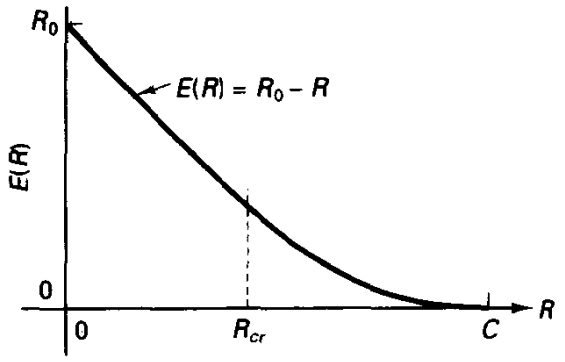

Figure 4.4.5 Error exponent $E(R)$ for general DMC.

4.4.4 Summary of Coding Potential for Block Codes on DMCs

We have developed two primary results from random coding arguments for block codes. The first, and easiest to develop, was

$$\overline{P(e|i)} < 2^{-n(R_0 - R)}, \tag{4.4.28}$$

which claims that the error probability for the ith codeword in the ensemble of codewords of block length n and rate R is *exponentially* decreasing as block length increases, at least if $R < R_0$. The second and stronger result is that

$$\overline{P(e|i)} < 2^{-nE(R)}, \tag{4.4.29}$$

where $E(R)$ is positive for all $R < C$.

Both results were developed for a specific codeword index i in the ensemble, but it is clear that the result is independent of the index i, so the bounds hold for the error probability averaged over the 2^{nR} codewords; that is,

$$\overline{P(e)} < 2^{-nE(R)}. \tag{4.4.30}$$

Since at least one code in the ensemble has error probability as good as the ensemble average (4.4.30), we then have the fundamental coding theorem for discrete memoryless channels:

Noisy Channel Coding Theorem

Given a DMC with capacity C, there exists a sequence of codes of increasing block length n, *each with fixed rate $R < C$, for which the error probability $P(e) = \sum P_i P(e|i)$ of a maximum likelihood decoder diminishes to zero exponentially fast in* n.

Various extensions of this fundamental result are possible. First, although we now know a code exists whose error probability averaged over all codewords is arbitrarily reliable, it is possible that the error probability conditioned on transmission of the mth codeword in this good code is, in fact, quite poor. Simple arguments are, however, possible to establish the existence of *uniformly good* codes, for which all codewords are reliable [9]. This is particularly important, since in real transmission systems the

prior message probabilities may be unknown. Second, we may ask whether our upper bound is the tightest possible. It turns out that at low rates the ensemble average error probability is dominated by poor codes (recall that the ensemble includes a code with all codewords identical, and this code's error probability would be nearly 1.) Expurgation arguments, which rid the ensemble of such bad codes, can in fact increase the size of the error exponent at low rates, although these same arguments cannot strengthen the bound at rates near capacity.

On the other side of the argument, we can ask what a lower bound on error probability is for rates less than capacity; this delimits a forbidden region for error probability as a function of code rate R, while our upper bound establishes a region where codes in fact are known to exist having a specified performance. The interested reader is referred to texts of Gallager [9] or Blahut [10] for the essential ideas. One important conclusion is that the exponent describing the lower bound coincides with $E(R)$ defined previously for rates approaching capacity, so we can be assured that our argument is essentially the strongest possible statement in this region.

Finally, similar results are obtainable for channels with discrete-input alphabets and continuous-output alphabets, as appropriate for, say, binary antipodal signaling on the AWGN channel. Specifically, we may claim that

$$\overline{P(e)} < 2^{-nE(R)}, \tag{4.4.31}$$

where $E(R)$ is positive for all $R < C$, with C now defined using the discrete-input, continuous-output form for mutual information. The zero-rate intercept of $E(R)$ in this case is again R_0, defined for this channel model in (4.3.32). We will examine these important cases later in the chapter.

Example 4.8 Calculations for the BSC

To illustrate the ideas we have just developed, consider the simple binary symmetric channel with parameter ϵ. Due to symmetry, it is known immediately that the equiprobable assignment on the channel inputs 0 and 1 is optimal in maximizing $E(R)$. By writing out the expression for $E_0(\rho, P)$ with equiprobable input distribution and then differentiating with respect to ρ and evaluating the partial derivative at $\rho = 1$, we have that the critical rate is

$$R_{\text{cr}} = 1 - h_2\left(\frac{\epsilon^{1/2}}{\epsilon^{1/2} + (1-\epsilon)^{1/2}}\right), \tag{4.4.32}$$

where $h_2(x)$ denotes the binary entropy function $h_2(x) = -x \log x - (1-x)\log(1-x)$. For the region $R < R_{\text{cr}}$, using the R_0 expression of Example 4.6, we obtain

$$E(R) = R_0 - R = 1 - \log[1 + (4\epsilon(1-\epsilon))]^{1/2} - R, \qquad R < R_{\text{cr}}. \tag{4.4.33}$$

For rates exceeding the critical rate, but less than capacity, we must resort to a solution of the parametric set of equations (4.4.25):

$$\begin{aligned} R &= 1 - h_2(\delta), \\ E(R) &= t(\delta) - h_2(\delta), \end{aligned} \tag{4.4.34}$$

where $t(\delta) = -\delta \log \epsilon - (1-\delta)\log(1-\epsilon)$.

To numerically interpret the result, suppose that the channel error parameter is $\epsilon = 0.1$. Then the channel capacity is, from (2.7.18a), $C = 0.53$ bit per usage. If we are

aggressive in pursuit of the ultimate limits on performance, we might opt for a coding system with $R = 0.5$ bit per symbol, near the capacity limit. This is well beyond the critical rate, which may be found to be 0.186 bit per channel use, so we must use the parametric form of the error exponent description. Setting $h_2(\epsilon) = 0.5$ yields $\delta = 0.11$, from which

$$E(R) = t(\delta) - h_2(\delta) = 8.1 \cdot 10^{-4}. \tag{4.4.35a}$$

If we wish our ensemble of codes of rate $R = 0.5$ to achieve error probability of 10^{-5}, we must have

$$2^{-n(0.00081)} < 10^{-5}, \tag{4.4.35b}$$

which implies that the block length must be $n = 20510$! Of course, the (very large) best code, if we could find it, might be far better that the average, but this is probably rather discouraging. The random coding arguments simply suggest that operation near the capacity limit with high reliability implies a rather large complexity. We note also that R here significantly exceeds $R_0 = 0.322$ bit/channel symbol, so the simpler bound based on R_0 is useless here.

If, on the other hand, the channel quality is somehow improved so that $\epsilon = 0.01$, channel capacity increases to 0.919 bit/symbol, and R_0 becomes 0.738 bit/symbol. Now operation at the same rate $R = 0.5$ with a target error probability of 10^{-5} implies a block length of (only) 68, as determined by $E(R)$ recalculation. The size of the code with these parameters is still $2^{(0.5)(68)} = 1.7 \cdot 10^{10}$ codewords! Use of the exponent $R_0 - R$ projects a required block length of 70; the closeness of these findings is traceable to the fact that $E(R) \approx R_0 - R$ at rate 0.5.

Example 4.9 Calculations for a 4-ary Erasure Channel

Suppose that QPSK modulation is employed and the channel is an AWGN channel. Instead of supplying the demodulator's best estimate of each symbol, we suppose that an erasure, or low-confidence output, is reported whenever the received two-dimensional vector is not sufficiently near one of the four desired signal locations in terms of phase angle. We suppose the SNR and the erasure policy are such that the correct symbol decision is obtained with probability 0.9, but that an erasure is declared with probability 0.1. We assume there is no chance for an incorrect symbol decision to be produced.

The equivalent DMC is a 4-input, 5-output symmetric erasure channel shown in Figure 4.4.6. The function $E_0(\rho, P)$ for this channel is shown in Figure 4.4.6 as well. It is straightforward to compute the channel capacity to be $C = 1.80$ bits/symbol and $R_0 = 1.62$ bits/symbol. We may determine the critical rate by evaluating the derivative of $E_0(\rho, P)$ at $\rho = 1$, and this is $R_{cr} = 1.40$ bits/symbol. Each has a graphical interpretation indicated in Figure 4.4.6.

If we develop codes for this channel having rate $R = 1$ bit per channel symbol, then, since $R < R_{cr}$, the error exponent would be given by $E(R) = 1.62 - 1.0$, and the ensemble of codes with block length $n = 16$ symbols would have ensemble error probability bounded by $2^{-16(0.62)} = 1.03 \cdot 10^{-3}$.

Example 4.10 Very Noisy Channels

Interesting closed-form results emerge when we consider ***very noisy channels*** [11], which basically are channels for which the conditional probability of output j is nearly the same for all inputs. More precisely,

$$P(j|k) = \gamma_j(1 + \mu_{jk}), \tag{4.4.36}$$

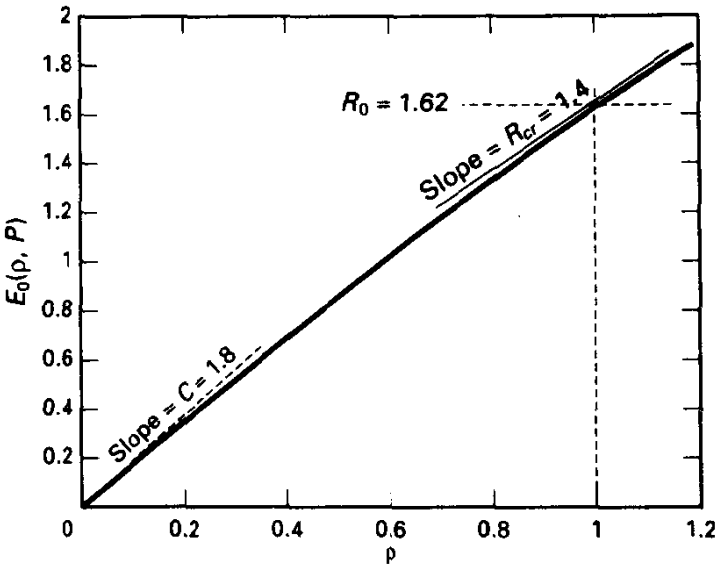

Figure 4.4.6 $E_0(\rho, P)$ for Example 4.8, 4-ary erasure channel.

where μ_{jk} is small compared to 1. Although we omit the details, Gallager [9] shows the following:

1. $R_0 = C/2$
2. $R_{cr} = C/4$
3. $E(R) = C/2 - R$ for $R < C/4$
4. $E(R) = [C^{1/2} - R^{1/2}]^2$ for $R_{cr} \leq R < C$

Thus, the error exponent is simply described for such channels.

A simple application would be a BSC with $\epsilon = 0.3$, for which $C = 0.12$ bit/symbol. Then it follows that $R_0 = 0.06$, and for codes with rate $\frac{1}{12}$, say, the exponent would be $E(R) = 0.0031$. These are in fact approximations to results that could be calculated by the techniques of Example 4.8.

Despite the powerful implications of the channel coding theorem, much practical work remains. We now know codes exist that are good, that is, codes whose performance with increasing block length but fixed rate is arbitrarily reliable. However, from the argument we know almost nothing about how to find them. (The Markov inequality again says, if we simply pick a code at random and use it, with high probability the performance will not be much worse than the ensemble average. However, such codes will in general lack sufficient structure to circumvent table-lookup encoding and exhaustive search decoding.) Since the original development of the noisy channel coding theorem and during its many subsequent refinements, the communication engineering emphasis has been on the description of codes that are reasonably implemented, particularly with respect to decoding effort. We shall devote Chapters 5 and 6 to the two prevalent classes of constructive codes, block codes, as discussed here, and trellis codes.

4.4.5 Remarks for Trellis Codes

We have been focusing on block codes since Section 4.1. However, the R_0 and capacity theory is pertinent to the ensemble of trellis codes as well. We will merely cite the relevant results here for the special case of binary convolutional codes; the interested reader is referred to Viterbi and Omura [12].

We consider a convolutional encoder with rate $R = 1/n$ to be a shift register having m delay elements along with n modulo-2 adders connected to bits in the register. In Chapter 6, we will discuss this in much more detail as a finite-state machine having 2^m states, and memory order m. The channel constraint length of the encoder is defined as $n_E = (m + 1)n$ channel bits, since a given information symbol possibly influences this number of consecutive channel symbols. This parameter bears a rough correspondence to the block length of a block code.

The most efficient decoder is implemented in the form of the Viterbi algorithm, but for now we assume that ML decoding is accomplished in any manner. For such an encoding system, the message error probability of an arbitrarily long message sequence must approach 1 for any reasonable channel (something bad will happen if we wait long enough!), so the proper figure of merit is the decoded symbol or bit error probability.

