelobes). The information-bearing SNR at the predetection point, usually denoted P_r/N or S/N, is the parameter of interest for obtaining E_b/N_0. For suppressed-carrier signaling, P_r/N and C/N have the same value; thus the following expressions are sometimes used interchangeably.

$$\frac{P_r}{N} \equiv \frac{S}{N} \equiv \frac{C}{N} \equiv \frac{C}{\kappa T^\circ W}$$

Is P_r/N always the same as C/N? No, the signal power and the carrier power are only the same when the carrier is totally modulated (e.g., wideband angle modulation). For example, consider a frequency-modulated (FM) carrier wave, as

$$s(t) = A \cos(\omega_0 t + K \int m(t)\, dt)$$

where $m(t)$ is a modulating signal, and K is a system constant. The average power in $m(t)$ is $\overline{m^2(t)}$. Increasing this modulating power only serves to increase the frequency deviation of $s(t)$, which means that the carrier is spread over a wider spectrum, but its average power $\overline{s^2(t)}$ remains equal to $A^2/2$, regardless of the power in the modulating signal. With such wideband angle-modulated signals, the carrier is "carrying" the information over an increased bandwidth, explaining why the information-bearing signal power is sometimes referred to as the carrier power.

For linear modulation, such as amplitude modulation (AM), the power in the carrier is quite different than the power in the modulating signal. For example, consider an AM carrier wave in terms of the modulating signal $m(t)$:

$$s(t) = [1 + m(t)]\ A \cos \omega_0 t$$

$$\overline{s^2(t)} = [1 + m(t)]^2 \frac{A^2}{2}$$

$$= \frac{A^2}{2}[1 + \overline{m^2(t)} + \overline{2\, m(t)}]$$

If we assume that $m(t)$ has a zero mean, then the average carrier power can be written as

$$\overline{s^2(t)} = \frac{A^2}{2} + \frac{A^2}{2}\,\overline{m^2(t)}$$

In this expression, there is an unmodulated carrier component; hence, the carrier power is not the same as the signal power. In summary, the parameters C/N and P_r/N are the same for suppressed-carrier signals (e.g., PSK or FSK), but this is not the case for a signal having an unmodulated carrier component represented by a spectral line at the carrier frequency (e.g., amplitude modulation).

We obtain P_r/N by dividing Equation (5.11) by noise power N, as follows:

$$\frac{P_r}{N} = \frac{\text{EIRP}\ G_r/N}{L_s} \tag{5.18}$$

Equation (5.18) applies to any one-way RF link. With *analog receivers,* the noise bandwidth (generally referred to as the effective or equivalent noise bandwidth) seen by the demodulator is usually greater than the signal bandwidth, and P_r/N is the main parameter for measuring signal detectability and performance quality. With *digital receivers,* however, correlators or matched filters are usually implemented, and signal bandwidth is taken to be equal to noise bandwidth. Rather than consider input noise power, a common formulation for digital links is to replace noise power with *noise power spectral density.* We can use Equation (5.17) with T° taken to be system effective temperature (developed later) to rewrite Equation (5.18) as

$$\frac{P_r}{N_0} = \frac{\text{EIRP}\ G_r/T^\circ}{\kappa L_s L_o} \tag{5.19}$$

where the system effective temperature T° is a function of the noise radiated into the antenna and the thermal noise generated within the receiver. Note that the receiving antenna gain G_r and system temperature T° are grouped together. The grouping G_r/T° is sometimes called the *receiver figure-of-merit.* The reason for treating these terms in this way is explained in Section 5.6.2.

It is important to emphasize that the system effective temperature T° is a parameter that *models* all the noise in the receiving system; the subject is treated in greater detail in Section 5.5. In Equation (5.19) we have introduced a term L_o to represent all other losses and degradation factors not specifically addressed by the other terms of Equation (5.18). The factor L_o allows for the large assortment of different losses and noise sources cataloged earlier. Equation (5.19) summarizes the key parameters of any link analysis: the received signal power-to-noise power spectral density (P_r/N_0), the effective transmitted power (EIRP), the receiver figure-of-merit (G_r/T°), and the losses (L_s, L_0). We are developing a methodical way to keep track of the gains and losses in a communications link. By starting with some source of power we can compute, using Equation (5.19), the net SNR arriving at the "face" of the detector (predetection point). We are working toward a "bookkeeping" system, much like the type used in a business that earmarks assets and liabilities and tallies up the bottom-line of profit (or loss). Equation (5.19) takes on such an entrepreneurial appearance. All of the numerator parameters (effective radiated power, receiver figure-of-merit) are like the assets of a business, and all the denominator parameters (thermal noise, space loss, other losses) are like the liabilities of a business.

Assuming that all the received power P_r is in the modulating (information-bearing) signal, we now relate E_b/N_0 and SNR from Equation (3.30) and write

$$\frac{E_b}{N_0} = \frac{P_r}{N}\left(\frac{W}{R}\right) \tag{5.20a}$$

$$\frac{E_b}{N_0} = \frac{P_r}{N_0}\left(\frac{1}{R}\right) \tag{5.20b}$$

and

$$\frac{P_r}{N_0} = \frac{E_b}{N_0} R \tag{5.20c}$$

where R is the bit rate. If some of the received power is carrier power (a signal-power loss), we can still employ Equation (5.20), except that the carrier power contributes to the loss factor, L_o in Equation (5.19). This fundamental relationship between E_b/N_0 and P_r/N_0 in Equation (5.20) will be required frequently in designing and evaluating systems. (See Chapter 9.)

5.4.1 Two E_b/N_0 Values of Interest

We have referred to E_b/N_0 as that value of bit energy per noise power spectral density required to yield a specified error probability. To facilitate calculating a margin or safety factor M, we need to differentiate between the *required* E_b/N_0 and the actual or *received* E_b/N_0. From this point on we will refer to the former as $(E_b/N_0)_{\text{reqd}}$ and to the latter as $(E_b/N_0)_r$. Figure 5.9 depicts an example with two operating points. The first is associated with $P_B = 10^{-3}$; let us call this operating point the system required error performance. Let us assume that an $(E_b/N_0)_{\text{reqd}}$ value of 10 dB will yield this required performance. Do you suppose we would build this system so that the demodulator received this 10-dB value *exactly?* Of course not; we would specify and design the system to have a safety margin, so that the $(E_b/N_0)_r$ actually received would be somewhat larger than the $(E_b/N_0)_{\text{reqd}}$. Thus we might design the system to operate at the second operating point on Figure 5.9; here $(E_b/N_0)_r = 12$ dB and $P_B = 10^{-5}$. For this example we can describe the safety margin or *link margin,* as providing a two-order-of-magnitude improved P_B, or as is more usual, we can describe the link margin in terms of providing 2 dB more E_b/N_0

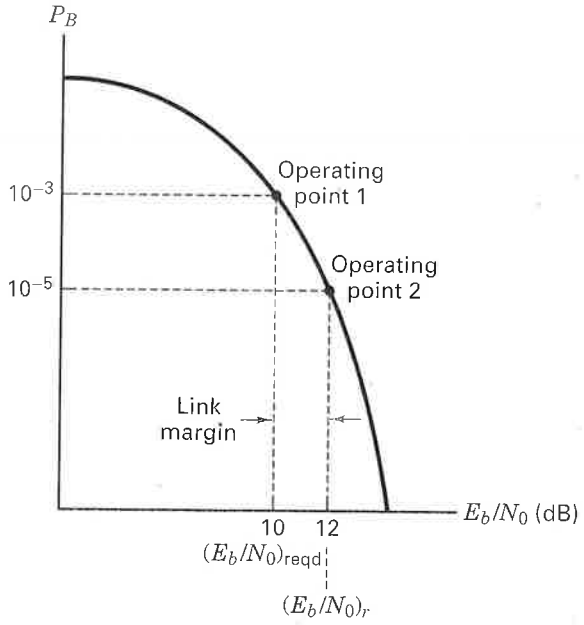


Figure 5.9 Two E_b/N_0 values of interest.

than is required. We can rewrite Equation (5.20c), introducing the link margin parameter M as

$$\frac{P_r}{N_0} = \left(\frac{E_b}{N_0}\right)_r \quad R = M\left(\frac{E_b}{N_0}\right)_{\text{reqd}} R \tag{5.21}$$

The difference in decibels between $(E_b/N_0)_r$ and $(E_b/N_0)_{\text{reqd}}$ yields the link margin:

$$M(\text{dB}) = \left(\frac{E_b}{N_0}\right)_r (\text{dB}) - \left(\frac{E_b}{N_0}\right)_{\text{reqd}} (\text{dB}) \tag{5.22}$$

The parameter $(E_b/N_0)_{\text{reqd}}$ reflects the differences from one system design to another; these might be due to differences in modulation or coding schemes. A larger than expected $(E_b/N_0)_{\text{reqd}}$ may be due to a suboptimal RF system, which manifests large timing errors or which allows more noise into the detection process than does an ideal matched filter.

Combining Equations (5.19) and (5.21) and solving for the link margin M yields

$$M = \frac{\text{EIRP}\, G_r/T^\circ}{(E_b/N_0)_{\text{reqd}}\, R\kappa L_s L_o} \tag{5.23}$$

Equation (5.23), the *link margin equation,* contains all of the parameters contributing to the link's error performance. Some of these parameters are defined with reference to particular locations. For example, E_b/N_0 is defined at the input to the receiver. More precisely, it is defined at the input to the detector (predetection point), where a demodulated waveform having a voltage amplitude proportional to received energy is the basis of a symbol decision. Similarly, any parameter describing received energy or power, whether useful or degrading, is also defined with reference to this predetection point. The receiver's figure-of-merit G_r/T° is defined at the input to the receiving antenna, where G_r is the gain of the receiving antenna, and T° is the effective system temperature. (See Section 5.5.5.) The effective radiated power EIRP is defined as the power associated with the electromagnetic wave at the output of the transmitting antenna. Each of these parameters, E_b/N_0, G_r/T°, and EIRP is defined with reference to a particular system location, and nowhere else.

5.4.2 Link Budgets Are Typically Calculated in Decibels

Since link budget analysis is typically calculated in decibels, we can express Equation (5.23) as

$$M(\text{dB}) = \text{EIRP}\,(\text{dBW}) + G_r(\text{dBi}) - \left(\frac{E_b}{N_0}\right)_{\text{reqd}} (\text{dB}) - R(\text{dB-bit/s}) - \kappa T^\circ\,(\text{dBW/Hz}) - L_s\,(\text{dB}) - L_o(\text{dB}) \tag{5.24}$$

Transmitted signal power EIRP is expressed in decibel-watts (dBW); noise power spectral density N_0 is in decibel-watts per hertz (dBW/Hz); antenna gain G_r is in decibels referenced to isotropic gain (dBi); data rate R is in decibels referenced to

1 bit/s (dB-bit/s); and all other terms are in decibels (dB). The numerical values of the Equation (5.24) parameters constitute the link budget, a useful tool for allocating communications resources. In an effort to maintain a positive margin, we might trade off any parameter with any other parameter; we might choose to reduce transmitter power by giving up excess margin, or we might elect to increase the data rate by reducing $(E_b/N_0)_{reqd}$ (through the selection of improved modulation and coding). Any one of the Equation (5.24) decibels, regardless of the parameter from which it stems, is just as good as any other decibel—a dB is a dB is a dB. The transmission system "does not know and does not care" where the decibels come from. As long as the needed amount of E_b/N_0 arrives at the receiver, the desired system error performance can be met. Well, let us add two other conditions for achieving a required error performance—synchronization must be maintained, and ISI distortion must be minimized or equalized. One might ask, Since the system has no preference as to where the decibels of E_b/N_0 come from, how shall we prioritorize the search for an adequate number of decibels? The answer is, we should look for the most cost-effective decibels. This goal will steer us toward reading the next several chapters on error-correction coding, since this discipline has historically provided a continual reduction in the cost of electronics to achieve error-performance improvements.

5.4.3 How Much Link Margin Is Enough?

The question of how much link margin should be designed into a system is asked frequently. The answer is that if all sources of gain, loss, and noise have been rigorously detailed (worst case), and if the link parameters with large variances (e.g., fades due to weather) match the statistical requirements for link availability, very little additional margin is needed. The margin needed depends on how much confidence one has in each of the link budget entries. For systems employing new technology or new operating frequencies, one needs more margin than for systems that have been repeatedly built and tested. Sometimes the link budget provides an allowance for fades due to weather directly, as a line item. Other times, however, the required value of margin reflects the link requirements for a given rain degradation. For satellite communications at C-band (uplink at 6 GHz, downlink at 4 GHz), where the parameters are well known and fairly well behaved, it should be possible to design a system with only 1 dB of link margin. Receive-only television stations operating with 16-ft-diameter dishes at C-band are frequently designed with only a fraction of a decibel of margin. However, telephone communications via satellite using standards of 99.9% availability require considerably more margin; some of the INTELSAT systems have 4 to 5 dB of margin. When nominal rather than worst-case computations are performed, allowances are usually made for unit-to-unit equipment variations over the operating temperature range, line voltage variations, and mission duration. Also, for space communications, there may be an allowance for errors in tracking a satellite's location.

Designs using higher frequencies (e.g., 14/12 GHz) generally call for larger (weather) margins because atmospheric losses increase with frequency and are highly variable. It should be noted that a by-product of the attenuation due to atmospheric loss is greater antenna noise. With low-noise amplifiers, small weather

changes can result in increases of 40 to 50 K in antenna temperature. Table 5.1 represents a link analysis proposed to the Federal Communications Commission (FCC) by Satellite Television Corporation for the Direct Broadcast Satellite (DBS) service. Notice that the downlink budget is tabulated for two alternative weather conditions: clear, and 5-dB loss due to rain. The signal loss due to atmospheric attenuation is only a small fraction of a decibel for clear weather and is the full 5 dB during rain. The next item in the downlink tabulation, home receiver $G/T°$, illustrates the additional degradation caused by the rain; additional thermal noise irradiates the receiving antenna, making the effective system noise temperature, $T°$, increase, and the home receiver $G/T°$ decrease (from 9.4 dB/K to 8.1 dB/K). Therefore, when extra margin is allowed for weather loss, additional margin should simultaneously be added to compensate for the increase in system noise temperature.

With regard to satellite links, in industry one often hears such expressions as "the link *can* be closed," meaning that the margin, in decibels, has a positive value and the required error performance will be satisfied, or "the link *cannot* be closed," meaning that the margin has a negative value and the required error performance will *not* be satisfied. Even though the words "the link closes" or "the link does not close" give the impression of an "on-off" condition, it is worth emphasizing that lack of link closure, or a negative margin, means that the error performance falls short of the system requirement; it does not necessarily mean that communications cease. For example, consider a system whose $(E_b/N_0)_{reqd} = 10$ dB, as shown in Figure 5.9, but whose $(E_b/N_0)_r = 8$ dB. Assume that 8 dB corresponds to $P_B = 10^{-2}$.