Viterbi [12, 13] showed[4] that for the ensemble of convolutional codes of rate R and channel constraint length n_E

$$\overline{P_b} < c_R 2^{-n_E E_t(R)}, \tag{4.4.37}$$

where c_R is a constant dependent on R but not n_E, and where $E_t(R)$ is a random-coding exponent for trellis codes. The functional form of $E_t(R)$ is given by

$$E_t(R) = \begin{cases} R_0, & R < R_0, \\ E_0(\rho^*, P), & R_0 \le R < C, \end{cases} \tag{4.4.38}$$

and ρ^* is the solution to

$$\rho^* R = E_0(\rho^*, P). \tag{4.4.39}$$

(Notice again the dual role of R_0 in assessing the performance of the ensemble of convolutional codes, at least for low rates.)

The graphical interpretation of $E_t(R)$ for rates in excess of R_0 is shown in Figure 4.4.7. It has also been shown that $E_t(R)$ is the best possible exponent for rates near C.

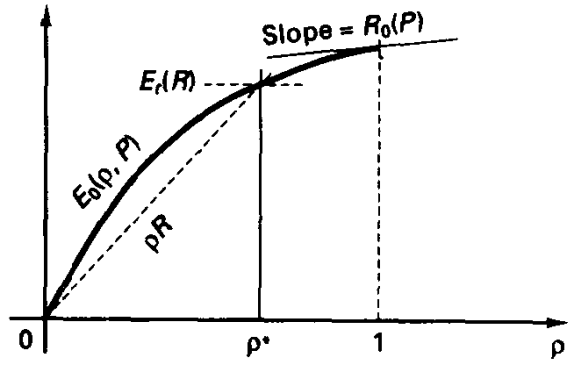

Figure 4.4.7 Determining $E_t(R)$ for $R > R_0$.

[4]The argument was strictly for time-varying trellis codes, but that is not an important issue here.

This ensemble coding exponent is sketched in Figure 4.4.8 for a typical DMC, where it is seen that the trellis coding exponent dominates the block coding exponent $E(R)$ obtained earlier. (Expurgated bounds also show a relative preference for convolutional codes [12].) This fact, coupled with the practical issue that maximum likelihood decoding is more convenient with trellis codes than with block codes, has prompted enormous interest in trellis codes in the last 20 years. Of course, there is some danger in comparing coding techniques merely based on such ensemble exponents. First, we have compared the classes so that block length n is equated with channel constraint length n_E. This leaves aside the relative decoding complexity and delay comparisons. Furthermore, there is a potentially large coefficient c_R in the bound for trellis codes. Finally, these are after all ensemble bounds, and best codes may not mimic this finding.

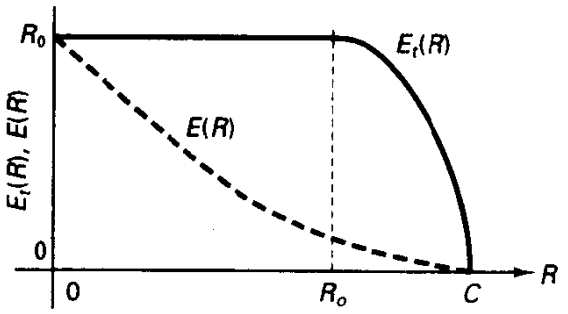

Figure 4.4.8 Error exponent for ensemble of time-varying trellis codes.

4.5 IMPLICATIONS OF R_0 AND C FOR SIGNALING ON AWGN CHANNELS

In the previous sections we established the significance of the parameters R_0 and C for general memoryless channels, and we now study implications for efficient system design and for required system resources that derive from these parameters. Specifically, we wish to determine how a channel encoder can optimally utilize a given modulation/demodulation technique, under power and/or bandwidth constraints, and what the potential gains are in system efficiency over uncoded transmission.

In this section we focus on the AWGN channel environment, treating a variety of modulation formats and detection scenarios. The channel is treated as a nondistorting, fixed-gain, white Gaussian noise channel. At the input to the demodulator, the energy available per uncoded information bit is E_b joules, and the two-sided noise power density is $N_0/2$ W/Hz. The demodulator may or may not make binary decisions on each code symbol.

We will first examine coding approaches of the form shown in Figure 4.5.1, which involve a coding process producing *binary* codewords with rate $R \leq 1$ bits per channel symbol and a binary modulation process.

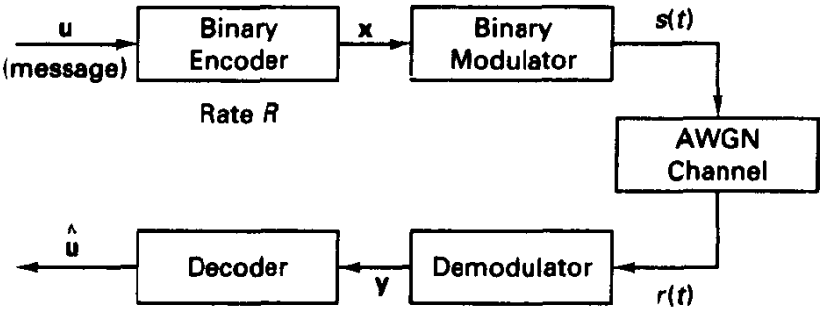

Figure 4.5.1 Binary coding framework.

4.5.1 R_0 and C Considerations for Binary Signaling, AWGN Channel, and Hard Decisions

When binary decisions are formed for each code symbol sent through the AWGN channel, the combination of modulator, channel, and demodulator becomes a simple BSC, with crossover probability ϵ determined by the modulation and detection format, as well as the ratio of energy per code symbol-to-noise density, E_s/N_0. This quantity is related to E_b/N_0 through the actual code rate by

$$\frac{E_s}{N_0} = R\frac{E_b}{N_0}, \tag{4.5.1}$$

since the energy associated with a given information bit is shared among code bits. The four cases of primary interest and their corresponding error probabilities are the following:

$$\epsilon = Q\left[\left(\frac{2E_s}{N_0}\right)^{1/2}\right] \quad \text{coherent PSK (antipodal)} \tag{4.5.2a}$$

$$\epsilon = Q\left[\left(\frac{E_s}{N_0}\right)^{1/2}\right] \quad \text{coherent FSK (orthogonal)} \tag{4.5.2b}$$

$$\epsilon = \frac{1}{2}e^{-E_s/N_0} \quad \text{DPSK} \tag{4.5.2c}$$

$$\epsilon = \frac{1}{2}e^{-E_s/2N_0} \quad \text{noncoherent FSK (orthogonal)} \tag{4.5.2d}$$

Notice that each error probability is an implicit function of code rate R for a given E_b/N_0, through (4.5.1). Given a power constraint and a fixed information rate constraint in bits per second, the optimization of R becomes an interesting design question.

For all BSCs, we have from Section 4.3 that

$$R_0 = 1 - \log\left[1 + (4\epsilon(1-\epsilon))^{1/2}\right], \tag{4.5.3}$$

which may be readily evaluated for these four cases as a function of E_s/N_0 and is shown in Figure 4.5.2. We note that all techniques have the same asymptote of 1 bit per symbol, as will any binary signal set. As indicated in (4.5.2), a given R_0 is obtained by

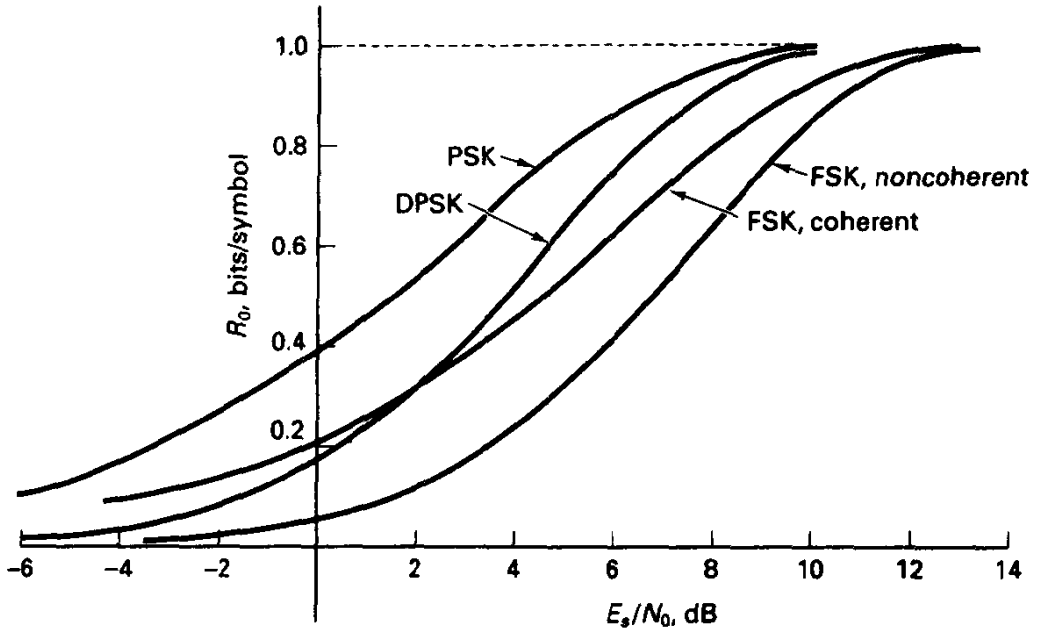

Figure 4.5.2 R_0 for binary modulation, hard-decision demodulation.

coherent antipodal systems with 3 dB less SNR than with coherent orthogonal; also, a 3-dB advantage exists for DPSK over noncoherent FSK.

For coherent PSK and noncoherent FSK, the channel capacities for the induced BSCs are given by (see Section 2.7)

$$C = 1 + \epsilon \log \epsilon + (1 - \epsilon) \log (1 - \epsilon). \tag{4.5.4}$$

where ϵ is given by (4.5.2a) or (4.5.2d). Comparison of these capacities with the corresponding R_0 will show that over much of the low-rate region, say at less than 0.5 bit per channel symbol, the capacity limit is about 3 dB below the R_0 limit in terms of E_s/N_0; that is, to achieve a given R_0 requires about 3 dB greater E_s/N_0, or at fixed E_s/N_0, the channel capacity is about twice the R_0 parameter. This occurs on any "very noisy channel," introduced in Example 4.10, when each channel use supplies low mutual information.

To discern the implications for modulation and coding design, we reason that provided $R < R_0$ (or $R < C$) arbitrarily small error probability may be achieved by increasing the encoder memory, represented by block length, or constraint length for trellis codes. For a given modulation/demodulation strategy, the primary questions are:

1. What is the required E_b/N_0 implied by the R_0 limit (or by C), if we code with rate R?
2. What code rate R should be adopted, if spectrum constraints are not present?

To address these we first equate $R = R_0$, which in turn is a function of RE_b/N_0 through (4.5.2) and (4.5.3); that is, we set

$$R = R_0 = f\left(R\frac{E_b}{N_0}\right) \tag{4.5.5}$$

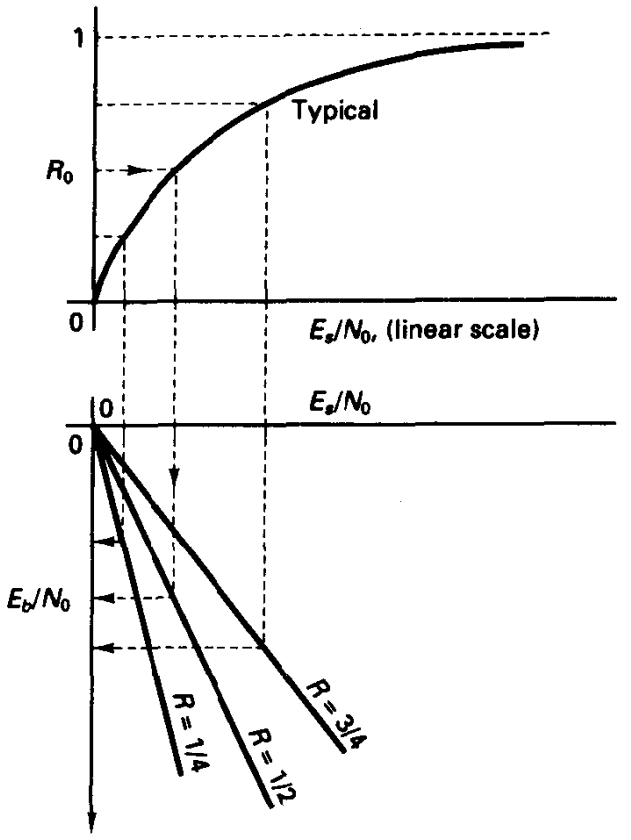

Figure 4.5.3 Nomograph for determing minimum E_b/N_0 as defined by $R < R_0$ for varying R.

and find the solution for E_b/N_0 for different code rates in the range $0 < R < 1$. The nomograph of Figure 4.5.3 illustrates a graphical solution; we pick some trial rate R, find the minimum E_s/N_0 producing $R_0 \geq R$, and then convert this to the required E_b/N_0 through the linear relation of (4.5.1). The locus of such solutions then provides guidance to the communication engineer for proper choice of code rate R and to the communication efficiency that may be expected, especially the relative efficiency of different options. We will illustrate the solution for the case of DPSK.