TABLE 5.1 Proposed Direct Broadcast Satellite (DBS) from Satellite Television Corp.

Uplink	
Earth station EIRP	86.6 dBW
Free-space loss (17.6 GHz, 48° elevation)	208.9 dB
Assumed rain attenuation	12.0 dB
Satellite $G/T°$	7.7 dB/K
Uplink $C/\kappa T°$	102.0 dB-Hz

	Atmospheric Condition	
Downlink	Clear	5-dB Rain Attenuation
Satellite EIRP	57.0 dBW	57.0 dBW
Free-space loss (12.5 GHz, 30° elevation)	206.1 dB	206.1 dB
Atmospheric attenuation	0.14 dB	5.0 dB
Home receiver $G/T°$ (0.75 m dish)	9.4 dB/K	8.1 dB/K
Receiver pointing loss (0.5° error)	0.6 dB	0.6 dB
Polarization mismatch loss (average)	0.04 dB	0.04 dB
Downlink $C/\kappa T°$	88.1 dB-Hz	82.0 dB-Hz
Overall $C/\kappa T°$	87.9 dB-Hz	82.0 dB-Hz
Overall C/N (in 16 MHz)	15.9 dB	10.0 dB
Reference threshold C/N	10.0 dB	10.0 dB
Margin over threshold	5.9 dB	0.0 dB

Thus there is a margin of –2 dB, and a bit error probability of 10 times the specified error probability. The link may still be useful, though degraded.

5.4.4 Link Availability

Link availability is usually a measure of long-term link utility stated on an average annual basis; for a given geographical location, the link availability measures the percentage of time the link can be closed. For example, for a particular link between Washington, D.C., and a satellite repeater, the long-term weather pattern may be such that a 10-dB weather margin is adequate for link closure 98% of the time; for 2% of the time, heavy rains result in greater than 10 dB SNR degradation, so that the link does not close. Since the effect of rain on SNR degradation is a function of signal frequency, link availability and required margin must be examined in the context of a particular transmission frequency.

Figure 5.10 summarizes worldwide satellite link availability at a frequency of 44 GHz. The plot illustrates percentage of the earth visible (the link closes, and a prescribed probability of error is met) as a function of margin for the case of three equispaced geostationary satellites. A *geostationary satellite* is located in a circular orbit in the same plane as the earth's equatorial plane and at the synchronous altitude of 35,800 km. The satellite's orbital period is identical with that of the earth's rotational period, and therefore the satellite appears stationary when viewed from the earth. Figure 5.10 shows a family of visibility curves with different required link availabilities, ranging from benign (95% availability) to fairly stringent (99% availability). In general, for a fixed link margin, visibility is inversely proportional to required availability, and for a fixed availability, visibility increases monotonically with margin [8]. Figures 5.11 to 5.13 illustrate, by

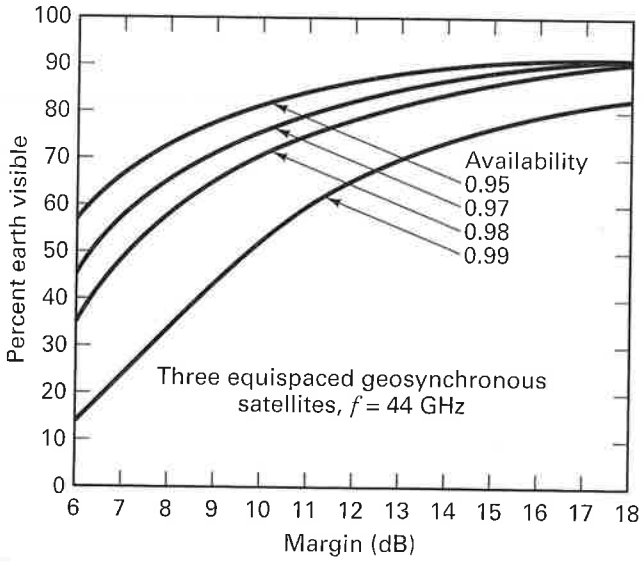


Figure 5.10 Earth coverage versus link margin for various values of link availability. (Reprinted from L. M. Schwab, "World-Wide Link Availability for Geostationary and Critically Inclined Orbits Including Rain Effects," *Lincoln Laboratory, Rep. DCA-9,* Jan. 27, 1981, Fig. 14, p. 38, courtesy of Lincoln Laboratory.)

Figure 5.11 Earth coverage (unshaded) for 0.99 link availability for three equispaced geostationary satellites, $f = 44$ GHz, link margin = 14 dB. (Reprinted from L. M. Schwab, "World-Wide Link Availability for Geostationary and Critically Inclined Orbits Including Rain Effects," *Lincoln Laboratory, Rep. DCA-9*, Jan. 27, 1981, Fig. 17, p. 42, courtesy of Lincoln Laboratory.)

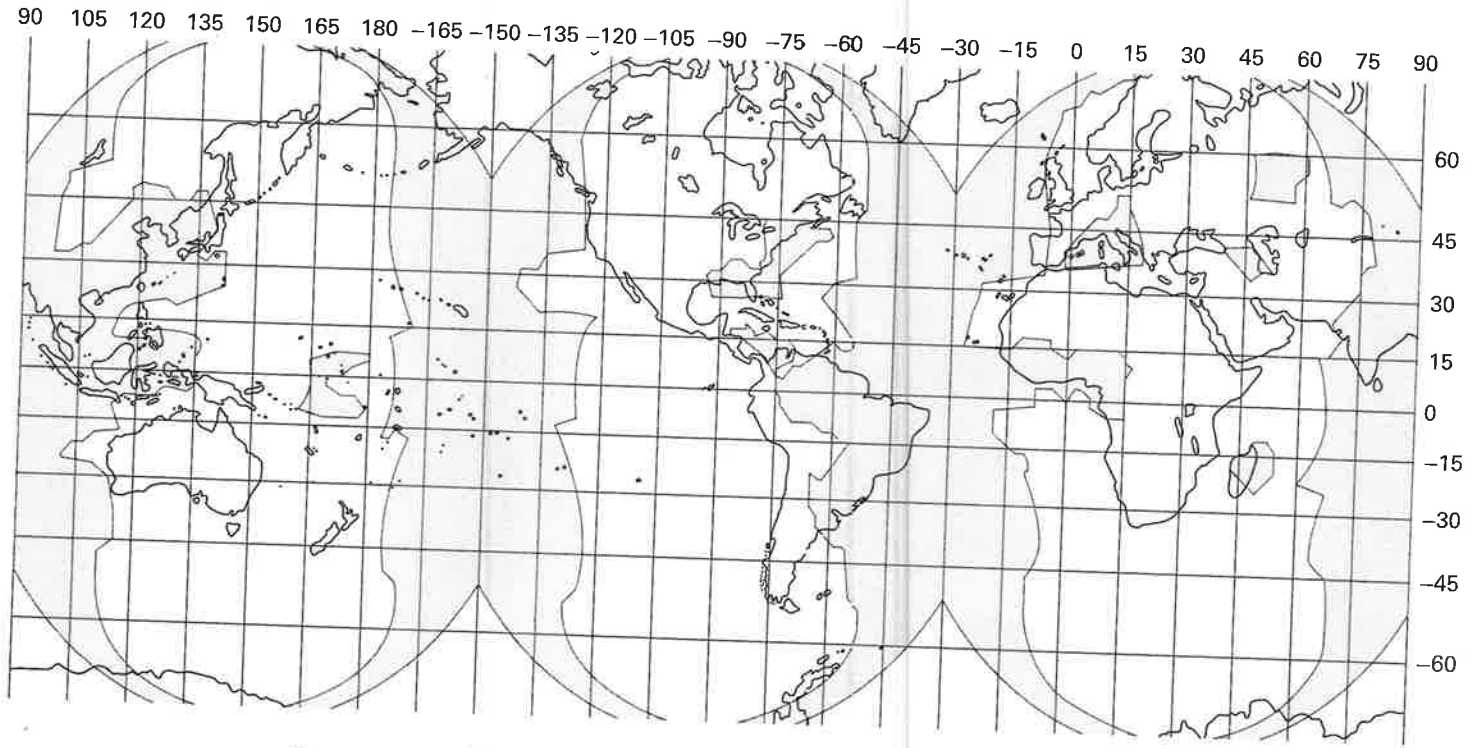


Figure 5.12 Earth coverage (unshaded) for 0.99 link availability for three equispaced geostationary satellites, $f = 44$ GHz, link margin = 10 dB. (Reprinted from L. M. Schwab, "World-Wide Link Availability for Geostationary and Critically Inclined Orbits Including Rain Effects," *Lincoln Laboratory, Rep. DCA-9,* Jan. 27, 1981, Fig. 18, p. 43, courtesy of Lincoln Laboratory.)

L. M. Schwab, "World-Wide Link Availability for Geostationary and Critically Inclined Orbits Including Rain Effects," *Lincoln Laboratory, Rep. DCA-9,* Jan. 27, 1981, Fig. 18, p. 43, courtesy of Lincoln Laboratory.)

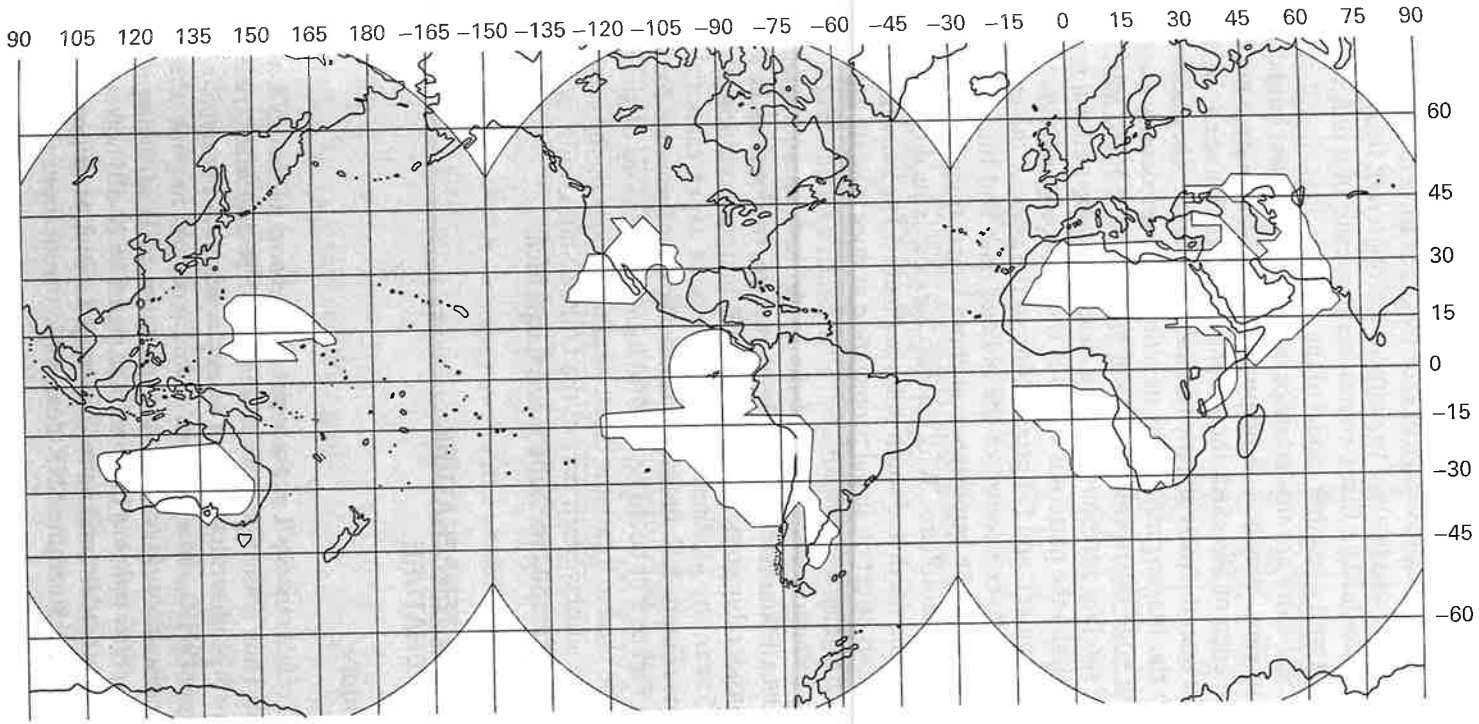


Figure 5.13 Earth coverage (unshaded) for 0.99 link availability for three equispaced geostationary satellites, $f = 44$ GHz, link margin = 6 dB. (Reprinted from L. M. Schwab, "World-Wide Link Availability for Geostationary and Critically Inclined Orbits Including Rain Effects," *Lincoln Laboratory, Rep. DCA-9,* Jan. 27, 1981, Fig. 19, p. 44, courtesy of Lincoln Laboratory.)

unshaded and shaded areas, the parts of the earth from which the 44-GHz link can and cannot be closed 99% of the time for three different values of link margin. Figure 5.11 illustrates the link coverage of such locations for a margin value of 14 dB. Notice that this figure can be used to pinpoint the regions of heaviest rainfall, such as Brazil and Indonesia. The figure represents the result of a link calculation performed in concert with a weather model of the earth.

In Figure 5.11 there are shaded strips on the east and west boundaries of each satellite's field of view. Why do you suppose the link availability is not met in these regions? At the edge of the earth the propagation path between the satellite and ground is longer than the path directly beneath the satellite. Degradation occurs in three ways: (1) the longer path results in reduced power density at the receiving antenna; (2) the edge of coverage sites will experience reduced satellite antenna gain, unless the satellite antenna pattern is designed to be uniform over its entire field of view (typically the pattern is –3 dB at the beam edge compared to the peak gain at the beam center); and (3) propagation to the edge of the earth traverses a thicker atmospheric layer because of the oblique path and the earth's curvature. The third item is of prime importance at those signal frequencies that are most attenuated by the atmosphere. Why do you suppose you do not see the same shaded areas near the north and south poles in Figure 5.11? Snowfall does not have the same deleterious effect on signal propagation as does rainfall; the phenomenon is known as the *freeze effect.*

Figure 5.12 illustrates the parts of the earth that can and cannot close the 44-GHz link 99% of the time with 10-dB link margin. Notice that the shaded areas have grown considerably compared to the 14-dB margin case; now, the east coast of the United States, the Mediterranean, and most of Japan cannot close the link 99% of the time. Figure 5.13 illustrates similar link performance for a margin of 6 dB. Whereas Figure 5.11 could be used to locate the regions of greatest rainfall, Figure 5.13 can be used to locate the driest weather regions on the earth; such areas are seen to be the southwestern part of the United States, most of Australia, the coast of Peru and Chile, and the Sahara desert in Africa.

5.5 NOISE FIGURE, NOISE TEMPERATURE, AND SYSTEM TEMPERATURE

5.5.1 Noise Figure

Noise figure, F, relates the SNR at the input of a network to the SNR at the output of the network. Thus noise figure measures the SNR degradation caused by the network. Figure 5.14 illustrates such an example. Figure 5.14a depicts the SNR at an *amplifier input* $(\text{SNR})_{\text{in}}$ as a function of frequency. At its peak, the signal is 40 dB above the noise floor. Figure 5.14b depicts the SNR at the *amplifier output* $(\text{SNR})_{\text{out}}$. The amplifier gain has increased the signal by 20 dB; however, the amplifier has added its own additional noise. The output signal, at its peak, is only 30 dB above the noise floor. Since the SNR degradation from input to output is 10 dB,

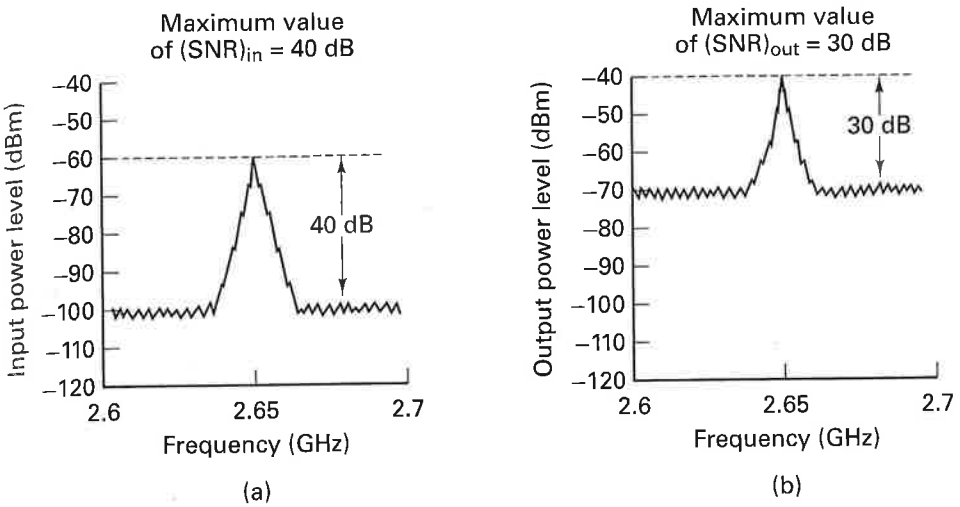


Figure 5.14 Amplifier signal and noise levels as a function of frequency. (a) Amplifier input. (b) Amplifier output.

this is tantamount to describing the amplifier as having a 10-dB noise figure. Noise figure is a parameter that expresses the noisiness of a two-port network or device, such as an amplifier, compared with a reference noise source at the input port. It can be written as

$$F = \frac{(\text{SNR})_{\text{in}}}{(\text{SNR})_{\text{out}}} = \frac{S_i/N_i}{GS_i/G(N_i + N_{ai})} \tag{5.25}$$

where

S_i = signal power at the amplifier input port,

N_i = noise power at the amplifier input port,

N_{ai} = amplifier noise referred to the input port,

and

G = amplifier gain.

Figure 5.15 is an example illustrating Equation (5.25). Figure 5.15a represents a *realizable amplifier* example with a gain $G = 100$, and internal noise power $N_a = 10\ \mu\text{W}$. The source noise, external to the amplifier, is $N_i = 1\ \mu\text{W}$. In Figure 5.15b we assume that the *amplifier is ideal,* and we ascribe the noisiness of the real amplifier, from part (a) of the figure, to an external source N_{ai} in series with the original source N_i. The value of N_{ai} is obtained by reducing N_a by the amplifier gain. As shown in Figure 5.15b, Equation (5.25) references all noise to the amplifier input, whether the noise is actually present at the input or is internal to the device. As can

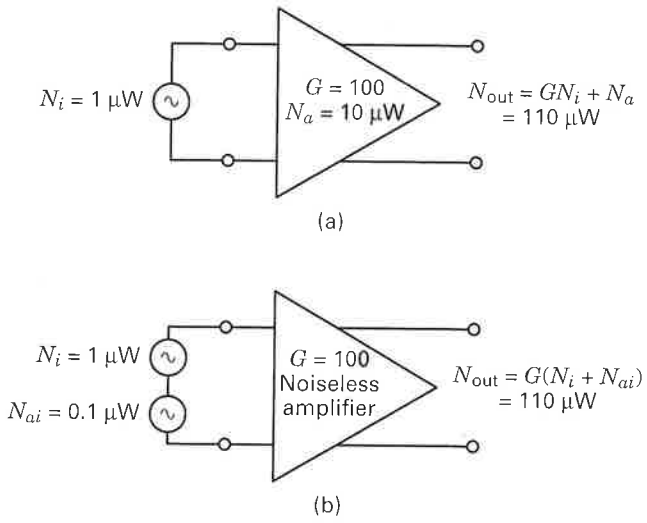


Figure 5.15 Example of noise treatment in amplifiers.

be seen in Figure 5.15, the noise power output from the real amplifier is identical to that of its electrically equivalent model.

Equation (5.25) reduces to

$$F = \frac{N_i + N_{ai}}{N_i} = 1 + \frac{N_{ai}}{N_i} \tag{5.26}$$

Notice from Equation (5.26) that the noise figure expresses the noisiness of a network relative to an input source noise; noise figure is *not* an absolute measure of noise. An ideal amplifier or network, one that contributes no noise ($N_{ai} = 0$), has a noise figure equal to unity (0 dB).

For the concept of noise figure to have utility, we need to be able to make equitable comparisons among devices on the basis of Equation (5.26). We must, therefore, choose a value of N_i as a *reference*. The noise figure of any device will then represent a measure of how much noisier the device is than the reference. In 1944, Friis [9] suggested that noise figure be defined for a noise source at a reference temperature of $T_0^\circ = 290$ K. That suggestion was subsequently adopted by the IEEE as part of its standard definition for noise figure [10]. From Equation (5.17) we see that the maximum available noise power spectral density from any source resistance is established by specifying its temperature. The value of 290 K was selected as the reference because it is a reasonable approximation of the source temperature for many links. Also, with T_0° chosen to be 290 K, the value of noise spectral density N_0 at T_0° results in an aesthetically pleasing number:

$$N_0 = \kappa T_0^\circ = 1.38 \times 10^{-23} \times 290 = 4.00 \times 10^{-21} \text{ W/Hz}$$

Or, expressed in decibels,

$$N_0 = -204 \text{ dBW/Hz}$$

Now that noise figure F has been defined with reference to a 290 K noise source, it is important to emphasize that the noise figure relationships in Equations (5.25) and (5.26) are only accurate when N_i is a 290 K noise source. For those cases where N_i is other than a 290 K noise source, we must rename F in Equations (5.25) and (5.26) to be termed *operational noise figure* F_{op}. The relationship between F_{op} and F is shown later in Equation (5.48).

5.5.2 Noise Temperature

Rearranging Equation (5.26), we can write

$$N_{ai} = (F - 1)N_i \tag{5.27}$$

From Equation (5.16) we can replace N_i with $\kappa T_0^\circ W$ and N_{ai} with $\kappa T_R^\circ W$, where T_0° is the reference temperature of the source and T_R° is called the *effective noise temperature* of the receiver (or network). We can then write

$$\kappa T_R^\circ W = (F - 1)\kappa T_0^\circ W$$

or

$$T_R^\circ = (F - 1)\, T_0^\circ$$

Or, since T_0° has been chosen to be 290 K,

$$T_R^\circ = (F - 1)290 \text{ K} \tag{5.28}$$

Equation (5.26) uses the concept of noise figure to characterize the noisiness of an amplifier. Equation (5.28) represents an alternative but equivalent characterization known as *effective noise temperature*. Note that the noise figure is a measurement relative to a reference. However, noise temperature has no such constraint.

We can think of available noise power spectral density and effective noise temperature, in the context of Equation (5.17), as equivalent ways of characterizing noise sources. Equation (5.28) tells us that the noisiness of an amplifier can be modeled as if it were caused by an additional noise source, as seen in Figure 5.15b, operating at some effective temperature called T_R°. For a purely resistive termination, T_R° is never less than ambient temperature unless it is cooled. It is important to note that for reactive terminations, such as uncooled parametric amplifiers or other low-noise devices, T_R° can be much less than 290 K, even though the ambient temperature is higher [11]. For the output of an amplifier as a function of its effective temperature, we can use Equations (5.16), (5.25), and (5.28) to write

$$N_{\text{out}} = GN_i + GN_{ai} \tag{5.29a}$$

$$= G\kappa T_g^\circ W + G\kappa T_R^\circ W = G\kappa(T_g^\circ + T_R^\circ)\, W \tag{5.29b}$$

$$= G\kappa T_g^\circ W + (F - 1)\, G\kappa T_0^\circ\, W \tag{5.29c}$$

where T_g° is the temperature of the source, and T_0° is 290 K.

5.5.3 Line Loss

The difference between amplifier networks and lossy line networks can be viewed in the context of the degradation mechanisms *loss* and *noise,* described earlier. Noisy networks in Sections 5.5.1 and 5.5.2 were discussed with amplifiers in mind. We saw that SNR degradation resulted from injecting additional (amplifier) noise into the link, as shown in Figure 5.15. However, in the case of a lossy line, we shall show that the SNR degradation results from the signal being attenuated while the noise remains fixed (for the case where the line temperature is equal to or less than the source temperature). The degradation effect will nonetheless be measured as an increase in noise figure or effective noise temperature.

Consider the lossy line or network shown in Figure 5.16. Assume the line is matched with its characteristic impedance at the source and at the load. We shall define power loss as

$$L = \frac{\text{input power}}{\text{output power}}$$

Then, the network gain G equals $1/L$ (less than unity for a lossy line). Let all components be at temperature T_g°. The total output noise power flowing from the network into the load is

$$N_{\text{out}} = \kappa T_g^\circ W$$

since the network output appears as a pure resistance at the temperature T_g°. The total power flowing from the load back into the network must also equal N_{out} to ensure thermal equilibrium. Recall that available noise power $\kappa T^\circ W$ is dependent only on temperature, bandwidth, and impedance matching; it is not dependent on the resistance value. N_{out} can be considered to be made up of two components, N_{go} and GN_{Li}, such that

$$N_{\text{out}} = \kappa T_g^\circ W = N_{go} + GN_{Li} \tag{5.30}$$

where

$$N_{go} = G\kappa T_g^\circ W \tag{5.31}$$

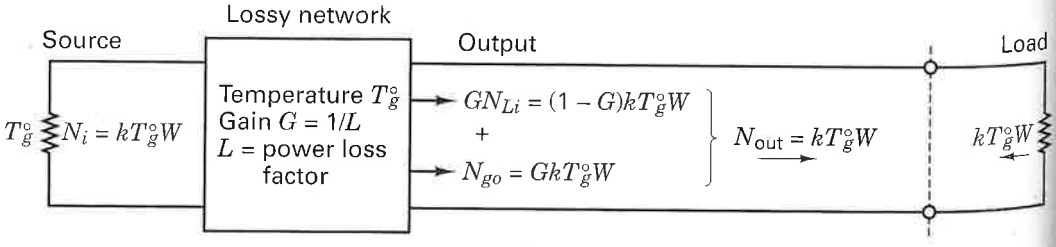


Figure 5.16 Lossy line: impedance matched and temperature matched at both ends.

is the component of output noise power due to the source and GN_{Li} is the component of output noise power due to the lossy network, where N_{Li} is the network noise relative to its input. Combining Equations (5.30) and (5.31), we can write

$$\kappa T_g^\circ W = G\kappa T_g^\circ W + GN_{Li} \tag{5.32}$$

then, solving for N_{Li} yields

$$N_{Li} = \frac{1-G}{G}\kappa T_g^\circ W = \kappa T_L^\circ W \tag{5.33}$$

Therefore, the effective noise temperature of the line is

$$T_L^\circ = \frac{1-G}{G}T_g^\circ \tag{5.34}$$

and since $G = 1/L$,

$$T_L^\circ = (L-1)T_g^\circ \tag{5.35}$$

Choosing $T_g^\circ = 290$ K as the reference temperature, we can write

$$T_L^\circ = (L-1)290\text{ K} \tag{5.36}$$

Using Equations (5.28) and (5.36), the *noise figure for a lossy line* can be expressed as

$$F = 1 + \frac{T_L^\circ}{290} = L \tag{5.37}$$

When the network is a lossy line, such that $F = L$ and $G = 1/L$, then N_{out} in Equation (5.29c) takes the following form:

$$N_{\text{out}} = \frac{\kappa T_g^\circ W}{L} + \left(1 - \frac{1}{L}\right)\kappa T_0^\circ W \tag{5.38}$$

Note that some authors use the parameter L to mean the reciprocal of the loss factor defined here. In such cases, noise figure $F = 1/L$.

Example 5.4 Lossy Line

A line at temperature $T_0^\circ = 290$ K is fed from a source whose noise temperature is $T_g^\circ = 1450$ K. The input signal power S_i is 100 picowatts (pW) and the signal bandwidth W is 1 GHz. The line has a loss factor $L = 2$. Calculate the $(\text{SNR})_{\text{in}}$, the effective line temperature T_L°, the output signal power S_{out}, and the $(\text{SNR})_{\text{out}}$.