Example 4.11 Finding E_b/N_0 Lower Bounds for Coded DPSK, Binary Decisions

Setting $R = R_0$ in (4.5.3), we obtain

$$4\epsilon(1-\epsilon)^{1/2} = 2^{1-R} - 1 \tag{4.5.6}$$

or

$$\epsilon(1-\epsilon) = 2^{-2R} - 2^{-R} + 2^{-2}. \tag{4.5.7}$$

Defining x as the right-hand side of (4.5.7), we then seek the solution to the quadratic equation

$$\epsilon^2 - \epsilon + x = 0, \tag{4.5.8}$$

which is $\epsilon^* = [1 - (1-4x)^{1/2}]/2$. (We discard solutions with crossover probability larger than $\frac{1}{2}$.) Having found ϵ^*, we use the DPSK error probability expression (4.5.2c)

$$\epsilon^* = \frac{1}{2}e^{-RE_b/N_0} \tag{4.5.9}$$

and achieve the result that the minimum SNR for operating at $R = R_0$, is

$$E_b/N_{0_{\min}} = \frac{-\log_e 2\epsilon^*}{R}. \tag{4.5.10}$$

For example, if $R = 0.5$, $\epsilon^* = 0.0449$ and $E_b/N_{0_{\min}} = 4.8$, or 6.8 dB.

This procedure may be repeated using channel capacity as the ultimate limit on communication rate, giving a (smaller) lower bound on E_b/N_0 (Exercise 4.5.3). We should also note that the R_0 and channel capacity calculations assumed a memoryless channel model with BSC parameter ϵ, which DPSK technically does not supply, due to the dependence between successive decisions. The memoryless condition can be achieved in practice by a small amount of code symbol interleaving and deinterleaving, a topic to be discussed in Chapter 5.

In Figure 4.5.4 we show the minimum E_b/N_0 loci implied by the R_0 limit for the four previously given modulation types. We interpret the curves as follows: if the minimum E_b/N_0 is not maintained at a specific code rate R, then the R_0 of the modulator/channel/demodulator is insufficient to keep $R_0 > R$.

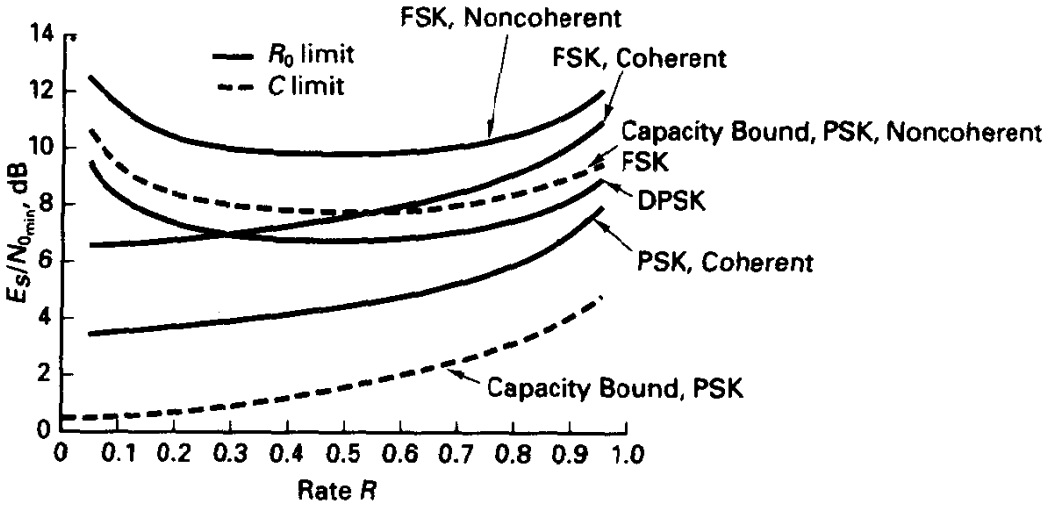

Figure 4.5.4 Required E_b/N_0 versus R for binary transmission, AWGN, hard decisions.

We observe several interesting features from the curves of Figure 4.5.4. First, note that the two coherently detected cases allow steadily smaller E_b/N_0 with decreasing code rate, or with increasing redundancy. Coherent PSK is everywhere 3 dB better than coherent detection of orthogonal signals, as expected. On the other hand, the noncoherently detected cases show a curious degradation at low rates and a broad optimum-rate region from about 0.3 to 0.7. Here again DPSK is 3 dB superior to noncoherent orthogonal detection. The reason for this low-rate degradation is quite subtle. Heuristically, the effect is due to the following facts:

1. As code rate lowers, the energy per symbol drops.
2. Noncoherent schemes are relatively poor in the low SNR region compared to the coherent counterpart.

In essence, the increased redundancy and increased intracodeword distance available with low rate codes are unable to overcome the increasingly poor quality of the binary decisions as R is lowered, and energy per symbol thereby decreases under a power constraint. For coherent detection, on the other hand, we always gain in efficiency when we decrease rate, although not substantially for rates below $R = \frac{1}{3}$. In graphical terms, the noncoherent R_0 curves are not convex, which admits an optimum rate solution, rather than a steadily declining energy requirement with rate. This is pursued in Exercise 4.5.4. We will see this same dichotomy between coherent and noncoherent performance again for unquantized and soft-decision demodulators.

Figure 4.5.4 also shows the limits imposed by keeping rate below the *capacity* limit in the case of coherent PSK and noncoherent FSK; this, of course, allows even smaller E_b/N_0 than R_0 bounds indicate. For low rates the improvement in the coherent antipodal case is 3 dB, due to the result that $R_0 \approx C/2$ in the low-rate region (again, the very noisy channel regime). For very low binary coding rates or large bandwidth-to-bit rate ratio, the capacity limit for antipodal signaling with *coherent hard-decision detection* is $E_b/N_{0_{\min}} = 0.6$ dB, which is roughly 2 dB above the usual Shannon capacity limit $E_b/N_0 > -1.6$ dB for the power-constrained Gaussian channel with infinite bandwidth and unquantized reception. This deficiency is not due to the use of binary inputs, rather than a more Gaussian-like signal, but purely due to hard-decision demodulation. Note also that the capacity limit for the noncoherent orthogonal case exhibits a broad minimum around $R = 0.5$, so this effect is not merely some artifact attached to R_0, but is endemic to noncoherent detection.

Comparison of binary coding with noncoherent detection of orthogonal signaling against binary coding with antipodal signaling and coherent detection, say at $R = 0.5$, gives a theoretical margin to the latter of about 5.3 dB using R_0 comparisons and an even larger margin if capacity is used as the figure of merit. This difference in coding potential is even larger than the difference in efficiency of uncoded signaling, say at $P_b = 10^{-5}$, which is about 4 dB as discussed in Chapter 3. This raises the question of whether noncoherent schemes, when combined with coding, are viable at all. On the AWGN channel with *binary signaling* there is a definite penalty as just discussed. When M-ary modulation is utilized, the margin shrinks quite a lot, as we will see in the next section, and for other channels such as fading and interference channels, the performance difference also becomes quite small, especially given the added difficulty of maintaining a phase reference for coherent detection in such environments.

4.5.2 Binary Signaling, Unquantized Demodulation

We now treat the same class of binary coding/modulation techniques, except that we assume the demodulator passes sufficient statistics to the decoder for each code symbol interval to support ML decoding. In the case of antipodal signals, the demodulator provides a single real-valued matched filter or correlator output, which is a Gaussian random variable. For noncoherent detection of orthogonal signals the decoder is supplied the two envelope-detected measurements of the 0 and 1 channels. In the case of DPSK, we assume that the vector dot product of consecutive phasor measurements in the DPSK receiver is supplied. (In the DPSK case, the vector dot product is techni-

cally not a sufficient statistic for decoding, and interleaving is required to render DPSK modulation/demodulation a memoryless channel.)

For the antipodal case, we have from (4.3.39) that R_0 is given by

$$R_0 = 1 - \log_2(1 + e^{-E_s/N_0}), \tag{4.5.11}$$

and for coherent orthogonal signaling with coherent detection

$$R_0 = 1 - \log_2(1 + e^{-E_s/2N_0}). \tag{4.5.12}$$

These are shown in Figure 4.5.5.

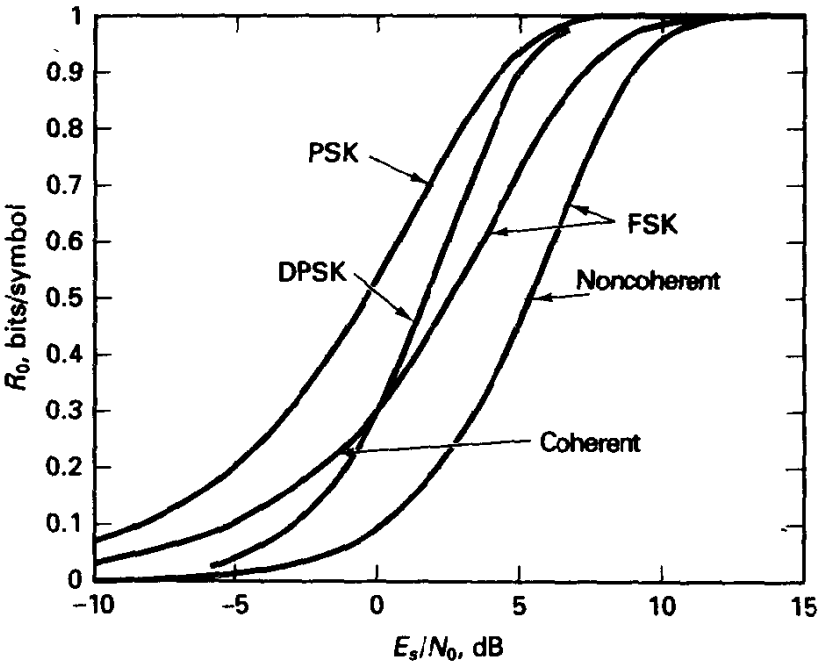

Figure 4.5.5 R_0 for binary modulation, AWGN, unquantized demodulation.

For operation at the antipodal R_0 limit, we choose $0 < R = R_0 < 1$ and find the solution to

$$R = 1 - \log_2(1 + e^{-RE_b/N_0}), \tag{4.5.13}$$

which is

$$\frac{E_b}{N_{0_{\min}}} = \frac{-\log_e(2^{1-R} - 1)}{R} \quad \text{(antipodal)} \tag{4.5.14}$$

and a 3 dB larger value for the orthogonal case. These lower bounds are plotted in Figure 4.5.6 as a function of R. We see that antipodal signaling monotonically approaches, as $R \to 0$, or as bandwidth expansion becomes large, $E_b/N_{0_{\min}} = 2\log_e 2 = 1.4$ dB, exactly 3 dB larger than the limit implied by capacity, another very noisy channel corollary. Furthermore, the unquantized performance limit implied by R_0 calculations is roughly 2 dB better than the binary-quantized limit for all rates of interest. (Compare

Figures 4.5.4 and 4.5.6.) This is an oft-quoted magic number—the information-theoretic penalty for making hard decisions on the AWGN channel with antipodal signaling is 2 dB. Experience with typical coding schemes typically confirms this difference, as we will see in Chapter 5.

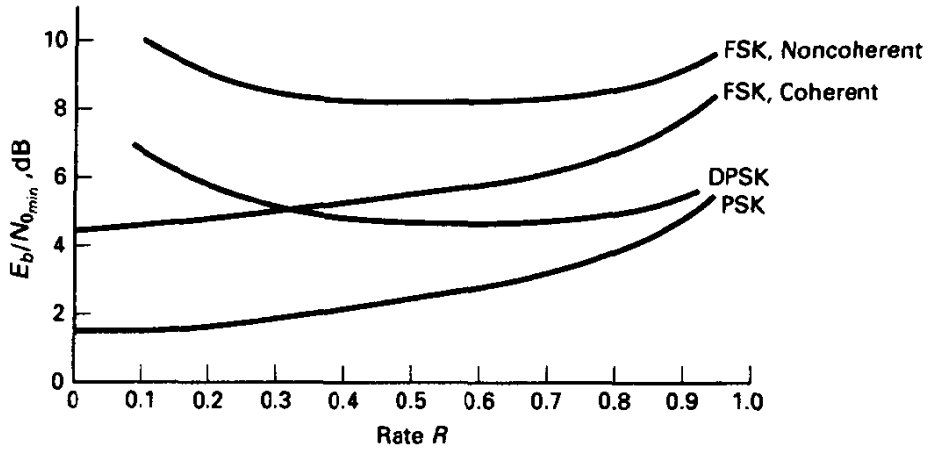

Figure 4.5.6 Minimal E_b/N_0 as defined by R_0, AWGN, unquantized demodulation.