Solution

$$\begin{aligned} N_i &= \kappa T_g^\circ W \\ &= 1.38 \times 10^{-23}\text{ W/K-Hz} \times 1450\text{ K} \times 10^9\text{ Hz} \\ &= 2 \times 10^{-11}\text{ W} = 20\text{ pW} \end{aligned}$$

$$(\text{SNR})_{\text{in}} = \frac{100\text{ pW}}{20\text{ pW}} = 5\ (7\text{ dB})$$

$$T_L^\circ = (L - 1)\,290\ K = 290\ K$$

$$S_{\text{out}} = \frac{S_i}{L} = \frac{100\ \text{pW}}{2} = 50\ \text{pW}$$

Using Equation (5.39), we obtain

$$N_{\text{out}} = \frac{\kappa T_g^\circ W}{L} + \left(1 - \frac{1}{L}\right)\kappa T_0^\circ W$$
$$= \frac{2 \times 10^{-11}}{2}\ \text{W} + \frac{1}{2}(4 \times 10^{-12})\ \text{W} = 12\ \text{pW}$$

and

$$(\text{SNR})_{\text{out}} = \frac{50\ \text{pW}}{12\ \text{pW}} = 4.17\ (6.2\ \text{dB})$$

5.5.4 Composite Noise Figure and Composite Noise Temperature

When two networks are connected in series, as shown in Figure 5.17a, their composite noise figure can be written as

$$F_{\text{comp}} = F_1 + \frac{F_2 - 1}{G_1} \tag{5.39}$$

where G_1 is the gain associated with network 1. When n networks are connected in series the relationship between stages expressed in Equation (5.39) continues, so that the *composite noise figure* for a sequence of n stages is written as

$$F_{\text{comp}} = F_1 + \frac{F_2 - 1}{G_1} + \frac{F_3 - 1}{G_1 G_2} + \cdots + \frac{F_n - 1}{G_1 G_2 \cdots G_{n-1}} \tag{5.40}$$

Can you guess from Equation (5.40) what the design goals for the front end of the receiver (especially the first stage or the first couple of stages) should be? At the front end of the receiver, the signal is most susceptible to added noise; therefore, the first stage should have as low a noise figure F_1 as possible. Also, because the noise figure of each subsequent stage is reduced by the gains of the prior stages, it behooves us to strive for as high a gain G_1 as possible. Simultaneously achieving the lowest F_1 and the highest G_1 represents conflicting goals; therefore, compromises are always necessary.

Equations (5.40) and (5.28) can be combined to express the composite effective noise temperature of a sequence of n stages:

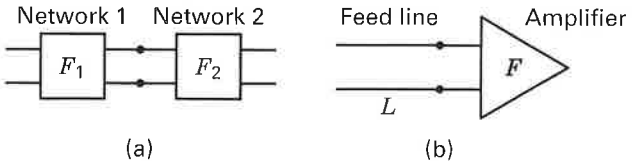


Figure 5.17 Networks connected in series.

$$T^\circ_{\text{comp}} = T^\circ_1 + \frac{T^\circ_2}{G_1} + \frac{T^\circ_3}{G_1 G_2} + \cdots + \frac{T^\circ_n}{G_1 G_2 \cdots G_{n-1}} \tag{5.41}$$

Figure 5.17b illustrates a feed line in series with an amplifier; this is a typical arrangement following a receiving antenna. Using Equation (5.39) to find F_{comp} for such a lossy line and amplifier arrangement, we can write

$$F_{\text{comp}} = L + L(F - 1) = LF \tag{5.42}$$

since the noise figure of the lossy line is L and the gain of the line is $1/L$. By analogy with Equation (5.36), we can write the composite temperature as

$$T^\circ_{\text{comp}} = (LF - 1)290 \text{ K} \tag{5.43}$$

We can also write the composite temperature of line and amplifier as follows:

$$\begin{aligned} T^\circ_{\text{comp}} &= (LF - 1 + L - L)290 \text{ K} \\ &= [(L - 1) + L(F - 1)]290 \text{ K} \\ &= T^\circ_L + LT^\circ_R \end{aligned} \tag{5.44}$$

5.5.4.1 Comparison of Noise Figure and Noise Temperature

Since noise figure F and effective noise temperature T° characterize the noise performance of devices, some engineers feel compelled to select one of these measures as the more useful. However, they each have their place. For terrestrial applications, F is almost universally used; the concept of SNR degradation for a 290 K source temperature makes sense, because terrestrial source temperatures are typically close to 290 K. Terrestrial noise figure values typically fall in the convenient range 1 to 10 dB.

For space applications, T° is the more common figure of merit. The range of values for commercial systems is typically between 30 and 150 K, giving adequate resolution for comparing performance between systems. A disadvantage of using noise figures for such low-noise networks is that the values obtained are all close to unity (0.5 to 1.5 dB), which makes it difficult to compare devices. For low-noise applications, F (in decibels) would need to be expressed to two decimal places to provide the same resolution or precision as does T°. For space applications, a reference temperature of 290 K is not as appropriate as it is for terrestrial applications. When using effective temperature, no reference temperature (other than absolute zero K) is needed for judging degradation. The effective input noise temperature is simply compared to the effective source noise temperature. In general, applications involving very low noise devices seem to favor the effective temperature measure over the noise figure.

5.5.5 System Effective Temperature

Figure 5.18 represents a simplified schematic of a receiving system, identifying those areas—the antenna, the line, and the preamplifier—that play a primary role in SNR degradation. We have already discussed the degradation role of the pream-

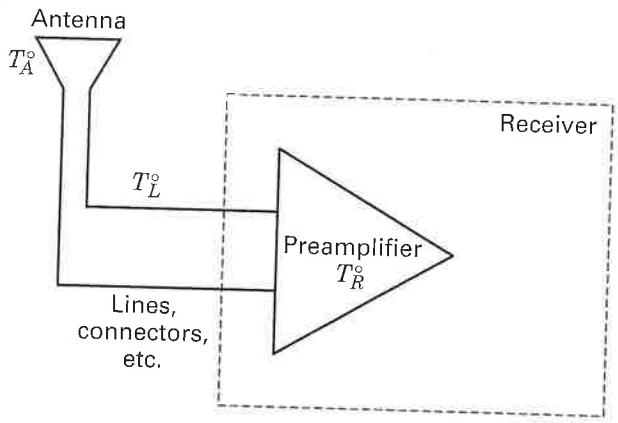


Figure 5.18 Major noise contributors of a receiving system.

plifier—additional noise is injected into the link. And we have discussed line loss—the signal is attenuated, while the noise is held fixed (for the case where the line temperature is less than or equal to the source temperature). The remaining source of degradation stems from both natural and man-made sources of noise and interference that enter the receiving antenna. Natural sources include lightning, celestial radio sources, atmospheric sources, and thermal radiation from the ground and other physical structures. Man-made noise includes radiation from automobile ignition systems and any other electrical machinery, and radio transmissions from other users that fall within the receiver bandwidth. The total noise contributed by these external sources can be characterized by $kT_{\text{ant}}W$, where T_{ant} is known as the *antenna temperature*. An antenna is like a lens. Its noise contributions are dictated by what the antenna is "looking at." If an antenna is pointed at a cool portion of the sky, very little thermal noise is introduced. The antenna temperature is a measure of the effective temperature integrated over the entire antenna pattern.

We now find *system temperature* T_S° by adding together all the system noise contributors (in terms of effective temperature). The summation is expressed as

$$T_S^\circ = T_A^\circ + T_{\text{comp}}^\circ \tag{5.45}$$

where T_A° is the antenna temperature and T_{comp}° is the composite temperature of the line and the preamplifier. Equation (5.45) illustrates the two primary sources of noise and interference degradation at a receiver. One source, characterized by T_A°, represents degradation from the "outside world" arriving via the antenna. The second source, characterized by T_{comp}°, is thermal noise caused by the motion of electrons in all conductors. Since the system temperature T_S° is a new composite, made up of T_A° and the composite effective temperature of the line and preamp, one might ask: Why doesn't Equation (5.45) appear to have the same sequential gain reduction factors as those in Equation (5.41)? We have assumed that the antenna has *no dissipative parts;* its gain, unlike an amplifier or attenuator, can be thought of as a processing gain. Whatever effective temperature is introduced at the antenna comes through, unaltered by the antenna; the antenna represents the source noise, or source temperature, at the input to the line.

Using Equation (5.44), we can modify Equation (5.45) as follows:

$$T_S^\circ = T_A^\circ + T_L^\circ + LT_R^\circ \tag{5.46}$$

$$= T_A^\circ + (L-1)290\text{ K} + L(F-1)290\text{ K}$$

$$= T_A^\circ + (LF-1)290\text{ K} \tag{5.47}$$

If LF is provided in units of decibels, we must first convert LF to a ratio, so that T_S° takes the form

$$T_S^\circ = T_A^\circ + (10^{LF/10} - 1)290\text{ K}$$

Equations (5.10), (5.11), and (5.45) through (5.47) describe power received P_r and system temperature T_S respectively. In each case, the parameters are referred to the output of the receiving antenna, a popular reference, preferred by system and antenna designers, and those working at the transmitter side of the link. Another convention, often used by receiver designers, describes received power P_r' and system temperature T_S' referenced to the input of the receiver. Assuming that the antenna and receiver are connected via a lossy line, then P_r and P_r' (also T_S and T_S') are related by the line-loss factor L. That is, $P_r = LP_r'$, and $T_S = LT_S'$. Note that the ratio of received power to system temperature, a receiving-system SNR parameter, is the same for both refereences. That is so because, $P_r/T_S = LP_r'/LT_S'$. One can similarly choose a reference anywhere in the receiver circuit-string (from the output of the receiving antenna to the predetection point) that results in the same receiving-system SNR model.

Example 5.5 Noise Figure and Noise Temperature

A receiver front end, shown in Figure 5.19a, has a noise figure of 10 dB, a gain of 80 dB, and a bandwidth of 6 MHz. The input signal power, S_i, is 10^{-11} W. Assume that the line is lossless and the antenna temperature is 150 K. Find T_R°, T_S°, N_{out}, $(SNR)_{in}$, and $(SNR)_{out}$.

Solution

First convert all decibel values to ratios:

$$T_R^\circ = (F-1)290\text{ K} = 2610\text{ K}$$

Using Equation (5.46) with $L = 1$ for a lossless line yields

$$T_S^\circ = T_A^\circ + T_R^\circ = 150\text{ K} + 2610\text{ K} = 2760\text{ K}$$

$$N_{out} = G\kappa T_A^\circ W + G\kappa T_R^\circ W = G\kappa T_S^\circ W$$

$$= 10^8 \times 1.38 \times 10^{-23} \times 6 \times 10^6(150\text{ K} + 2610\text{ K})$$

$$= \underbrace{1.2\ \mu\text{W}}_{\text{source contribution}} + \underbrace{21.6\ \mu\text{W}}_{\text{front-end contribution}} = 22.8\ \mu\text{W}$$

$$(SNR)_{in} = \frac{S_i}{\kappa T_A^\circ W} = \frac{10^{-11}}{1.24 \times 10^{-14}} = 806.5\ (29.1\text{ dB})$$

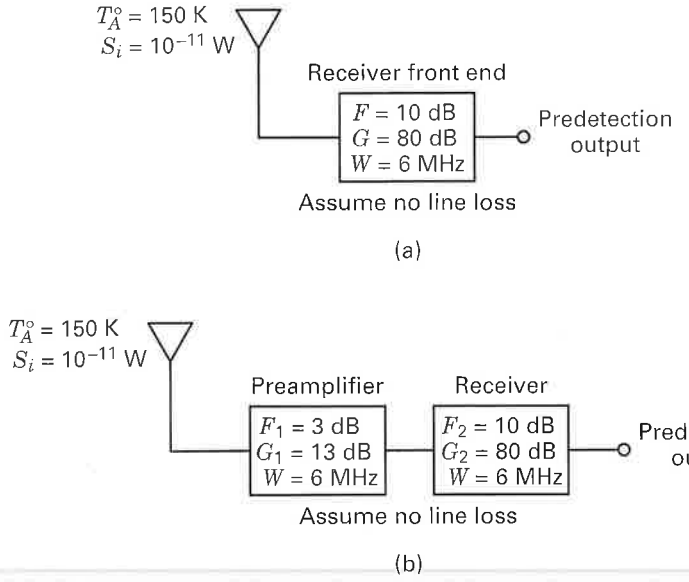


Figure 5.19 Improving a receiver front end with a low-noise preamplifier.

$$(\text{SNR})_{\text{out}} = \frac{S_{\text{out}}}{N_{\text{out}}} = \frac{10^8 \times 10^{-11}}{22.8 \times 10^{-6}} = 43.9\ (16.4\ \text{dB})$$

Notice in this example that the amplifier noise is significantly larger than the source noise and represents the major cause of SNR degradation.

Example 5.6 Improving SNR with a Low-Noise Preamplifier

Use a preamplifier, as shown in Figure 5.19b, with a noise figure of 3 dB, a gain of 13 dB, and a bandwidth of 6 MHz to improve the SNR of the receiver in Example 5.5. Find T°_{comp} for the composite preamplifier and receiver. Find T°_S, F_{comp}, N_{out}, and $(\text{SNR})_{\text{out}}$. Assume zero line loss.

Solution

Again, convert all decibel values to ratios before proceeding:

$$T^\circ_{R1} = (F_1 - 1)290\ \text{K} = 290\ \text{K}$$

$$T^\circ_{R2} = (F_2 - 1)290\ \text{K} = 2610\ \text{K}$$

$$T^\circ_{\text{comp}} = T^\circ_{R1} + \frac{T^\circ_{R2}}{G_1} = 290\ \text{K} + \frac{2610\ \text{K}}{20} = 420.5\ \text{K}$$

$$T^\circ_S = T^\circ_A + T^\circ_{\text{comp}} = 150\ \text{K} + 420.5\ \text{K} = 570.5\ \text{K}$$

$$F_{\text{comp}} = F_1 + \frac{F_2 - 1}{G_1} = 2 + \frac{9}{20} = 2.5\ (4\ \text{dB})$$

$$N_{out} = G\kappa T_A^\circ W + G\kappa T_{comp}^\circ W = G\kappa T_S^\circ W$$
$$= 20 \times 10^8 \times 1.38 \times 10^{-23} \times 6 \times 10^6(150\text{ K} + 420.5\text{ K})$$
$$= \underbrace{24.8\ \mu\text{W}}_{\substack{\text{source}\\\text{contribution}}} + \underbrace{69.6\ \mu\text{W}}_{\substack{\text{front-end}\\\text{contribution}}} = 94.4\ \mu\text{W}$$

$$(\text{SNR})_{out} = \frac{S_{out}}{N_{out}} = \frac{10^{-11} \times 20 \times 10^8}{94.4 \times 10^{-6}} = 212.0\ (23.3\text{ dB})$$

With the added preamplifier the (predetection) output noise has increased (from 22.8 μW in Example 5.5) to 94.4 μW. Even though the noise power has increased, the lower system temperature has resulted in a 6.9-dB improvement in SNR (from 16.4 dB in Example 5.5, to 23.3 dB here). The price we pay for this improvement is the need to provide an F_{comp} improvement of 6 dB (from 10 dB in Example 5.5, to 4 dB in this example).

The unwanted noise is, in part, *injected via the antenna* ($\kappa T_A^\circ W$), and in part, *generated internally* in the receiver front end ($\kappa T_{comp}^\circ W$). The amount of system improvement that can be rendered via front-end design depends on what portion of the total noise the front end contributes. We saw in Example 5.5 that the front end contributed the major portion of the noise. Therefore, in Example 5.6, providing a low-noise preamplifier improved the system SNR significantly. In the next example, we show the case where the major portion of the noise is injected via the antenna; we shall see that introducing a low-noise preamplifier in such a case will not help the SNR appreciably.

Example 5.7 Attempting SNR Improvement When the Value of T_A° Is Large

Repeat Examples 5.5 and 5.6 with one change: let $T_A^\circ = 8000$ K. In other words, the preponderant amount of noise is being injected via the antenna; the antenna might have a very hot body (the sun) fully occupying its field of view. Calculate the SNR improvement that would be provided by the preamplifier used in Example 5.6 and Figure 5.19b, and compare the result with that of Example 5.6.

Solution

Without preamplifier

$$N_{out} = G\kappa W(T_A^\circ + T_R^\circ)$$
$$= 10^8 \times 1.38 \times 10^{-23} \times 6 \times 10^6(8000\text{ K} + 2610\text{ K})$$
$$= \underbrace{66.2\ \mu\text{W}}_{\substack{\text{source}\\\text{contribution}}} + \underbrace{21.6\ \mu\text{W}}_{\substack{\text{front-end}\\\text{contribution}}} = 87.8\ \mu\text{W}$$

$$(\text{SNR})_{out} = \frac{S_{out}}{N_{out}} = \frac{10^8 \times 10^{-11}}{87.8 \times 10^{-6}} = 11.4\ (10.6\text{ dB})$$

With preamplifier

$$N_{out} = 20 \times 10^8 \times 1.38 \times 10^{-23} \times 6 \times 10^6(8000\text{ K} = 420.5\text{ K})$$

$$= \underbrace{1324.8\ \mu W}_{\substack{\text{source}\\ \text{contribution}}} + \underbrace{69.6\ \mu W}_{\substack{\text{front-end}\\ \text{contribution}}} = 1394.4\ \mu W$$

$$(\text{SNR})_{\text{out}} = \frac{20 \times 10^8 \times 10^{-11}}{1.39 \times 10^{-3}} = 14.4\ (11.6\ \text{dB})$$

Therefore, for this case, the SNR improvement is only 1 dB, a far cry from the 6.9 dB accomplished earlier. When the noise is mostly due to devices within the receiver, it is possible to improve the SNR by introducing low-noise devices. However, when the noise is mostly due to external causes, improving the receiver front end will not help much.

Noise figure is a definition, predicated on a reference temperature of 290 K. When the source temperature is other than 290 K, as is the case in Examples 5.5, 5.6, and 5.7, it is necessary to define an *operational* or *effective noise figure* that describes the actual $(\text{SNR})_{\text{in}}$ versus $(\text{SNR})_{\text{out}}$ relationship. Starting with Equations (5.25) and (5.27), the operational noise figure can be found as follows:

$$\begin{aligned} F_{\text{op}} &= \frac{S_i/kT_AW}{GS_i/G(kT_AW + N_{ai})} \\ &= \frac{kT_AW + N_{ai}}{kT_AW} = 1 + \frac{(F-1)kT_0W}{kT_AW} \\ &= 1 + \frac{T_0}{T_A}(F-1) \end{aligned} \tag{5.48}$$

5.5.6 Sky Noise Temperature

The receiving antenna collects random noise emissions from galactic, solar, and terrestrial sources, constituting the sky background noise. The sky background appears as a combination of galactic effects that decrease with frequency, and atmospheric effects that start becoming significant at 10 GHz and increase with frequency. Figure 5.20 illustrates the sky temperature, as measured from the earth, due to both these effects. Notice that there is a region, between 1 and 10 GHz, where the temperature is lowest; the galaxy noise has become quite small at 1 GHz, and for satellite communications the blackbody radiation noise due to the absorbing atmosphere is not significant below 10 GHz. (For other applications, e.g., passive radiometry, it is still a problem.) This region, known as the *microwave window* or *space window,* is particularly useful for satellite or deep-space communication. The low sky noise is the principal reason that such systems primarily use carrier frequencies in this part of the spectrum. The galaxy and atmospheric noise curves in Figure 5.20 are made up of a family of curves each at a different elevation angle θ. When θ is zero, the receiving antenna points at the horizon and the propagation path encompasses the longest possible atmospheric layer; when θ is 90°, the receiving antenna points to the zenith, and the resulting propagation path contains the shortest possible atmospheric layer. Thus the upper curve of the family repre-

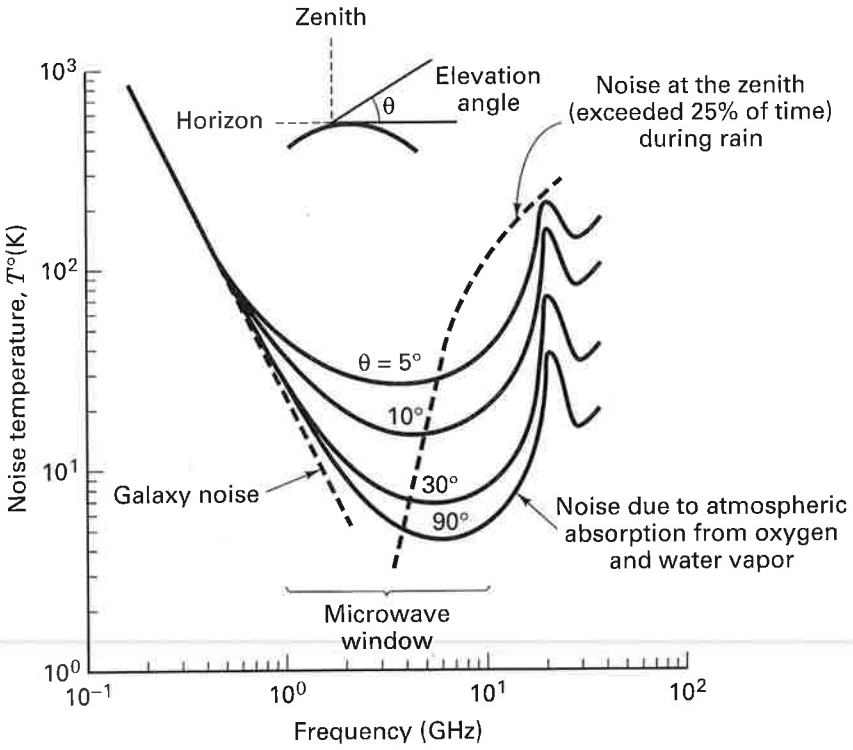


Figure 5.20 Sky noise temperature.

sents the near worst-case (clear-weather) sky noise temperature versus frequency, and the lower curve represents the most benign case. Also shown in Figure 5.20 is a plot of noise temperature versus frequency *due to rain.* Since the intensity of any rainstorm can only be expressed statistically, the noise temperatures shown are values that are exceeded 25% of the time (at the zenith). Which spectral region appears the most benign for space communications when rainfall is taken into account? It is the region at the low end of the space window. For this reason, systems such as the Space Ground Link Subsystem or SGLS (military) and the Unified S-Band Telemetry, Tracking, and Control System (NASA) are located in the 1.8 to 2.4-GHz band.

5.5.6.1 Radio Maps of the Sky

Various researchers have mapped the galactic noise radiation as a function of frequency. Figure 5.21 is such a radio temperature map, after Ko and Kraus [12], indicating the temperature contours of the sky in the region of 250 MHz when viewed from the earth. In general, the sky is composed of localized galactic sources (sun, moon, planets, etc.), each having its own temperature. The map is effectively

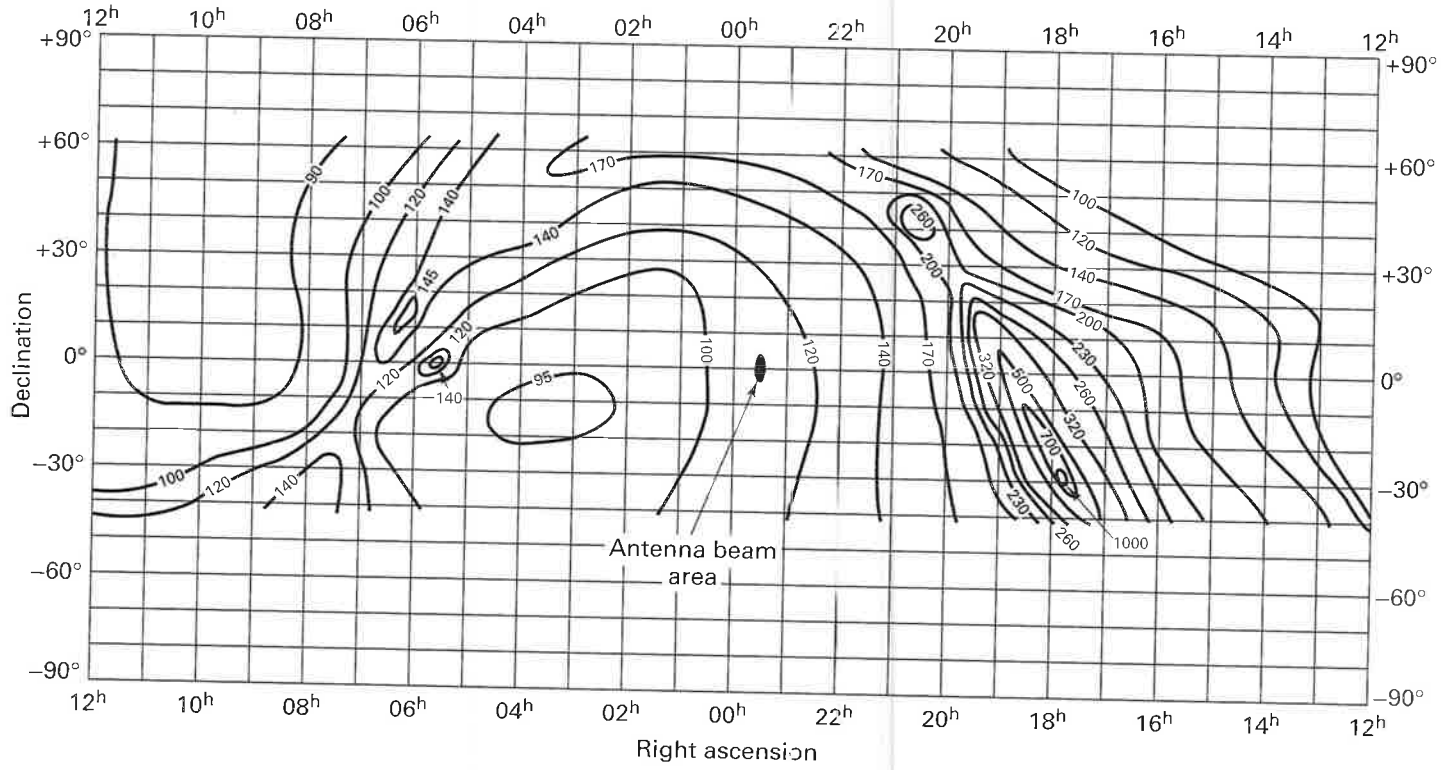


Figure 5.21 Radio map of the sky background at 250 MHz. (Reprinted from H. C. Ko and J. D. Kraus, "A Radio Map of the Sky at 1.2 Meters," *Sky Telesc.*, vol. 16, Feb. 1957, p. 160, with permission from *Sky and Telescope* astronomy magazine, Cambridge, Mass.)

Ko and J. D. Kraus, "A Radio Map of the Sky at 1.2 Meters," *Sky Telesc.,* vol. 16, Feb. 1957, p. 160, with permission from *Sky and Telescope* astronomy magazine, Cambridge, Mass.)

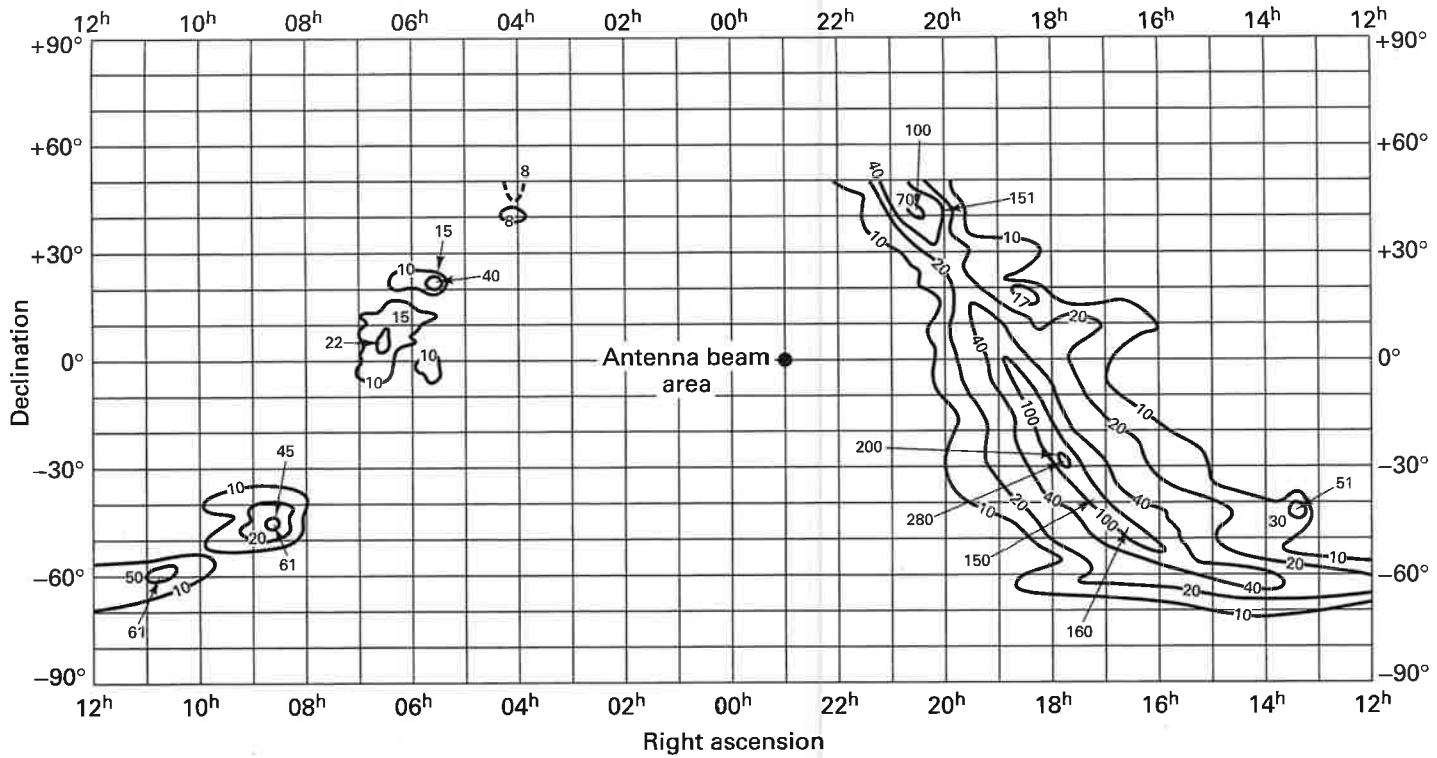


Figure 5.22 Radio map of the sky background at 600 MHz. (Reprinted with permission from J. H. Piddington and G. H. Trent, "A Survey of Cosmic Radio Emission at 600 Mc/s," *Aust. J. Phys.,* vol. 9, Dec. 1956, Fig. 1, pp. 483–486.)

a weighted sum of the individual galactic source temperatures plus a constant sky background. The coordinates of the map, *declination and right ascension,* can be thought of as celestial latitude and longitude with an earth reference (right ascension is calibrated in units of hour-angle, where 24 hours corresponds to a complete rotation of the earth). On Figure 5.21, the temperature contours range from a low of 90 K to a high of 1000K. The measurements were made so as to exclude the sun (night sky). The antenna beam in the center of the map indicates the size of the sky area over which measurements are made (each measurement is an average over that beam area). The narrower the beam, the finer the resolution of the temperature contours; the wider the beam, the coarser the resolution.