For *unquantized noncoherent detection* of orthogonal signals, the decoder is provided the pair of real values from the two orthogonal channels. One of these is Rayleigh distributed, while the other is Rician, and the variables are independent. The p.d.f.'s necessary for computing R_0 via (4.3.32c) are found in (2.2.19) and (2.2.21). Doing this computation gives

$$\begin{aligned} R_0 &= 1 - \log\left(1 + \int\int f(y_0, y_1|s_0)^{1/2} f(y_0, y_1|s_1)^{1/2}\, dy_0\, dy_1\right) \\ &= 1 - \log\left(1 + \left[e^{-\mu^2/4\sigma^2} \int \frac{y}{\sigma^2} e^{-y^2/2\sigma^2} I_0^{1/2}\left(\frac{\mu y}{\sigma^2}\right) dy\right]^2\right). \end{aligned} \tag{4.5.15}$$

This expression must be evaluated numerically, using $\mu = E_s^{1/2}$ and $\sigma^2 = N_0/2$, but as a check we observe that when μ^2/σ^2 becomes small (small SNR) the integral becomes that of a Rayleigh p.d.f.; hence R_0 approaches zero. In Figure 4.5.5, we show the resulting R_0 versus E_s/N_0, together with the coherently detected counterpart and note the more rapid drop in R_0 for the noncoherent case at low SNR (this is somewhat obscured by the logarithmic presentation). It has been shown analytically by Jordan [14] that the noncoherent R_0 falls as the second power of E_s/N_0 in the low SNR region, rather than a first-power dependence for the coherent case that a series expansion of (4.5.12) will show.

Similarly for DPSK, we numerically evaluate R_0 using (4.3.32c) and the p.d.f. for the vector inner product of two complex Gaussian random variables [15]. Substitution into the expression for R_0 gives

$$R_0 = 1 - \log\left[1 + \frac{1}{2} e^{-E_s/N_0} g\left(\frac{E_s}{N_0}\right)\right], \tag{4.5.16}$$

where

$$g(z) = \left[e^{-z} \sum_{m=0}^{\infty} \frac{z^m}{m!} \right]^{\frac{1}{2}} \int_0^{\infty} \left[e^{-y} \sum_{n=0}^{m} \frac{y^n}{n!} \right]^{1/2} dy. \tag{4.5.17}$$

This expression can be evaluated using numerical integration and is also shown in Figure 4.5.5 alongside the coherent antipodal case; again note the difference in behavior as SNR becomes small.

Following the procedure described earlier, we can numerically solve for the minimum E_b/N_0 that will keep $R_0 > R$ at various rates for DPSK and noncoherent orthogonal cases. The results are shown in Figure 4.5.6, again with the coherently detected counterparts. We observe behavior similar to that in the hard-quantized case: the noncoherent schemes suffer at low rates, and for these cases an optimum coding rate exists. Also, in the noncoherent cases, notice that the unquantized case is superior to the binary-quantized case, as it should be, but by a lesser amount than in the coherent detection mode.

We will not discuss at length the capacity implications for unquantized transmission, for they basically tell a similar story. To indicate the methodology, we will consider the antipodal case. We recall that the channel capacity for these binary modulation techniques under an energy constraint is given by the mixed expression for mutual information, under the adoption of equiprobable inputs:

$$C = \int_{\mathbf{y}} \sum_{i=0}^{1} P(\mathbf{s}_i) f(\mathbf{y}|\mathbf{s}_i) \log \left[\frac{f(\mathbf{y}|\mathbf{s}_i)}{f(\mathbf{y})} \right] d\mathbf{y}, \tag{4.5.18}$$

where $f(\mathbf{y}|\mathbf{s}_i)$ is a one-dimensional Gaussian p.d.f. in the case of antipodal signals and where $f(\mathbf{y}|\mathbf{s}_i)$ is a two-dimensional Gaussian p.d.f. in the case of orthogonal signals. Centering of the p.d.f.'s is at the signal in the conditioning statement, and the variance in each signal-space coordinate is $N_0/2$. In Figure 4.5.7 we plot the antipodal (unquantized) capacity versus E_s/N_0, along with the capacity for the additive Gaussian noise channel without a binary input constraint. Note that for low SNR per code symbol the two capacities are substantially equivalent, so binary transmission induces no loss of performance. For larger SNR, however, we must resort to nonbinary modulation, for example, QAM, to efficiently utilize the resources. We will return to this shortly.

4.5.3 Binary Signaling with Soft-quantized Demodulation

We have just seen that binary (hard-decision) demodulation can be very detrimental to the potential performance of coded communication systems, typically manifesting itself as a roughly 2-dB loss in efficiency relative to unquantized demodulation on the AWGN channel,[5] at least as predicted by R_0 theory. If we recognize that in modern decoders the computations will be performed with finite-precision calculations and that we would ordinarily wish to minimize the associated complexity, the degree of acceptable quantization becomes of interest. Decoding with finely quantized receiver output data is referred to in the literature as ***soft-decision decoding.***

[5]The penalty may be even more profound for other channels, notably fading channels, as discussed in Section 4.6.

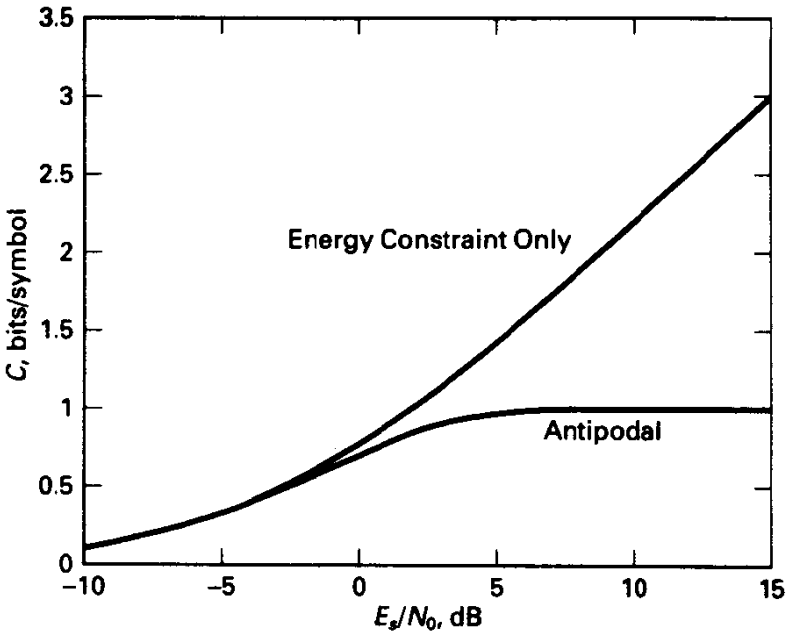

Figure 4.5.7 Capacity for antipodal signals and energy-constrained channel.

We will first consider the case of antipodal signaling with coherent detection, with a $Q = 2^b$-level quantizer operating on the demodulator output. For simplicity, the quantizer is a uniform quantizer with step size of Δ. The quantized channel becomes a 2-input, Q-output discrete memoryless channel, for which earlier expressions can be used to evaluate R_0, or capacity as well, once the transition probabilities are specified for the DMC. The latter are a function of E_s/N_0 and the quantizer scale factor.

To operate effectively over a range of SNRs, as signal and/or noise levels vary, we need to find a procedure for scaling the range of the quantizer. For multiamplitude constellations, it is necessary to scale the quantizer according to the received signal level, rather than according to the noise level, although either approach works well in the antipodal case if proper scaling is adopted. Thus, we choose $\Delta = cE_s^{1/2}$, or more precisely c times the mean of the demodulator output, where c is a scale factor to be optimized. Our experience is that for 4-level quantization, the proper scale factor is $c \approx 0.6$, while for 8-level quantization, $c \approx 0.3$ gives good performance. These judgments depend slightly on the SNR assumed, but are appropriate for SNRs giving R_0 near 0.5 bit/symbol. Lee [16] provides necessary conditions for the design of R_0-optimal quantizers for decoding, which could be applied to this problem as well.

Figure 4.5.8 shows the R_0 curves for 2-, 4-, and 8-level quantization of binary PSK, as well as the unquantized case. The quantizer zone probabilities at $E_s/N_0 = 0$ dB are given in Table 4.1. Of course, to properly utilize this soft-decision information, the zone probabilities must be known fairly accurately to compute metrics; otherwise mismatch exists.

Observe that 8-level (3-bit) quantizing provides essentially the same efficiency as unquantized demodulation, losing perhaps 0.25 dB; 4-bit quantization could be claimed to provide completely adequate discretization of the receiver outputs. In the other direction,

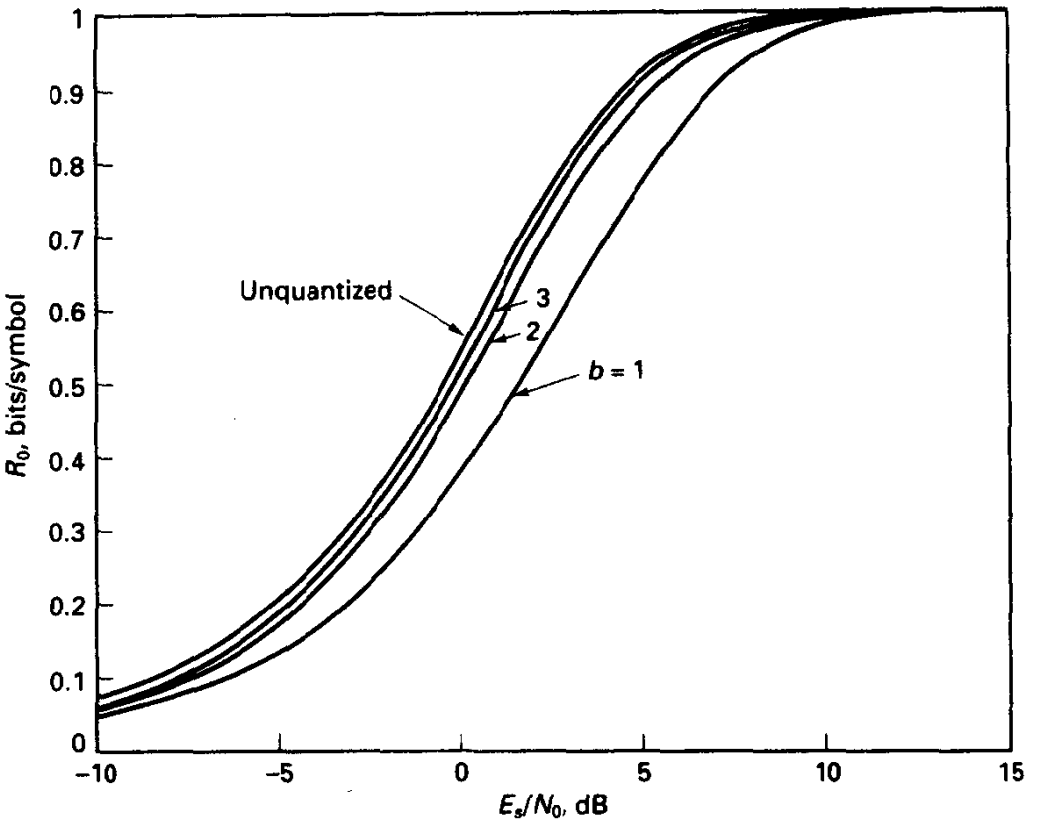

Figure 4.5.8 R_0 for quantized antipodal signaling, AWGN.

TABLE 4.1. QUANTIZER CONDITIONAL PROBABILITIES AT $E_s/N_0 = 0$ dB FOR $Q = 8$, 4, AND 2. STEP SIZE IS $0.3E_s^{1/2}$ FOR $Q = 8$.

0.556	0.158	0.125	0.082	0.046	0.021	0.008	0.004
0.714	0.207	0.067	0.012				
0.921	0.079						

Zones are indexed from closest to farthest signal hypothesis.

the loss with 4-level quantization is typically only 0.7 dB, and, as we have seen earlier, the penalty for hard decisions is about 2 dB. We should be careful in generalizing this, however, to other signal schemes; 16-QAM transmission will require roughly 5 bits per coordinate axis to achieve near-unquantized performance.

4.5.4 Summary for Binary Transmission, AWGN Channel

Before closing the discussion of coding potential for binary schemes, it is well to summarize what coding offers. Recall that for uncoded transmission the best binary schemes were antipodal signaling if coherent detection is allowed and differential PSK when non-

coherent detection is required. To achieve an error probability of 10^{-5} requires that $E_b/N_0 \approx 10$ dB for both, with coherent antipodal slightly superior. By employing channel coding of these binary modulation schemes, on the other hand, we can *potentially* operate with E_b/N_0 approaching -1.6 dB in the coherent case. (R_0 calculations suggest that $E_b/N_0 \approx 2$ dB is technologically feasible.) Both cases admit arbitrarily low error probability, as opposed to, say, 10^{-5} bit error probability, so it is difficult to compare directly the coded and uncoded options. Nonetheless, there is an apparent 8- to 10-dB energy savings offered through coding, which has prompted the enormous interest in coding since the original realization of this fact. If we repeat the calculation for the constraint of binary orthogonal signaling with noncoherent detection, the required E_b/N_0 to attain $P_b = 10^{-5}$ without channel coding is 13.4 dB, from Chapter 3. Figure 4.5.6, an R_0 assessment, illustrates that arbitrarily reliable communication is possible with this modulator/demodulator scheme with $E_b/N_0 \approx 8.5$ dB, a savings of about 5 dB.

To reap this benefit, we must be prepared to expand the bandwidth of our transmitted signal and to accept potentially large complexity in encoding and decoding. Regarding bandwidth, we can define the bandwidth expansion ratio, relative to uncoded transmission with the same modulator, as $1/R$, since the number of binary digits per unit time is increased by the encoder. Thus, the designer should realize that spectrum economy suffers dramatically in the low-rate region. Given this situation, it is well to realize from Figure 4.5.6 that almost all the available energy efficiency is accrued by using rate $\frac{1}{3}$ codes. For noncoherent detection, it is best not to use low-rate codes anyway.

4.5.5 R_0 and C for Coding with M-ary Modulation, AWGN Channels

We continue the analysis of various communication options by extending the discussion to M-ary modulation, wherein the coding schemes now produce M-ary code symbols. We continue to specify the code rate R in information bits per code symbol; under this definition the encoder rate can be greater than unity. As in the previous section we assume that the channel is nonfading, nondistorting, and corrupted only by AWGN. We will focus on the unquantized demodulation case; hard-decision decoding is handled readily using the DMC methodology we have outlined.

We shall begin with coherent detection. The principal cases of interest are M-ary orthogonal (and its biorthogonal and simplex relatives) and M-ary PSK/QAM schemes in two dimensions. The former exhibit very good energy efficiency, at the expense of bandwidth, while the QAM schemes are more bandwidth efficient in exchange for a larger SNR requirement.

M-ary Orthogonal, Biorthogonal, and Simplex Designs, Coherent Detection, Unquantized Demodulation

For coherent detection of M orthogonal signals, the basis-function form of the demodulator developed in Section 3.3 produces M random variables, all Gaussian and independent. One demodulator output statistic has mean $\mu = E_s^{1/2}$, and the remainder have zero mean. The variance in each channel is $\sigma^2 = N_0/2$. The ML decoder will form a symbol metric

for testing code symbol j that utilizes only that dimension of the demodulator output:

$$\lambda(x_j, \mathbf{y}_j) = -(y_{x_j} - E_s^{1/2})^2, \tag{4.5.19}$$

which is equivalent to simply using the real-valued output of one demodulator channel as a metric.

In (4.3.38), we presented an R_0 expression for any signal constellation in the presence of AWGN and unquantized demodulation. For the orthogonal constellation, all distances between distinct signals are $d_{ij} = (2E_s)^{1/2}$, and substitution into (4.3.38) gives

$$\boxed{R_0 = \log M - \log\left[1 + (M-1)e^{-E_s/2N_0}\right] \quad \text{bits/symbol, orthogonal}} \tag{4.5.20}$$

Figure 4.5.9 depicts the value of R_0 for $M = 2$, 8, and 32 versus E_s/N_0. In each case, R_0 approaches $\log_2 M$ bits at high SNR.

To link this to communication efficiency and to compare different options, we note that $E_s = E_b R$ since each code symbol, by definition, conveys R information bits for coded transmission with code rate R.[6] We again ask, "What is the smallest E_b/N_0

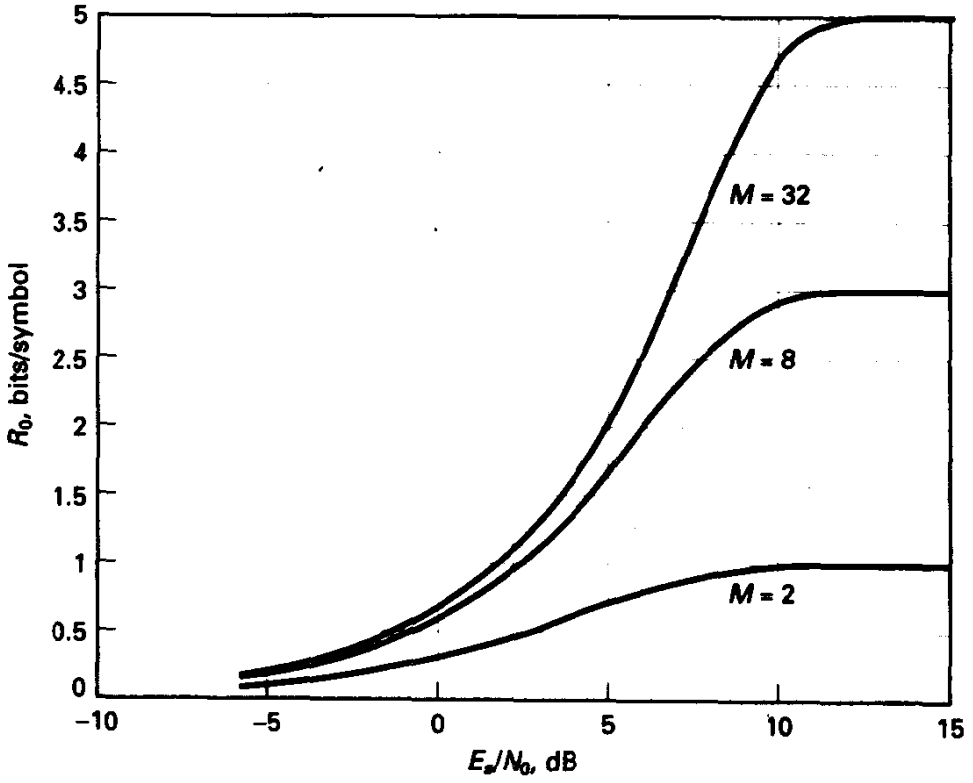

Figure 4.5.9 R_0 for M orthogonal signals, AWGN, unquantized coherent detection.

[6]Notice R may exceed 1 in these cases.

that will maintain R_0 greater than R?" To do so, we find the solution to

$$R = R_0 = \log M - \log\left[1 + (M-1)e^{-RE_b/N_0}\right] \tag{4.5.21}$$

for R in the range $0 < R < \log M$. The results are shown in Figure 4.5.10, after normalizing both R_0 and R by $\log M$ for plotting convenience. Notice that the theoretical minimum E_b/N_0 implied by R_0 limits decreases with decreasing rate R and with increasing alphabet size M. One penalty of this energy improvement is increased bandwidth occupancy. If we measure the relative spectral efficiency by the number of information bits per signal-space dimension, then the spectral efficiency becomes

$$\eta = R\frac{\log_2 M}{M} \quad \text{bits per dimension} \tag{4.5.22}$$

This exhibits the bandwidth penalty we have earlier seen for uncoded M-ary orthogonal signaling, exacerbated here by channel coding. To indicate the bandwidth efficiency of various alternatives, values of η^{-1} are marked along the R_0 curves. This makes more explicit the bandwidth expansion penalty paid if we wish to extract the ultimate energy efficiency from the system. Bandwidth and complexity constraints will normally prevail well before the asymptotic gain is encountered.

Using the expression for R_0 in (4.3.38), which involves only signal-space coordinates, it is simple to obtain the following expressions for the M-ary biorthogonal and

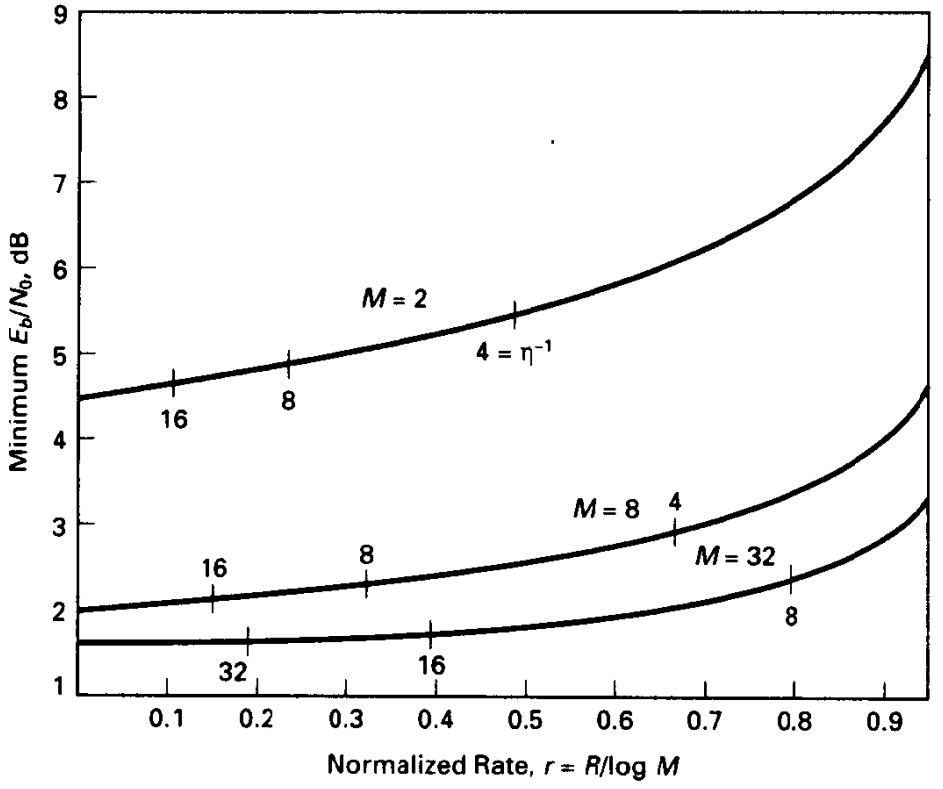

Figure 4.5.10 Minimum E_b/N_0 for M orthogonal signals, coherent detection.

simplex cases:

$$R_0 = \log M - \log\left[1 + (M-1)e^{-(M-1)E_s/MN_0}\right] \quad \text{simplex} \tag{4.5.23}$$

and

$$R_0 = \log M - \log\left[1 + (M-2)e^{-E_s/N_0} + e^{-2E_s/N_0}\right] \quad \text{biorthogonal.} \tag{4.5.24}$$

As these expressions attest, for $M \geq 8$ there are only minor differences in the R_0 values between orthogonal, biorthogonal, and simplex. Of course, the biorthogonal construction requires smaller bandwidth occupancy. Massey [2] has shown that the simplex design is optimal in the sense of maximizing R_0 among all M-ary signal sets having equal E_s/N_0, without regard to dimensionality.

M-PSK and *M*-QAM

Similar calculations are easily formulated for M-ary PSK and QAM constellations, needing only the signal-space coordinates and the intra signal distances to calculate R_0. For PSK, the result of (4.3.38) becomes

$$R_0 = -\log_2\left[\frac{1}{M}\sum_{i=0}^{M-1} e^{-(E_s/N_0)\sin^2(i\pi/M)}\right]. \tag{4.5.25}$$

Figure 4.5.11 presents results for $M = 2$, 4, 8, and 16 as a function of E_s/N_0, from which two conclusions should be drawn. First, for small E_s/N_0, there is no energy benefit in utilizing large PSK signal sets for coding purposes, and this is true of general modulation sets in this regime—binary antipodal signaling is the most suitable choice. As E_s/N_0 increases, however, we can reliably achieve greater throughput (larger R_0) by adopting larger PSK sets.

Second, the potential benefits of coding are evident from such plots. Suppose we wish to send two bits per modulator interval using a two-dimensional constellation. The most natural design is QPSK without coding; that is, each message bit pair is a minimessage. We have seen that to achieve a symbol error probability of, say, $P_s = 10^{-5}$ requires roughly 10 dB in E_b/N_0, or roughly 13 dB in E_s/N_0. R_0 theory would suggest that arbitrarily reliable transmission is possible if we supply a modulator/channel/demodulator with $R_0 > R$. If $R = 2$ bits per interval, Figure 4.5.11 shows that use of 8-PSK modulation can meet this requirement at $E_s/N_0 = 7.5$ dB, representing a potential savings of about 5.5 dB. (If we performed the comparison at $P_s = 10^{-9}$, the gains would have been even larger.) Furthermore, bigger PSK constellations than 8-PSK are apparently of no substantial benefit in achieving 2 bits/symbol throughput.

For small M, PSK constellations are essentially the best in two dimensions. However, as observed in Chapter 3, as M increases, the minimum distance drops rapidly due to the points-on-a-circle constraint, and M-ary QAM constellations are typically more efficient for large M. Figure 4.5.11 shows the R_0 curve for 16-QAM,[7] and 16-QAM is

[7] We have used the equiprobable probability assignment in evaluating R_0 here. Slight improvement is possible with signal probability biased toward the small energy signals, since this reduces average energy slightly.