Figure 5.22 is another such radio map at 600 MHz, after Piddington and Trent [13]. At this frequency the galaxy noise is reduced compared to Figure 5.21, as predicted in Figure 5.20; the low is 8 K and the high is 280 K. If you examine Figures 5.21 and 5.22 for the region of greatest noise radiation, where on the map do you see the most activity, and what is its significance? It is seen as an elongated region in the right-hand midsection of each map; the longitudinal axis of the elongation designates the location of our *galactic plane,* where such cosmic noise radiation is most intense.

5.6 SAMPLE LINK ANALYSIS

In Section 5.4 we developed the basic link parameter relationships. In this section we use these relationships to calculate a sample link budget, as shown in Table 5.2. The table may appear to house a formidable listing of terms; one can get the false impression that the link budget represents a complex compilation. Just the opposite is true, and we introduce Figure 5.23 to underscore this assertion. In this figure we have reduced the set of line items from Table 5.2 to a few key parameters. The goal of a link analysis is to determine whether or not the required error performance is met, by examining the E_b/N_0 actually received and comparing it with the E_b/N_0 required to meet the system specification. The principal items needed for this determination are the EIRP (how much effective power is transmitted), the G/T° figure of merit (how much capability the receiver has for collecting this power), L_s (the largest single loss, the space loss), and L_o (other contributing losses and degradations). That is *all* there is to it!

5.6.1 Link Budget Details

The link budget example in Table 5.2 consists of three columns of numbers. Only the middle column represents the link budget. The other columns consist of ancillary information, such as antenna beamwidth, or computations to support the main tabulation. Losses are bracketed in the usual bookkeeping way. If a value is not bracketed, it represents a gain. Subtotals are shown enclosed in a box. Starting from the top of the middle column, we algebraically sum all of the gains and losses. The final link margin result is shown in a double box in item 21. The computations

TABLE 5.2 Earth Terminal to Satellite Link Budget Example: Frequency = 8 GHz, Range = 21,915 Nautical Miles.

1.	Transmitter power (dBW)	(100.00W)	20.0	P_t
2.	Transmitter circuit loss (dB)		⟨2.0⟩	L_o
3.	Tramsmitter antenna gain (peak dBi)		51.6	G_t
	Dish diameter (ft)	20.00		
	Half-power beamwidth (degrees)	0.45		
4.	Terminal EIRP (dBW)		69.6	EIRP
5.	Path loss (dB)	(10° elev.)	⟨202.7⟩	L_s
6.	Fade allowance (dB)		⟨4.0⟩	L_o
7.	Other losses (dB)		⟨6.0⟩	L_o
8.	Received isotropic power (dBW)		−143.1	
9.	Receiver antenna gain (peak dBi)		35.1	G_r
	Dish diameter (ft)	3.00		
	Half-power beamwidth (degrees)	2.99		
10.	Edge-of-coverage loss (dB)		⟨2.0⟩	L_o
11.	Received signal power (dBW)		−110.0	P_r
	Receiver noise figure at antenna port (dB)			11.5
	Receiver temperature (dB-K)			35.8 (3806 K)
	Receiver antenna temperature (dB-K)			24.8 (300 K)
12.	System temperature (dB-K)			36.1 (4106 K)
13.	System G/T° (dB/K)	−1.0		G/T°
14.	Boltzmann's constant (dBW/K-Hz)			−228.60
15.	Noise spectral density (dBW/Hz)		⟨−192.5⟩	$N_0 = kT^\circ$
16.	Received P_r/N_0 (dB-Hz)		82.5	$(P_r/N_0)_r$
17.	Data rate (dB-bit/s)	(2 Mbits/s)	⟨63.0⟩	R
18.	Received E_b/N_0 (dB)		19.5	$(E_b/N_0)_r$
19.	Implementation loss (dB)		⟨1.5⟩	L_o
20.	Required E_b/N_0 (dB)		⟨10.0⟩	$(E_b/N_0)_{reqd}$
21.	Margin (dB)		8.0	M

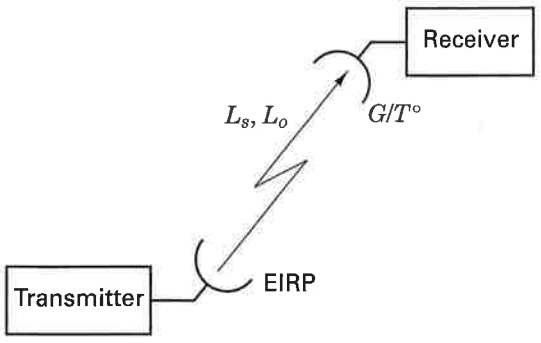


Figure 5.23 Key parameters of a link analysis.

are performed as in Equation (5.24), which is repeated here, with the exception that the terms G_r and T° are grouped together at G_r/T° instead of being listed separately:

$$M(\text{dB}) = \text{EIRP}\,(\text{dBW}) + \frac{G_r}{T^\circ}(\text{dB/K}) - \left(\frac{E_b}{N_0}\right)_{\text{reqd}}(\text{dB}) - R(\text{dB-bits/s}) - \kappa\,(\text{dBW/K-Hz}) - L_s\,(\text{dB}) - L_o\,(\text{dB})$$

Let us examine the 21 line items listed in Table 5.2.

1. Transmitter power is 100 W (20 dBW).
2. Circuit loss between the transmitter and antenna is 2 dB.
3. Transmitting antenna gain is 51.6 dBi.
4. The net tally of items 1 to 3 yields the EIRP = 69.6 dBW.
5. The path loss has been calculated for the range shown in the table title, corresponding to a 10° elevation angle at the earth terminal.

6 **and 7.** Here are allowances made for weather fades and a variety of other, unspecified losses.

8. Received isotropic power refers to the power that would be received, –143.1 dBW, if the receiving antenna were isotropic.
9. The peak gain of the receiving antenna is 35.1 dBi.
10. Edge-of-coverage loss is due to the off-axis antenna gain (compared to peak gain) and to the increased range for users at the extreme edge of communication coverage (a nominal 2-dB loss is shown here.)
11. The input power to the receiver, tallied from items 8, 9, and 10, is –110 dBW.
12. System temperature is found using Equation (5.46). However, in this example we are assuming a lossless line from the receiver antenna to the front end, so that the line loss factor L is equal to 1, and system temperature is computed in column 3, as $T_S^\circ = T_A + T_R$.
13. We form the receiver figure-of-merit ratio G/T° by combining the gain of the receiver antenna G_r (see item 9) with system temperature T_S. This ratio is placed in the left column as a parameter of interest, rather than in the middle column. This is because G_r is accounted for in link budget item 9, and T_S is accounted for in item 15. If G/T° were to be placed in the center column, it would represent a double tabulation.
14. Boltzmann's constant is –228.6 dBW/K-Hz.
15. Boltzmann's constant in decibels (item 14), plus system temperature in decibels (item 12), yields noise power spectral density.
16. Finally, we can form the received signal-to-noise spectral density, 82.5 dB-Hz, by subtracting noise spectral density in decibels (item 15), from received signal power in decibels (item 11).
17. The data rate is listed in dB-bit/s.

18. Since $E_b/N_0 = (1/R)\ (P_r/N_0)$, we need to subtract R in decibels (item 17), from P_r/N_0 in decibels (item 16), yielding $(E_b/N_0)_r = 19.5$ dB.
19. An implementation loss, here taken to be 1.5 dB, accounts for the difference between theoretically predicted detection performance, and the performance of the actual detector.
20. This is our required E_b/N_0, a result of the modulation and coding chosen, and the probability of error specified.
21. The difference between the received and the required E_b/N_0 in decibels (taking implementation loss into account), yields the final margin.

The gain or loss items shown in a link budget, generally follow the convention of first presenting an *ideal* or *simplistic* result, followed by a gain or loss term that modifies the simplistic yielding an *actual result.* In other words, the link budget typically follows a *modular* approach to partitioning the gains and losses in a way that easily adapts to the needs of any system. Consider the following examples of this format. In Table 5.2, item 1 gives the transmitter power that would be launched from a transmitter with an isotropic transmitting antenna (the simplistic). However, only after applying the modules of circuit loss and transmitter-antenna gain do we see, in item 4, the EIRP (actually) launched. Similarly, item 8 shows the power received by an isotropic antenna (the simplistic). However, only after applying the modules of receiver-antenna gain and edge-of-coverage loss do we see, in item 11, the (actually) received signal power.

5.6.2 Receiver Figure of Merit

An explanation of why receiving antenna gain and system temperature are often grouped together as $G/T°$, is as follows. In the early days of satellite communications development, the G_r and the $T_S°$ were specified separately. A contractor who agreed to meet these specifications would need to allow himself some safety margin for meeting each specification. Even though the user was generally only interested in the "bottom-line" performance, and not in the explicit value of G_r or $T_S°$, the contractor would not be able to exploit potential trade-offs. The net result was an overspecified (more costly) system than was necessary. Recognition of such overspecification resulted in specifying the antenna and receiver front end as a single figure-of-merit parameter $G/T°$ (sometimes called the *receiver sensitivity*), such that cost-effective trade-offs between the antenna design and the receiver design might be employed.

5.6.3 Received Isotropic Power

Another recognized area of overspecification in receiver design is in the separate specification of the required P_r/N_0 (or E_b/N_0) and receiver $G/T°$. If P_r/N_0 and $G/T°$ are specified separately, the system contractor is forced to meet each value. The contractor will plan for a margin in both places. As in the $G/T°$ case of the preceding section, there are advantages in specifying P_r/N_0 and $G/T°$ as one parameter;

this new parameter, called the *received isotropic power* (RIP), can be written as follows:

$$\text{RIP (dBW)} = \frac{P_r}{N_0}\text{(dB-Hz)} - \frac{G}{T^\circ}\text{(dB/K)} + \kappa\text{ (dBW/K-Hz)} \tag{5.49}$$

Or, in terms of ratios, it also can be written as

$$\text{RIP} = \frac{P_r}{\kappa T^\circ}\left(\frac{\kappa T^\circ}{G_r}\right) = \frac{P_r}{G_r} \tag{5.50}$$

It is important to note that P_r/N_0 refers to the predetection signal-to-noise spectral density ratio (SNR) *required* for a particular error probability when using a particular modulation scheme (it usually includes an allowance for *detector implementation losses*). Let us designate the theoretically required SNR to yield a particular P_B as $(P_r/N_0)_{\text{th-rq}}$. We can therefore write

$$\frac{P_r}{N_0} = L_o'\left(\frac{P_r}{N_0}\right)_{\text{th-rq}} \tag{5.51}$$

where L_o' is called the implementation loss and accounts for the hardware and operational losses in the detection process. Combining Equations (5.50) and (5.51), we can write

$$\text{RIP} = L_o'\left(\frac{P_r}{\kappa T^\circ}\right)_{\text{th-rq}}\frac{\kappa T^\circ}{G_r} \tag{5.52}$$

Specifying the RIP required to meet the system error performance allows the contractor to commit to meeting a single parameter value. The contractor is allowed to trade off P_r/N_0 versus G/T° and L_o' performance. As G/T° is improved, the detector performance can be degraded, and vice versa.

5.7 SATELLITE REPEATERS

Satellite repeaters retransmit the messages they receive (with a translation in carrier frequency). A *regenerative* (digital) repeater regenerates, that is, demodulates and reconstitutes the digital information embedded in the received waveforms before retransmission; however, a *nonregenerative* repeater only amplifies and retransmits. A nonregenerative repeater, therefore, can be used with many different modulation formats (simultaneously or sequentially without any switching), but a regenerative repeater is usually designed to operate with only one, or a very few, modulation formats. A link analysis for a regenerative satellite repeater treats the uplink and downlink as two separate point-to-point analyses. To calculate the overall bit error performance of a regenerative repeater link, it is necessary to determine separately the bit error probability on the uplink and downlink. Let P_u and P_d be the probability of a bit being in error on the uplink and downlink, respectively. A bit will be correct in the end-to-end link if either the bit is correct on both the

up- and downlink, or if it is in error on both the up- and downlink. Therefore, the overall probability that a bit is correct is

$$P_c = (1 - P_u)(1 - P_d) + P_u P_d \tag{5.53}$$

and the overall probability that a bit is in error is

$$P_B = 1 - P_c = P_u + P_d - 2P_u P_d \tag{5.54}$$

For low values of P_u and P_d, the overall bit error performance is approximated simply by summing the individual uplink and downlink bit error probabilities:

$$P_B \approx P_u + P_d \tag{5.55}$$

5.7.1 Nonregenerative Repeaters

Link analysis for a nonregenerative repeater treats the entire "round trip" (uplink transmission to the satellite and downlink retransmission to an earth terminal) as a single analysis. Features that are unique to nonregenerative repeaters, are the dependence of the overall SNR on the uplink SNR and the sharing of the repeater downlink power in proportion to the uplink power from each of the various uplink signals and noise. Henceforth, reference to a repeater or transponder will mean a *nonregenerative repeater,* and for simplicity, we will assume that the transponder is operating in its linear range.