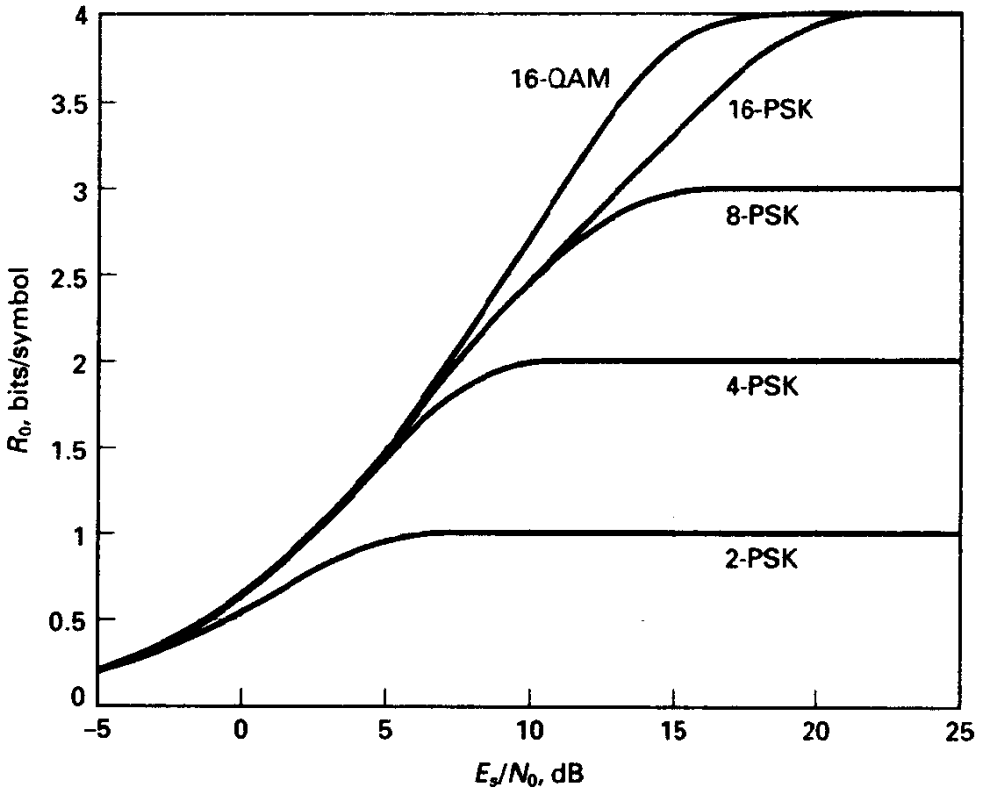

Figure 4.5.11 R_0 for two-dimensional constellations, AWGN channel.

indeed superior to 16-PSK for a given average signal energy. (A peak-energy comparison is more favorable to 16-PSK, however.) 16-QAM would apparently be an efficient modulation scheme for sending $R = 3$ (not 4) bits of information per interval.

We know that channel capacity is the ultimate limit on reliable throughput, although pushing the rate beyond R_0 proves practically difficult, as we will see. Nonetheless, let's reconsider what capacity arguments suggest about coded two-dimensional signaling.

First, recall from Chapter 2 that the channel capacity of the N-dimensional additive Gaussian noise channel, in bits per channel use, is

$$C_N = \frac{N}{2} \log_2 \left(1 + \frac{2E_s}{NN_0} \right), \tag{4.5.26}$$

where E_s is the allowed energy per N-dimensional input vector, and $N_0/2$ is the noise spectral density, equivalent to the noise variance per dimension. In the two-dimensional case this becomes $C = \log(1 + E_s/N_0)$. This maximum mutual information is achieved when the inputs are independent, zero-mean Gaussian variables, each with variance E_s/N joules.

In principle, efficient coding could transpire by building large sets of signal vectors having the preceding prescription. Indeed, random coding is one way to proceed, especially for large block lengths. However, for practical reasons, we wish to form code sequences from sequences of some elementary signals; that is, we wish to build large sets from small modulator constellations. Thus, we ask for the channel capacity for two-

dimensional constellations and in addition prescribe that the channel inputs be selected with equal probability. The channel capacity is given by an extension of (4.5.18):

$$C^* = \int \sum_{i=0}^{M-1} \frac{1}{M} f(\mathbf{y}|\mathbf{s}_i) \log_2 \left[\frac{f(\mathbf{y}|\mathbf{s}_i)}{\sum_k \frac{1}{M} f(\mathbf{y}|\mathbf{s}_k)} \right] d\mathbf{y}. \tag{4.5.18}$$

This capacity is a function of the signal constellation and E_s/N_0. Slightly larger mutual information is available if the inputs are used with unequal probabilities, as the preceding Gaussian distribution would suggest.

In Figure 4.5.12, C^* is shown for certain two-dimensional constellations discussed in Chapter 3. Notice that the various capacity curves saturate at $\log_2 M$ bits per modulator symbol, implied by the fact that there is no possibility of communicating more than $\log_2 M$ bits/symbol reliably with an M-ary signal set. This same behavior was seen for R_0.

The important observation for code design is that to achieve a certain capacity in bits per symbol, say R bits/symbol, it is basically sufficient to code (build codewords) with a good constellation having 2^{R+1} symbols. This was first apparently recognized by Ungerboeck [17] and formed the basis of the folk theorem that constellation expansion by 2 is sufficient. We can argue that in the capacity sense another 1 dB or so savings in E_s/N_0 is available with still bigger constellations with code symbols selected nonequiprobably, but in practice this has yet to show any payoff. (Shaping [18] of constellations can help by altering the p.d.f. on input selection.)

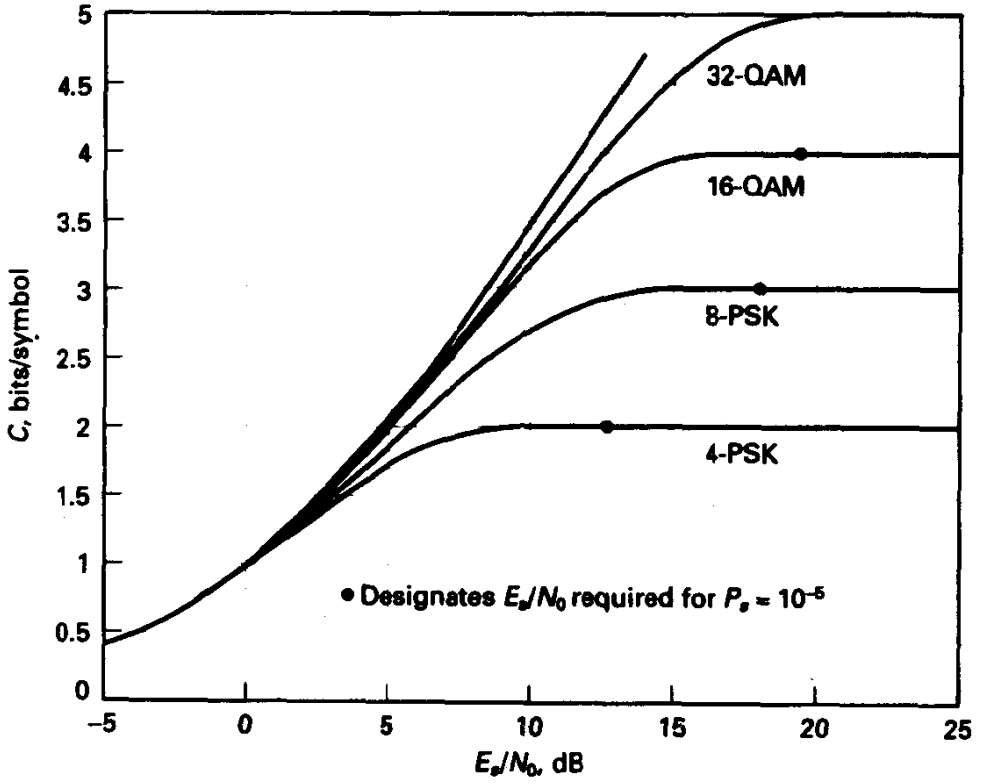

Figure 4.5.12 Capacity for two-dimensional constellations, AWGN channel.

On Figure 4.5.12, we have shown for each constellation the E_s/N_0 necessary to achieve $P_s = 10^{-5}$ and observe, following Ungerboeck, that roughly 8-dB improvement in energy efficiency is potentially available through coding. Specifically, communicating $R = 2$ bits/symbols using an 8-ary constellation can be accomplished in principle with $E_s/N_0 = 4.8$ dB, whereas achieving $P_s = 10^{-5}$ with uncoded QPSK requires $E_b/N_0 = 10$ dB, or $E_s/N_0 = 13$ dB, and the potential saving is $13 - 4.8 = 8.2$ dB. The really important observation is that this can occur *without increase in bandwidth* by adopting an expanded signal constellation to provide redundancy. This is in marked contrast to incorporating redundancy by sending *more* symbols from the same original constellation and thereby increasing the bandwidth for a given fixed information rate. The constructive side of this process of course remains, which we will discuss in Chapters 5 and 6. The rate region is larger for a given E_s/N_0 than that defined by R_0.

R_0 for M-ary Signals, Noncoherent Demodulation

We now take up the case of noncoherent demodulation in coded M-ary systems, pertinent when the demodulator may not be able to attain a stable phase reference with which to perform coherent demodulation. Noncoherent systems typically utilize either orthogonal signaling or a differential phase-shift-keying (DPSK) modem.

Let's examine first the case first of M-ary orthogonal signaling. The demodulator produces for each codeword position $0 \le i \le n-1$ a *vector* of measurements, $\mathbf{y}_i = (y_{i_0}, \ldots, y_{i_{M-1}})$, corresponding to the outputs of the M noncoherent correlators or matched filters. For the AWGN model, one of these variables will be Rician and the remaining variables Rayleigh, with all jointly independent. In symbol-by-symbol detection, we would choose the index of the largest random variable in $\mathbf{y}_i$ as our decision. In sequence transmission, however, we would like the decoder to utilize as much of the demodulator output as necessary for optimal codeword decisions. Given a codeword $\mathbf{x}_i = (x_{i_0}, x_{i_1}, \ldots, x_{i_{n-1}})$, the p.d.f. for the sequence of demodulator output vectors, $\tilde{\mathbf{y}} = (\mathbf{y}_0, \ldots, \mathbf{y}_{n-1})$, may be written as

$$f(\tilde{\mathbf{y}}|\mathbf{x}_i) = \prod_{j=0}^{n-1} f(\mathbf{y}_j|x_{i_j}), \tag{4.5.27a}$$

where

$$f(\mathbf{y}_j|x_{i_j} = k) = \frac{y_{j_k}}{\sigma^2} I_0\left(\frac{\mu y_{j_k}}{\sigma^2}\right) e^{-(y_{j_k}^2+\mu^2)/2\sigma^2} \prod_{m \ne k} \frac{y_{j_m}}{\sigma^2} e^{-y_{j_m}^2/2\sigma^2}. \tag{4.5.27b}$$

This is nothing more than the product of a Rician p.d.f. and $M-1$ Rayleigh p.d.f.'s, with the indexing controlled by the code symbol specified for the codeword under consideration.

If we substitute (4.5.27) into (4.3.32c), use the symmetry of the modulator and channel to realize that an equiprobable assignment maximizes the R_0 expression, and simplify the integrand by recognizing density functions whose integrals are 1, we can determine that

$$R_0 = \log_2 M - \log_2 \left\{ \left[1 + (M-1) e^{-E_s/N_0} \left[\int_0^\infty \frac{y}{\sigma^2} e^{-y^2/2\sigma^2} I_0^{1/2}\left(\frac{\mu y}{\sigma^2}\right) dy \right]^2 \right] \right\}. \tag{4.5.28}$$

where $\sigma^2 = N_0/2$ and $\mu = E_s^{1/2}$. This result was first derived in Jordan [14], citing Cheek and Reiffen. Note that for $M = 2$, we obtain the result of Section 4.5.2. In [14] it was also observed that for small E_s/N_0 with M becoming large,

$$R_0 \approx \frac{1}{4}\left(\frac{E_s}{N_0}\right)^2 \log_2 e \quad \text{bits per channel symbol,} \tag{4.5.29a}$$

whereas for coherent detection the result at low SNR and large M is

$$R_0 \approx \frac{1}{2}\left(\frac{E_s}{N_0}\right) \log_2 e \quad \text{bits per channel symbol.} \tag{4.5.29b}$$

This again demonstrates that noncoherent detection is relatively inefficient in the small SNR regime.

Figure 4.5.13 illustrates R_0 for orthogonal signaling with noncoherent detection for $M = 2$, 8, and 32, computed numerically from (4.5.28); these results should be compared with those of coherent detection (Figure 4.5.9). The implication for coding can be appreciated by finding the minimum energy solution consistent with keeping R_0 greater than a given rate R. We obtain this by setting $R = R_0$ and solving for E_b/N_0. This is shown in Figure 4.5.14 as a function of code rate R, and we see an important departure between the noncoherent and coherent situations. Specifically, noncoherent detection performs best at modest rates, say $R \approx \log_2 M/2$ bits per symbol, and at this optimal code rate R, the difference between coherent and noncoherent performance

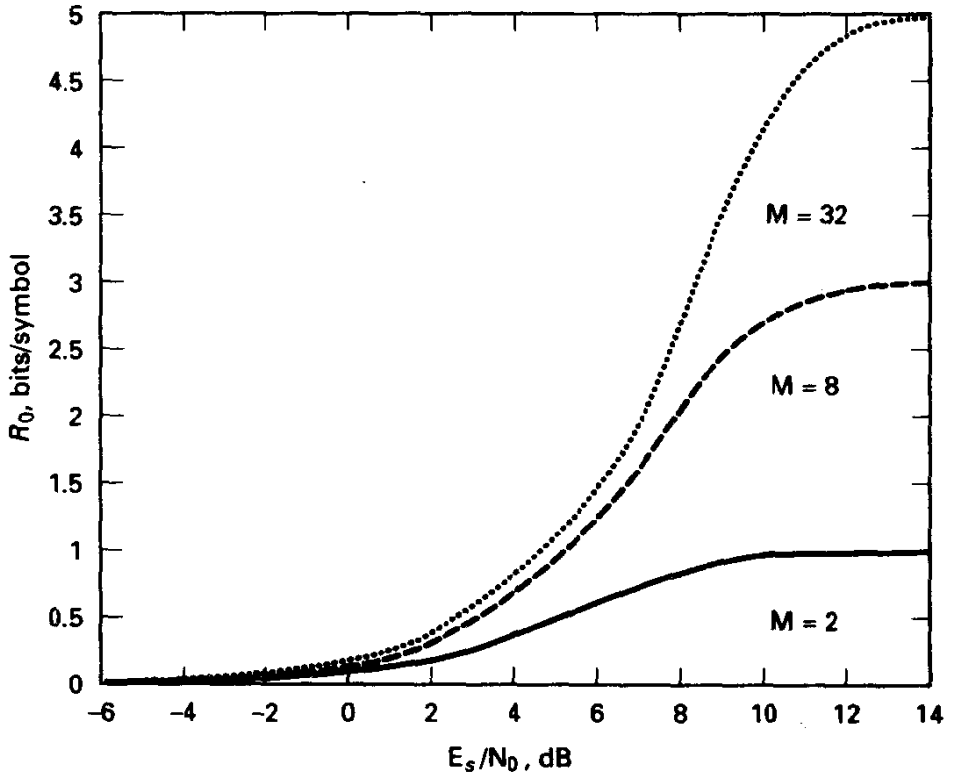

Figure 4.5.13 R_0 for M orthogonal signals, AWGN, noncoherent detection.

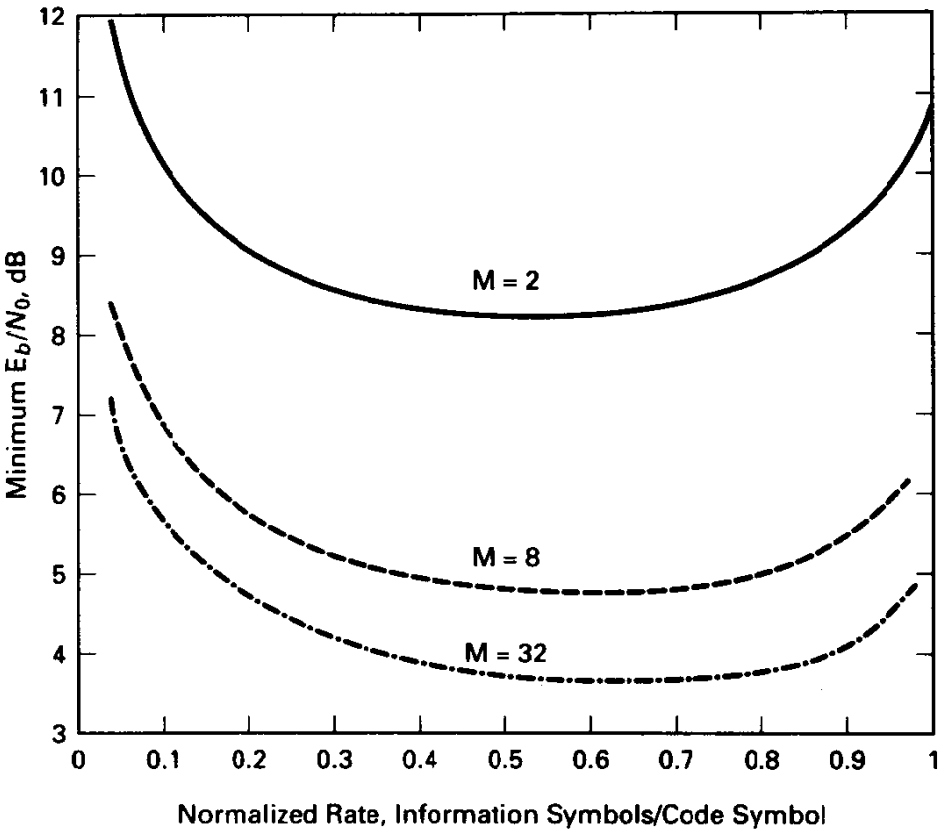

Figure 4.5.14 Minimum E_b/N_0 versus rate, M orthogonal signals, noncoherent detection AWGN.

predicted by R_0 is about 3 dB. The fact that noncoherent detection does not benefit by letting code rate diminish indefinitely on the AWGN channel is sometimes referred to as the ***noncoherent combining penalty.*** We have earlier seen similar effects in the situation of fast-frequency-hopping spread-spectrum modulation (Section 3.8). Notice that the use of large M mitigates this noncoherence penalty somewhat.

DPSK modulation is another noncoherent technique that by itself is rather efficient in use of spectrum for large M, but, as we have seen, the (uncoded) energy performance is roughly 3 dB poorer than PSK for larger M. A technical detail in the analysis of coded DPSK is that successive demodulator outputs are not independent, because two intervals join to form the statistic for a given interval. Thus, we cannot directly apply memoryless channel analysis techniques. One means of handling this mathematical difficulty, and a wise engineering choice as well, is to interleave, or scramble, the sequence at the transmitting end and then reorder symbols after demodulation so that consecutive symbols are essentially independent. We shall say more about interleaving in later chapters. Apart from this, there is a question about what variables the demodulator should supply the decoder for optimal decoding, or what are the sufficient statistics. In uncoded transmission, the demodulator bases its decision on phase differences; it is thereby sensible to pass the analog phase difference to the decoder for each interval. Recently, Bello [19] has shown that the proper data the demodulator should supply for

interval i is

$$l_i = \mathrm{Re}\,[r_i r_{i-1}^*], \tag{4.5.30}$$

where * denotes conjugation. This can be simplified to

$$l_i = A_i A_{i-1} \cos(\theta_i - \theta_{i-1}), \tag{4.5.31}$$

where A_n and θ_n are, respectively the amplitude and phase of the complex number represented by the quadrature channel outputs. Notice that the measurement amplitudes are important in the decoding process, rather than merely the angular difference.

4.6 CAPACITY AND R_0 FOR THE RAYLEIGH FADING CHANNEL

Although Rayleigh fading exacts a very large penalty on energy efficiency for uncoded transmission on the flat-fading Rayleigh fading channel, as demonstrated in Section 3.6, properly designed coded systems can recoup virtually all this loss under certain assumptions. Such channels, and others to be examined in the next section, are especially amenable to channel coding, with coding gains much larger than for the AWGN channel. This potential is foretold by analyzing C and R_0.

The Rayleigh channel exhibits two important distinctions from the AWGN channel and its hard-decision derivatives. First, we have assumed that the fading is slow, relative to a symbol duration, and this will apparently mean that the channel amplitude and phase modification of the signal are strongly dependent over many successive channel transmissions. Thus, as it stands, the assumed fading channel is far from memoryless. Second, the decoder can profit from side information in the form of the actual channel amplitude scale factor, a_j, for the jth channel symbol. In the AWGN case, it was important for the demodulator to have proper internal scaling in cases such as QAM if symbol-by-symbol decisions are intended, but the decoder cannot further benefit by being told the amplitude, because it is merely a constant scale factor in the metric.

A traditional means of addressing the memory of the channel ***interleaving and deinterleaving*** as shown in Figure 4.6.1. The interleaver is inserted between the channel coding operation and the modulator and the deinterleaver between the demodulator and decoder. For now, think of these devices as scramblers that permute the order of symbols sent over the channel in such a fashion that, once descrambled, the action of the channel appears memoryless. Some delay is incurred in this process, which is the major limitation on its practicality, and we will discuss the details in Chapter 5.

Actually, such scrambling does not alter the total information available to the decoder, but merely rearranges it in time. The rationale behind interleaving is that shorter block-length codes can be immunized against the effect of a single bad fading episode; instead, our codeword decision will be predicated on many independent channel states, and a law of large numbers can be exploited. We will discuss at the end of the section an information-theoretic view on interleaving, but for now we will assume that the interleaver is ideal, producing a memoryless channel as seen by the encoder/decoder pair. *If* side information on amplitude is to be supplied the decoder, it is necessary that this information be carried along with the demodulator outputs in the deinterleaving operation.

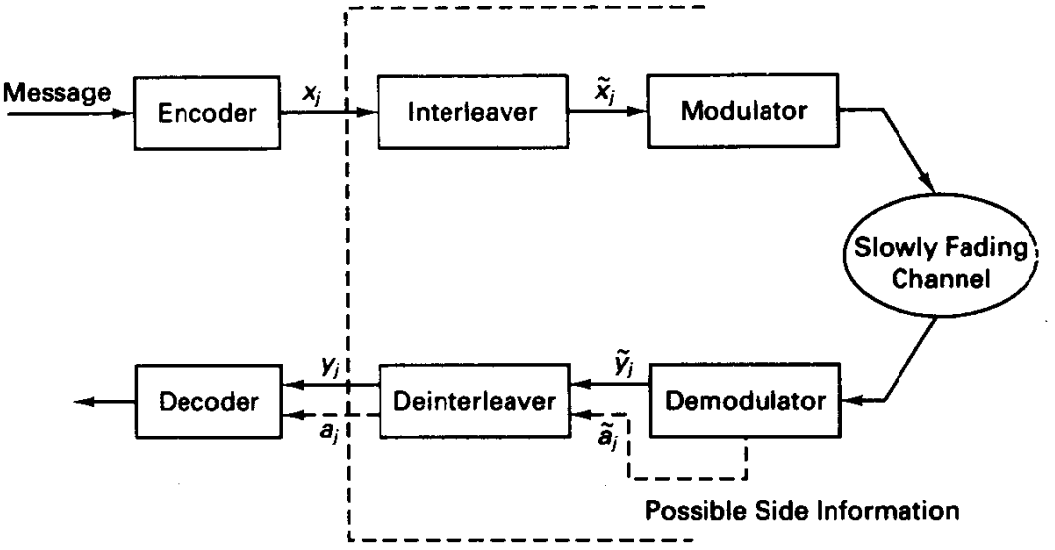

Figure 4.6.1 Generic interleaving technique for slowly fading channel. Encoder/decoder "see" a memoryless channel.

There are multiple variations on the basic model of Figure 4.6.1, including all combinations of the following:

1. Whether the demodulator provides hard or soft decisions
2. Whether side information in the form of channel amplitude is available to the decoder.

Using a memoryless channel model, we can define capacity and R_0 parameters as before. We will begin with binary signaling.