A satellite transponder is limited in transmission capability by its downlink power, the earth terminal's uplink power, satellite and earth terminal noise, and channel bandwidth. One of these usually is a dominant performance constraint; most often the downlink power or the channel bandwidth proves to be the major system limitation. Figure 5.24 illustrates the important link parameters of a linear satellite repeater channel. The repeater transmits all uplink signals (or noise, in the absence of signal) without any processing beyond amplification and frequency translation. Let us assume that there are multiple simultaneous uplinks within the receiver's bandwidth W and that they are separated from one another through the use of a technique called *frequency-division multiple access* (FDMA). FDMA is a communications resource-sharing technique whereby different users occupy disjoint portions of the transponder bandwidth; FDMA is treated in Chapter 11. The satellite effective downlink power EIRP_s is constant and since we are assuming a linear transponder, EIRP_s is shared among the multiple uplink signals (and noise) in proportion to their respective input power levels.

The transmission starts from a ground station (bandwidth $\leq W$), say terminal i, with a terminal $\text{EIRP}_{ti} = P_{ti}G_{ti}$. Simultaneously, other signals are being transmitted to the satellite (from other terminals). The EIRP from the kth terminal will henceforth be referred to simply as P_k. At the satellite, a total signal power $P_T = \Sigma A_k P_k$ is received, where A_k reflects the uplink propagation loss and the satellite receive antenna gain for each terminal. $N_s W$ is the satellite uplink noise power, where N_s is the composite noise power spectral density due to noise radiated into the satellite antenna *and* generated in the satellite receiver. The total satellite downlink $\text{EIRP}_s = P_s G_{ts}$, where P_s is the satellite transponder output power and G_{ts}

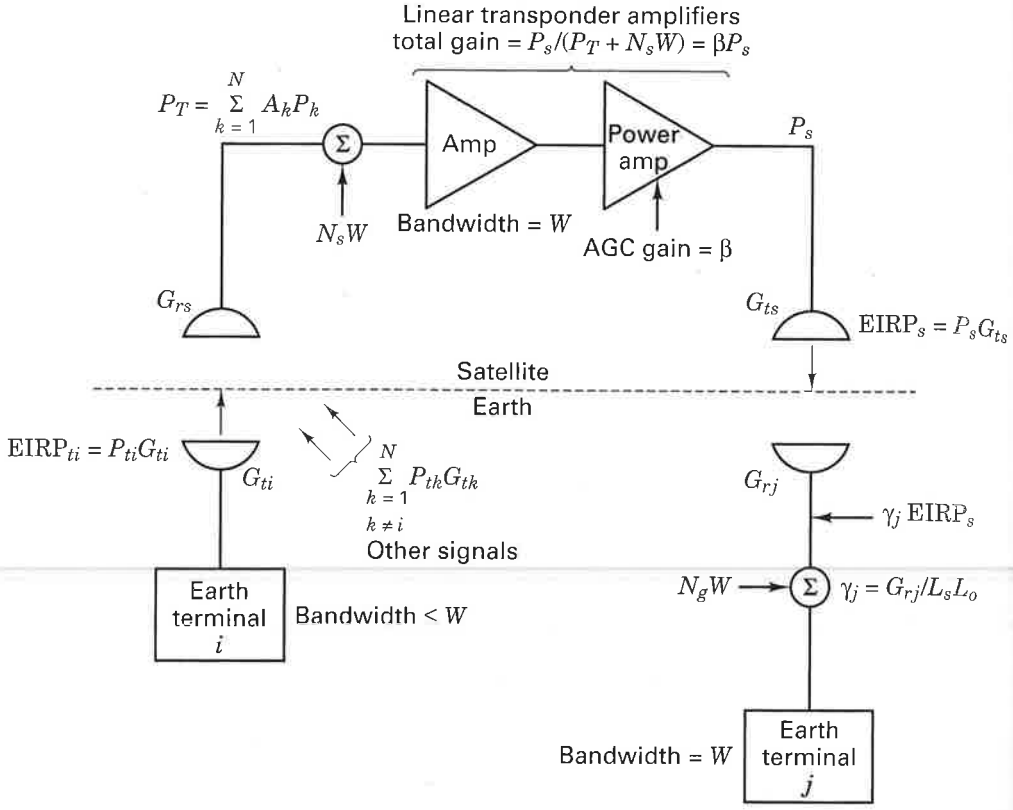


Figure 5.24 Nonregenerative satellite repeater.

is the satellite transmitting antenna gain, can be expressed by the following identity [14]:

$$\text{EIRP}_s = \text{EIRP}_s \beta [A_i P_i + (P_T - A_i P_i) + N_s W] \tag{5.56}$$

Both the left and right sides of Equation (5.56) express the total satellite EIRP. On the right side, the term $\beta[A_i P_i + (P_T - A_i P_i) + N_s W]$ constitutes the fractional apportionment of EIRP_s for the various users and uplink noise, such that the composite expression equals unity. The usefulness of this identity should become clear shortly. The total power gain of the transponder can be expressed as βP_s. Since P_s is fixed and the input signals can vary, $\beta = 1/(P_T + N_s W)$ represents an automatic gain control (AGC) term. The total received uplink signal power, P_T, has purposely been written as $A_i P_i + (P_T - A_i P_i)$ to separate signal i power from the remainder of the simultaneous signals in the transponder. The total power received at the jth earth terminal, with bandwidth W, can be written as

$$P_{rj} = \text{EIRP}_s \gamma_j \beta [A_i P_i + (P_T - A_i P_i) + N_s W] + N_g W \tag{5.57}$$

where $\gamma_j = G_{rj}/L_s L_o$ accounts for downlink losses and receiving antenna gain for the jth earth terminal. $\text{EIRP}_s\gamma_j$ represents the portion of EIRP_s that is received by the jth earth terminal, and N_g is the downlink noise power spectral density generated and introduced into that terminal receiver. Equation (5.57) describes the essence of downlink power apportionment among the various users and noise in a repeater. Let us rewrite Equation (5.57) by replacing β with its equivalent $1/(P_T + N_s W)$, as follows:

$$P_{rj} = \text{EIRP}_s\gamma_j \left(\frac{A_i P_i}{P_T + N_s W} + \frac{P_T - A_i P_i}{P_T + N_s W} + \frac{N_s W}{P_T + N_s W} \right) + N_g W \qquad (5.58)$$

To facilitate our discussion, let us amplify Equation (5.58) with words, yielding

$$P_{rj} = \text{EIRP}_s\gamma_j \left(\frac{S_i \text{ U/L power}}{\text{total } (S + N) \text{ U/L power}} + \frac{\text{balance } S \text{ U/L power}}{\text{total } (S + N) \text{ U/L power}} + \frac{\text{U/L noise power}}{\text{total } (S + N) \text{ U/L power}} \right) + N_g W$$

where S stands for signal power, N for noise power, and U/L for uplink.

From Equation (5.58), can you recognize an important relationship that must exist among multiple users sharing a nonregenerative transponder? The users must *cooperate with one another,* by not exceeding agreed-upon uplink transmission power levels. Equation (5.58) states that the portion of the downlink EIRP dedicated to any one user (or to uplink noise) is determined by the ratio of that user's uplink power to the total uplink signal plus noise power. Hence if one of the sharing users should choose to "cheat" by increasing his or her uplink power, the effect would be an enhancement of this user's downlink signal level, at the expense of the other users' downlink signal levels. Notice from Equation (5.58) that the uplink noise shares the downlink resource along with the other users. This coupling of uplink noise onto the downlink is a feature unique to nonregenerative repeaters.

From Equation (5.58) we can express the P_r/N for signal i received at the jth terminal as

$$\left(\frac{P_r}{N}\right)_{ij} \approx \frac{\text{EIRP}_s\gamma_j[A_i P_i/(P_T + N_s W)]}{\text{EIRP}_s\gamma_j[N_s W/(P_T + N_s W)] + N_g W} \qquad (5.59)$$

and we can write the overall P_r/N_0 for signal i received at the jth terminal as [14]

$$\left(\frac{P_r}{N_0}\right)_{ij} = \frac{\text{EIRP}_s\gamma_j\beta A_i P_i}{\text{EIRP}_s\gamma_j\beta N_s + N_g} \qquad (5.60)$$

Equations (5.58) to (5.60) illustrate that the uplink repeater noise degrades the overall SNR in two ways—it "steals" downlink EIRP, and it contributes to the total system noise. When the satellite uplink noise dominates—that is, when $P_T << N_s W$, the link is said to be *uplink limited,* and most of the downlink EIRP_s is wastefully allocated to uplink noise power. When this is the case and when $\text{EIRP}_s\gamma_j >> N_g W$, we can rewrite Equation (5.60) as

$$\left(\frac{P_r}{N_0}\right)_{ij} \approx \frac{\text{EIRP}_s \gamma_j A_i P_i / N_s W}{(\text{EIRP}_s \gamma_j / W) + N_g} \approx \frac{A_i P_i}{N_s} \tag{5.61}$$

Equation (5.61) illustrates that in the case of an uplink limited channel, the overall P_r/N_0 ratio essentially follows the uplink SNR. The more common situation is the *downlink limited* channel, in which case $P_T >> N_s W$, and the satellite EIRP is limited. In this case, Equation (5.60) can be rewritten as

$$\left(\frac{P_r}{N_0}\right)_{ij} \approx \frac{\text{EIRP}_s \gamma_j A_i P_i / P_T}{N_g} \tag{5.62}$$

The power of the transponder is then shared primarily among the various uplink transmitted signals; very little uplink noise is transmitted on the downlink. The performance of the repeater, in this case, is constrained only by its downlink parameters.

Table 5.3 illustrates a link analysis example (full round trip) for a nonregenerative repeater. The uplink portion by itself does not constitute a link budget since the transmission is not demodulated at the satellite. Without demodulation, *there are no bits* and therefore there is no way to measure the bit-error performance. After the full round trip, the signal is demodulated at the earth terminal; only then does the link analysis yield the margin. The example in Table 5.3 represents a case where the satellite transponder is servicing 10 simultaneous users. In the block marked "A" is shown the ratio $P_r/(P_T + N_s W)$, which dictates the apportionment of the downlink EIRP for the signal of interest. In this example, with all users transmitting the same power level, each of the signals is allocated 9.8% of the downlink EIRP. In the block marked "B" we see the apportionment of the downlink EIRP. The total is 1514.7 W; the user of interest gets 148.5 W; the other nine users get a total of 1336.1 W; and the uplink noise is apportioned 30.1 W.

An estimate of the performance described in Equation (5.60) can be obtained by using the uplink and downlink values of E_b/N_0 (or P_r/N_0), combined as follows, in the *absence of intermodulation noise* [15].

$$\left(\frac{E_b}{N_0}\right)_{\text{ov}}^{-1} = \left(\frac{E_b}{N_0}\right)_{u}^{-1} + \left(\frac{E_b}{N_0}\right)_{d}^{-1} \tag{5.63}$$

where the subscripts ov, *u*, and *d*, indicate overall, uplink, and downlink values of E_b/N_0, respectively.

Most commercial satellite transponder designs are nonregenerative. However, it seems clear that future commercial systems will require on-board processing, switching, or selective message addressing, and will use regenerative repeaters to transform the received waveforms to message bits. Besides the potential for sophisticated data processing, one of the principal advantages of regenerative compared to nonregenerative repeaters is that the uplink is decoupled from the downlink so that the uplink noise is not retransmitted on the downlink. There are significant performance improvements possible with regenerative satellite repeaters in terms of the E_b/N_0 values needed on the uplinks and downlinks, relative to the values needed for the conventional nonregenerative designs in use today. Improvements of as much as 5 dB on the uplink and 6.8 dB on the downlink (using coherent QPSK modulation, with $P_B = 10^{-4}$) have been demonstrated [16].

TABLE 5.3 Link Budget Example For a Nonregenerative Satellite Repeater with 10 Users: Uplink Frequency = 375 MHz, Downlink Frequency = 275 MHz, Range = 22,000 Nautical Miles

		Uplink			Downlink		
Transmitter power (dBW)		27.0	(500.0 W)		13.0	(20.0 W)	
Transmitter circuit losses (dB)		1.0			1.0		
Transmitter antenna gain (peak-dBi)		19.0			19.8		
Dish diameter (ft)	10.00			15.00			
Half-power beamwidth (degrees)	19.16			17.42			
EIRP (dBW)		45.0			31.8	(1514.7 W)	
Path loss (dB)		176.1			173.4		
Transmitted signal power (dBW)					21.7	(148.5 W)	B
Transmitted other signal power (dBW)					31.3	(1336.1 W)	
Transmitted U/L noise power (dBW)					14.8	(30.1 W)	
Other losses (dB)		2.0			2.0		
Received isotropic signal power (dBW)		−133.1			−153.7		
Received isotropic U/L noise power (dBW)					−160.6		
Receiver antenna gain (peak dBi)		22.5			16.3		
Dish diameter (ft)	15.00			10.00			
Half-power beamwidth (degrees)	12.77			26.13			
Received signal power (dBW)		−110.6			−137.4		
Received U/L noise power (dBW)					−144.3		
Receiver antenna temperature (dB-K)		24.6	(290 K)		20.0	(100 K)	
Receiver noise figure at antenna port (dB)		10.8			2.0		
Receiver temperature (dB-K)		35.1	(3197 K)		22.3	(170 K)	
System temperature (dB K)		35.4	(3487 K)		24.3	(270 K)	
System $G/T°$ (dB/K)		−12.9			−8.0		
Boltzmann's constant (dBW/K-Hz)		−228.6			−228.6		
Noise spectral density (dBW/Hz)		−193.2			−204.3		
System bandwidth (dB-Hz)		75.6	(36.0 MHz)		75.6	(36.0 MHz)	
Noise power (dBW)		−117.6			−128.7		
U/L noise + D/L noise power (dBW)					−128.6		
Simultaneous accesses	10						
Received other signal power (dBW)		−101.1					
Other signals + noise (dBW)		−101.0					
$P_r/(P_T + N_sW)$ (dB)		−10.1	(0.098)	A			
P_r/N (dB)		7.0			−8.7		
Overall P_r/N (dB)					−8.8		
P_r/N_0 (dB-Hz)		82.6			66.9		
Overall P_r/N_0 (dB-Hz)					66.8		
Data rate (dB-bit/s)					50.0	(100,000 bits/s)	
Available E_b/N_0 (dB)					16.8		
Required E_b/N_0 (dB)					10.0		
Margin (dB)					6.8		

5.7.2 Nonlinear Repeater Amplifiers

Power is severely limited in most satellite communication systems, and the inefficiencies associated with linear power amplification stages are expensive to bear. For this reason, many satellite repeaters employ nonlinear power amplifiers. Efficient power amplification is obtained at the cost of signal distortion due to nonlinear operation. The major undesirable effects of the repeater nonlinearities are:

1. Intermodulation (IM) noise due to the interaction of different carriers. The harm is twofold; useful power can be lost from the channel as IM energy (typically 1 to 2 dB), and spurious IM products can be introduced into the channel as interference. The latter problem can be quite serious.
2. AM-to-AM conversion is a phenomenon common to nonlinear devices such as traveling wave tubes (TWT). At the device input, any signal-envelope fluctuations (amplitude modulation) undergo a nonlinear transformation and thus result in amplitude distortion at the device output. Hence, a TWT operating in its nonlinear region would not be the optimum power-amplifier choice for an amplitude-based modulation scheme (such as QAM).
3. AM-to-PM conversion is another phenomenon common to nonlinear devices. Fluctuations in the signal envelope produce phase variations that can affect the error performance for any phase-based modulation scheme (such as PSK or DPSK).
4. In hard limiters, weak signals can be suppressed, relative to stronger signals, by as much as 6 dB [2]. In saturated TWTs, the suppression of weak signals is due not only to limiting, but also to the fact that the signal coupling mechanism of the tube is optimized in favor of the stronger signals. The effect can cause weak signals to be suppressed by as much as 18 dB [17].

Conventional nonregenerative repeaters are generally operated *backed-off* from their highly nonlinear saturated region; this is done to avoid appreciable IM noise and thus to allow efficient utilization of the system's entire bandwidth. However, backing off to the linear region is a compromise; some level of IM noise must be accepted to achieve a useful level of output power.

5.8 SYSTEM TRADE-OFFS

The link budget example in Table 5.3 is a resource allocation document. With such a link tabulation, one can examine potential system trade-offs and attempt to optimize system performance. The link budget is a natural starting point for considering all sorts of potential trade-offs: margin versus noise figure, antenna size versus transmitter power, and so on. Table 5.4 represents an example of a computer exercise for examining a possible trade-off between the earth station transmitting power and the system noise margin at the receiving terminal. The first row in the table is taken from the Table 5.3 link budget. Suppose a system engineer is concerned that a 500-W transmitter is not practical because of some physical con-

TABLE 5.4 Potential Trade-Off: P_t versus Margin

P_t (W)	$(P_r/N_0)_u$ (dB-Hz)	$(P_r/N_0)_d$ (dB-Hz)	$(P_r/N_0)_{ov}$ (dB-Hz)	Margin (dB)
500.0	82.6	66.9	66.8	6.8
250.0	79.6	66.8	66.6	6.6
125.0	76.6	66.6	66.2	6.2
62.5	73.6	66.3	65.5	5.5
31.3	70.5	65.7	64.5	4.5
15.6	67.5	64.8	62.9	2.9
7.8	64.5	63.3	60.8	0.8
3.9	61.5	61.4	58.4	−1.6
2.0	58.4	59.0	55.7	−4.3
1.0	55.4	56.4	52.9	−7.2
0.5	52.4	53.6	49.9	−10.1

straints within the transmitting earth terminal or that such a transmitter makes the system "uplink rich" (a poor design point). The engineer might then consider a trade-off of transmitter power versus thermal noise margin. The listing of candidate trade-offs is a trivial task for a computer. Table 5.4 was generated by repeating the link budget computation multiple times, and at each iteration, reducing P_t by one-half.

The result is a selection of transmitters (in steps of 3 dB) and uplink, downlink, and overall SNRs, and margin, associated with each transmitter value. The system engineer need only peruse the list to find a likely candidate. For example, if the engineer were satisfied with a margin of 3 to 4 dB, it appears he could reduce the transmitter from 500 W to 20 or 30 W. Or, he might be willing to provide a transmitter with, say, $P_t = 100$ W, since he may want to consider additional trade-offs (perhaps because of having misgivings about one of the other subsystems, say the antenna size). The engineer would then start a new tabulation with $P_t = 100$ W, and again perform a succession of link budget computations, to produce a similar enumeration of other possible trade-offs.

Notice from Table 5.4 that one can recognize the uplink-limited and downlink-limited regions, discussed earlier. In the first few rows, where the uplink SNR is high, a 3-dB degradation in uplink SNR results in only a few tenths of a decibel degradation to the overall SNR. Here the system is *downlink limited;* that is, the system is constrained primarily by its downlink parameters and is hardly affected by the uplink parameters. In the bottom few rows of the table, we see that a 3-dB degradation to the uplink affects the overall SNR by almost 3 dB. Here the system is *uplink limited;* that is, the system is constrained primarily by the uplink parameters.

5.9 CONCLUSION

Of the many analyses that support a developing communication system, the link budget stands out in its ability to provide overall system insight. By examining the link budget, one can learn many things about the overall system design and performance. For example, from the link margin, one learns whether the system will meet

its requirements comfortably, marginally, or not at all. It will be evident if there are any hardware constraints, and whether such constraints can be compensated for in other parts of the link. The link budget is often used for considering system trade-offs and configuration changes, and in understanding subsystem nuances and interdependencies. Together with other modeling techniques, the link budget can help predict weight, size, and cost. We have considered how to formulate this budget and how it might be used for system trade-offs. The link budget is one of the system manager's most useful documents; it represents a "bottom-line" tally in the search for optimum error performance of the system.

REFERENCES

1. Panter, P. F., *Communication Systems Design: Line-of-Sight and Tropo-Scatter Systems,* R. E. Krieger Publishing Co, Inc., Melbourne, Fla., 1982.
2. Jones, J. J., "Hard Limiting of Two Signals in Random Noise," *IEEE Trans. Inf. Theory,* vol. IT9, January 1963.
3. Silver, S., *Microwave Antenna Theory and Design,* MIT Radiation Laboratory Series, Vol. 12, McGraw-Hill Book Company, New York, 1949.
4. Kraus, J. D., *Antennas,* McGraw-Hill Book Company, New York, 1950.
5. Johnson, J. B., "Thermal Agitation of Electricity in Conductors," *Phys. Rev.,* vol. 32, July 1928, pp. 97–109.
6. Nyquist, H., "Thermal Agitation of Electric Charge in Conductors," *Phys. Rev.,* vol. 32, July 1928, pp. 110–113.
7. Desoer, C. A., and Kuh, E. S., *Basic Circuit Theory,* McGraw-Hill Book Company, New York, 1969.
8. Schwab, L. M., "World-Wide Link Availability for Geostationary and Critically Inclined Orbits Including Rain Attenuation Effects," *Lincoln Laboratory, Rep. DCA-9,* Jan. 27, 1981.
9. Friis, H. T., "Noise Figure of Radio Receivers," *Proc. IRE,* July 1944, pp. 419–422.
10. IRE Subcommittee 7.9 on Noise, "Description of the Noise Performance of Amplifiers and Receiving Systems," *Proc. IEEE,* Mar. 1963, pp. 436–442.
11. Blackwell, L. A., and Kotzebue, K. L., *Semiconductor Diode Parametric Amplifiers,* Prentice-Hall, Inc., Englewood Cliffs, N.J., 1961.
12. Ko, H. C., and Kraus, J. D., "A Radio Map of the Sky at 1.2 Meters," *Sky Telesc.,* vol. 16, Feb. 1957, pp. 160–161.
13. Piddington, J. H., and Trent, G. H., "A Survey of Cosmic Radio Emission at 600 Mc/s," *Aust. J. Phys.,* vol. 9, Dec. 1956, pp. 481–493.
14. Spilker, J. J., *Digital Communications by Satellite,* Prentice-Hall, Inc., Englewood Cliffs, N.J., 1977.
15. Pritchard, W. L., and Sciulli, J. A., *Satellite Communication Systems Engineering,* Prentice-Hall, Inc., Englewood Cliffs, N.J., 1986.
16. Campanella, S. J., Assal, F., and Berman, A., "Onboard Regenerative Repeaters," *Int. Conf. Commun.,* Chicago, vol. 1, 1977, pp. 6.2-121–66.2-125.
17. Wolkstein, H. J., "Suppression and Limiting of Undesired Signals in Travelling-Wave-Tube Amplifiers," Publication ST-1583, *RCA Rev.,* vol. 22, no. 2, June 1961, pp. 280–291.

PROBLEMS

5.1. **(a)** What is the value in decibels of the free-space loss for a carrier frequency of 100 MHz and a range of 3 miles?

(b) The transmitter output power is 10 W. Assume that both the transmitting and receiving antennas are isotropic and that there are no other losses. Calculate the received power in dBW.

(c) If in part (b) the EIRP is equal to 20 W, calculate the received power in dBW.

(d) If the diameter of a dish antenna is doubled, calculate the antenna gain increase in decibels.

(e) For the system of part (a), what must the diameter of a dish antenna be in order for the antenna gain to be 10 dB? Assume an antenna efficiency of 0.55.

5.2. A transmitter has an output of 2 W at a carrier frequency of 2 GHz. Assume that the transmitting and receiving antennas are parabolic dishes each 3 ft in diameter. Assume that the efficiency of each antenna is 0.55.

(a) Evaluate the gain of each antenna.

(b) Calculate the EIRP of the transmitted signal in units of dBW.

(c) If the receiving antenna is located 25 miles from the transmitting antenna over a free-space path, find the available signal power out of the receiving antenna in units of dBW.

5.3. From Table 5.1 we see that the proposal from Satellite Television Corporation called for a direct broadcast satellite (DBS) EIRP of 57 dBW and a downlink transmission frequency of 12.5 GHz. Assume that the only loss is the downlink space loss shown. Suppose that the downlink information consists of a digital signal with a data rate of 5×10^7 bits/s. Assume that the required E_b/N_0 is 10 dB, the system temperature at your home receiver is 600 K, and that your rooftop dish has an efficiency of 0.55. What is the minimum dish diameter that you can use in order to close the link? Do you think the neighbors will object?

5.4. An amplifier has an input and output resistance of 50 Ω, a 60-dB gain, and a bandwidth of 10 kHz. When a 50-Ω resistor at 290 K is connected to the input, the output rms noise voltage is 100 μV. Determine the effective noise temperature of the amplifier.

5.5. An amplifier has a noise figure of 4 dB, a bandwidth of 500 kHz, and an input resistance of 50 Ω. Calculate the input signal voltage needed to yield an output SNR = 1 when the amplifier is connected to a signal source of 50 Ω at 290 K.

5.6. Consider a communication system with the following specifications: transmission frequency = 3 GHz, modulation format is BPSK, bit-error probability = 10^{-3}, data rate = 100 bits/s, link margin = 3 dB, EIRP = 100 W, receiver antenna gain = 10 dB, distance between transmitter and receiver = 40,000 km. Assume that the line loss between the receiving antenna and the receiver is negligible.

(a) Calculate the maximum permissible noise power spectral density in watts/hertz referenced to the receiver input.

(b) What is the maximum permissible effective noise temperature in kelvin for the receiver if the antenna temperature is 290 K?

(c) What is the maximum permissible noise figure in dB for the receiver?

5.7. A receiver preamplifier has a noise figure of 13 dB, a gain of 60 dB, and a bandwidth of 2 MHz. The antenna temperature is 490 K, and the input signal power is 10^{-12} W.

(a) Find the effective temperature, in kelvin, of the preamplifier.

(b) Find the system temperature in kelvin.

(c) Find the output SNR in decibels.

5.8. Assume that a receiver has the following parameters: gain = 50 dB, noise figure = 10 dB, bandwidth = 500 MHz, input signal power = 50×10^{-12} W, source temperature, $T_A^\circ = 10$ K, line loss = 0 dB. You are asked to insert a preamplifier between the antenna and the receiver. The preamplifier is to have a gain of 20 dB and a bandwidth of 500 MHz. Find the preamplifier noise figure that would be required to provide a 10-dB improvement in overall system SNR.

5.9. Find the maximum allowable effective system temperature, T_S°, required to *just close* a particular link with a bit error probability of 10^{-5} for a data rate of $R = 10$ kbits/s. The link parameters are as follows: transmission frequency = 12 GHz, EIRP = 10 dBW, receiver antenna gain = 0 dB, modulation type is noncoherently detected BFSK, other losses = 0 dB, and the distance between transmitter and receiver = 100 km.

5.10. Consider a receiver made up of the following three stages: The input stage is a preamplifier with a gain of 20 dB and a noise figure of 6 dB. The second stage is a 3-dB lossy network. The output stage is an amplifier with a gain of 60 dB and a noise figure of 16 dB.

(a) Find the composite noise figure for the receiver.

(b) Repeat part (a) with the preamplifier removed.

5.11. **(a)** Find the effective input noise temperature, T_R°, of a receiver comprised of three amplifier stages connected in series with power gains, from input to output, of 10, 16, and 20 dB, and effective noise temperatures, from input to output, of 1800, 2700, and 4800 K.

(b) What would the gain of the first stage have to be to reduce the contribution to T_R° of all stages after the first to 10% of the first-stage contribution?

5.12. The effective temperature of a particular multiple-stage receiver is required to be 300 K. Assume that the effective temperatures and gains of stages 2 through 4 are as follows: $T_2^\circ = 600$ K, $T_3^\circ = T_4^\circ = 2000$ K, $G_2 = 13$ dB, and $G_3 = G_4 = 20$ dB.

(a) Compute the required gain, G_1, for the first stage, under the conditions that $T_1^\circ = 200, 230, 265, 290, 295$, and 300 K.

(b) Plot the G_1 versus T_1° trade-off.

(c) Regarding contributions to the effective temperature of the receiver, why is it reasonable in this case to ignore all stages beyond the fourth stage?

(d) In a practical engineering trade-off between T_1° and G_1 what range of T_1° values do you think should be considered?

5.13. A receiver consists of a preamplifier followed by multiple amplifier stages. The composite effective temperature of all the amplifier stages is 1000 K, referenced to the preamplifier output.

(a) Compute the receiver effective noise temperature, referenced to the preamplifier input, for a single-stage preamplifier with a noise temperature of 400 K and gains of 3, 6, 10, 16, and 20 dB.

(b) Repeat part (a) for a two-stage preamplifier with 400-K noise per stage and gains of 3, 6, 10, and 13 dB per stage.

(c) Plot the receiver effective temperature versus the gain of the first stage for parts (a) and (b).

5.14. **(a)** Equation (5.42) shows the composite noise figure for a network made up of a lossy line followed by an amplifier. Develop a general expression for the composite noise figure of three such networks connected in series.

(b) Consider a network that is made up of an amplifier followed by a lossy line. Develop a general expression for the composite noise figure of three such networks connected in series.