4.6.1 Coding Potential for Binary Signaling on the Rayleigh Channel

Hard Decisions with No Side Information

Suppose that a binary modulator sends one symbol per unit time over a fully interleaved Rayleigh fading channel, and let E_s/N_0 be interpreted as the *mean* symbol energy-to-noise density ratio at the receiver, averaged over the fading distribution. If the demodulator supplies its best estimate of each symbol to the decoder (again referred to as hard decisions), the channel error probability is given by the expressions of Section 3.6. For example, in the case of binary orthogonal signals with noncoherent detection,

$$P_s = \epsilon = \frac{1}{2 + E_s/N_0}. \tag{4.6.1}$$

If no further side information is supplied the decoder, the R_0 expressions developed in earlier sections for the BSC apply:

$$\begin{aligned} R_0(\epsilon) &= 1 - \log[1 + (4\epsilon(1-\epsilon))^{1/2}] \\ C(\epsilon) &= 1 - h_2(\epsilon). \end{aligned} \tag{4.6.2}$$

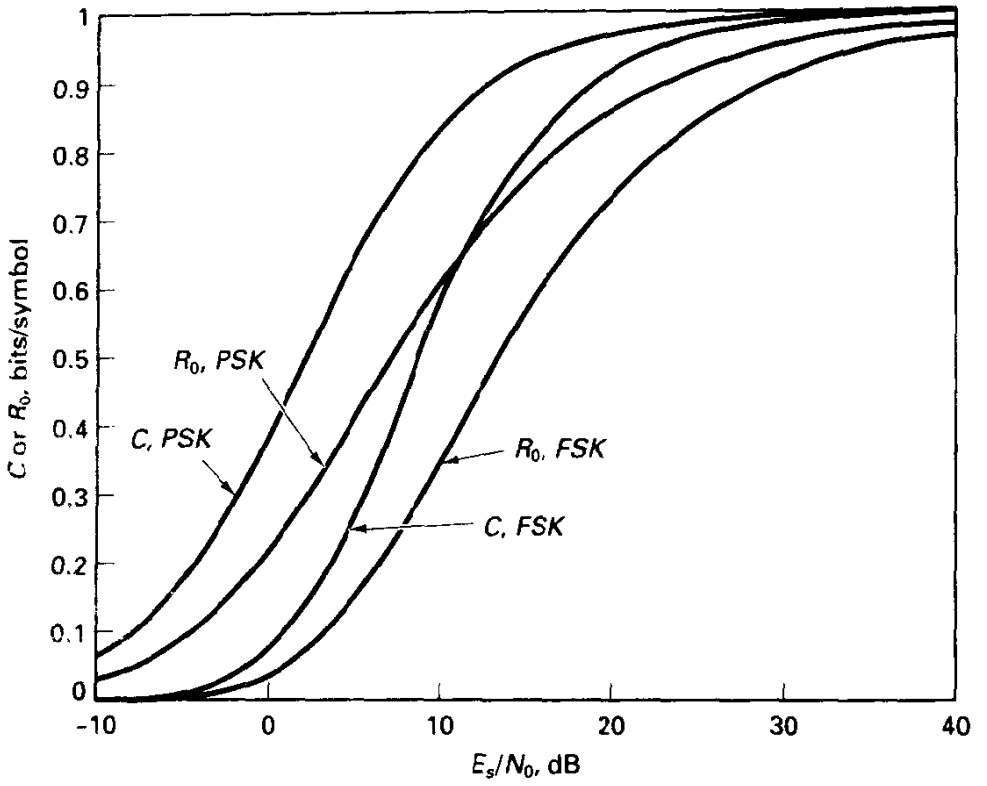

Figure 4.6.2 Capacity for R_0 on Rayleigh channel, coherent PSK and noncoherent FSK, hard decisions.

Figure 4.6.2 presents results for binary orthogonal signals with noncoherent detection and for antipodal signaling with coherent detection, these representing two typical choices. By comparing these results with those of Figure 4.5.2 we determine the energy penalty attached to the Rayleigh channel. Note that for throughput approaching 1 bit per symbol (nearing uncoded transmission) the penalties are indeed large, while as the code rate decreases, the energy penalty diminishes, as measured by C or R_0. For example, to obtain $R_0 = 0.5$ bit/symbol with coherent PSK, the nonfading channel can operate with about 5.7 dB smaller mean SNR than a Rayleigh, fully interleaved channel when hard-decision demodulation is performed. The comparison for noncoherent FSK gives a similar difference. Thus, although fading still exacts a penalty, it is much smaller than the 20- to 40-dB penalties attached to fading with uncoded transmission. Furthermore, if we are allowed still lower coding rate, the penalty is even smaller, as we see from these figures.

As with the AWGN channel, we may view code rate $0 < R \leq 1$ as a design variable and determine the minimum E_b/N_0 (average) necessary to maintain C or R_0 above this code rate. For example, we let $C(x)$ represent the functional dependence of capacity on the quality parameter $E_s/N_0 = RE_b/N_0$ and find E_b/N_0 by solving

$$R = C[\epsilon(RE_b/N_0)] \tag{4.6.3}$$

as R varies, defining the minimum E_b/N_0 allowed at this rate. The same could be done for the R_0 parameter. Results are shown in Figure 4.6.3, and in particular we find

that minimum energy operation on the Rayleigh channel points to rather small binary code rates. Provided that bandwidth expansion allows such small code rates, comparison of Figures 4.6.3 and 4.5.4 shows that the penalty exacted by the Rayleigh channel is minimal, on the order of a decibel for PSK. Keep in mind that full interleaving is assumed throughout.

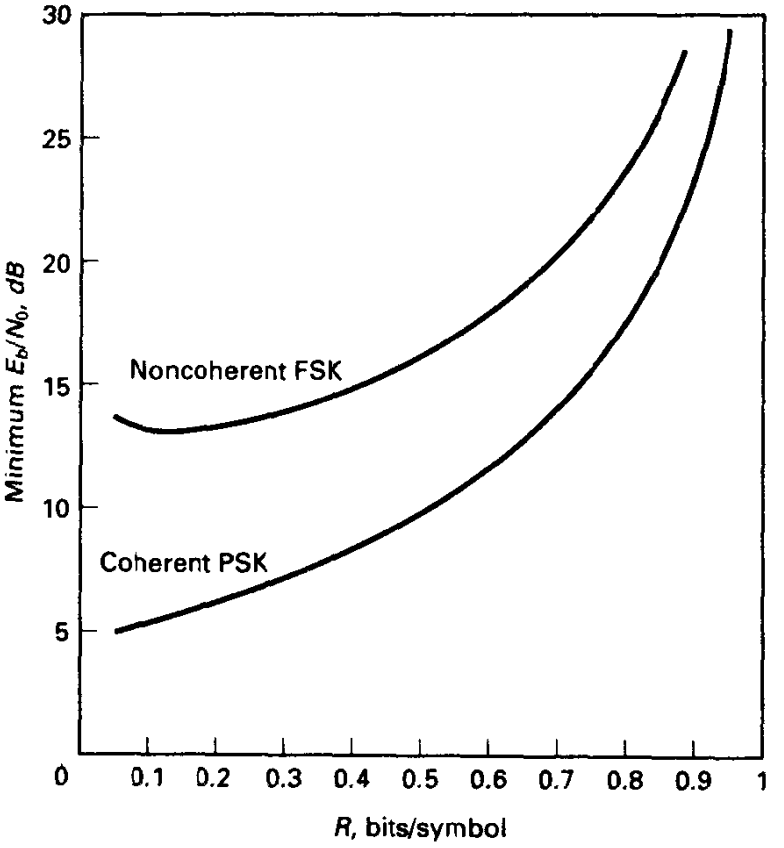

Figure 4.6.3 Minimum E_b/N_0 maintaining $R_0 > R$ for interleaved Rayleigh channel, binary signaling, hard decisions, no side information.

Hard Decisions with Perfect Side Information

Next, suppose that the decoder is supplied perfect knowledge of the channel gain a_j for each transmission. (This is difficult to estimate perfectly in a noisy environment, but slow-fading conditions make this estimation somewhat easier.) The decoder's metric for the discrete channel is based on the log-likelihood function, now including a_j in the "observables":[8]

$$\lambda(y_j, x_j; a_j) = \log f(y_j, a_j | x_j). \tag{4.6.4}$$

(Some may find it more natural to include a_j in the conditioning for y_j, since a_j is a given in the problem, and this view helps in writing conditional p.d.f.'s. As shown in

[8]We choose to express the metric in a form where side information is carried along after the semicolon to highlight its auxiliary or optional role.

Appendix 4A1, when the input and channel amplitude are independent random variables, as assumed, then

$$f(y_j, a_j|x_j) = cf(y_j|x_j, a_j),$$

where c does not depend on x_j, so either may be used.)

Since a_j determines the crossover probability in a given transmission, and transmissions are independent after interleaving, we can immediately write that

$$f(y_j, a_j|x_j) = \epsilon(a_j)^{d_H(x_j, y_j)}[1 - \epsilon(a_j)]^{1-d_H(x_j, y_j)}, \tag{4.6.5}$$

where $\epsilon(a_j)$ represents the crossover probability of the channel, given a specific amplitude a_j. The ML metric, with side information, then becomes, after dispensing with bias terms,

$$\lambda(y_j, x_j; a_j) = d_H(x_j, y_j) \log\left[\frac{\epsilon(a_j)}{1 - \epsilon(a_j)}\right], \tag{4.6.6}$$

where for coherent PSK

$$\epsilon(a_j) = Q\left[\left(\frac{a_j^2 2E_s}{N_0}\right)^{1/2}\right] \tag{4.6.7a}$$

and for noncoherent FSK

$$\epsilon(a_j) = \frac{1}{2}e^{-a_j^2 E_s/2N_0}. \tag{4.6.7b}$$

(In these expressions E_s/N_0 remains as the average symbol energy-to-noise density ratio.) Note that in contrast to the no-side-information Hamming metric, the side-information metric incorporates a scaling of each Hamming distance calculation, based on the instantaneous channel crossover probability. Badly faded intervals are basically ignored, while symbols with good SNR are weighted strongly. Hagenauer [20] has earlier derived this combining policy and studied binary coding on the Rayleigh channel in detail.

We will defer discussion of channel capacity until Example 4.12.

Unquantized Demodulation, Perfect Side Information (PSI)

Now suppose that the binary demodulator produces unquantized demodulator output vectors $\mathbf{y}_j$ at each time j for the channel decoder, rather than making binary decisions. Let us assume again that the demodulator can also supply side information to the decoder in the form of a_j. To determine the optimal metric, we write the log-likelihood function as usual. In the case of antipodal signaling, the optimal decoder metric is a weighted correlation, as in Example 4.4:

$$\lambda(y_j, x_j; a_j) = a_j y_j \tilde{x}_j, \tag{4.6.8}$$

where $\tilde{x}_j$ is the ± 1 version of the binary signal. For noncoherent FSK transmission, the optimal metric applied to $\mathbf{y}_j = (y_{j_0}, y_{j_1})$ follows from the Rician–Rayleigh joint p.d.f.:

$$\lambda(\mathbf{y}_j, x_j = m; a_j) = \log f(\mathbf{y}_j, a_j|x_j = m) \to \log I_0\left(\frac{a_j \mu y_{j_m}}{\sigma_j^2}\right), \tag{4.6.9a}$$

where $\mu = E_s^{1/2}$ and $\sigma^2 = N_0/2$. Therefore, the decoder utilizes only the demodulator output corresponding to the symbol under test. The amplitude scaling appears in a

more complicated manner for this second case, and the preceding metric would often be approximated as

$$\lambda(\mathbf{y}_j, x_j = m; a_j) \approx a_j^2 y_{j_m}^2, \tag{4.6.9b}$$

generalizing an approximation for the nonfading channel.

Expressions for capacity and R_0 mimic those found in Section 4.5 for the AWGN channel, since interleaving has rendered the channel memoryless, but we now must include the fading amplitude in the averaging calculation:

$$C_{\text{PSI. UQ}} = \sum_{x=0}^{1} \int_a \int_{\mathbf{y}} \frac{1}{2} f(\mathbf{y}, a|x) \log \left[\frac{f(\mathbf{y}, a|x)}{f(\mathbf{y}, a)} \right] d\mathbf{y}\, da \tag{4.6.10a}$$

and

$$R_{0_{\text{PSI. UQ}}} = \int_{a=0}^{\infty} \int_{\mathbf{y}} \left[\sum_{x=0}^{1} \frac{1}{2} f(\mathbf{y}, a|x)^{1/2} \right]^2 d\mathbf{y}\, da. \tag{4.6.10b}$$

We have used equiprobable input distributions due to the symmetry of the problem. To calculate (4.6.10), it is convenient to use

$$f(\mathbf{y}, a|x) = f(\mathbf{y}|x, a) f(a|x) = f(\mathbf{y}|x, a) f(a) \tag{4.6.10c}$$

since amplitude is presumed independent of the channel input.

A simpler intuitive understanding of the channel capacity calculation is possible when perfect side information is available, deriving from straightforward information theory statements. Recall that capacity is the maximum mutual information between channel input and output, which now includes the presumed known channel side information, which we denote more generally by S, for channel state:

$$C = \max_{P(x)} I(X; Y, S). \tag{4.6.11}$$

Using the facts that ([8], see also Exercise 4.6.1)

$$I(X; Y, S) = I(X; Y|S) + I(X; S) \tag{4.6.12}$$

and that X and channel state S are assumed independent, we have

$$I(X; Y, S) = I(X; Y|S), \tag{4.6.13}$$

which is expressed as

$$\begin{aligned} I(X; Y|S) &= \int_s I(X; Y|S = s) f_S(s)\, ds \\ &= \int_s \sum_{x_i} \int_{\mathbf{y}} P(x_i|s) f(\mathbf{y}|x_i, s) \log \left[\frac{f(\mathbf{y}|x_i, s)}{f(\mathbf{y}|s)} d\mathbf{y} \right] f_S(s)\, ds \end{aligned} \tag{4.6.14}$$

and is just the mutual information for each channel state averaged over the distribution of states, here fading amplitudes. Now, if the input distribution that maximizes mutual information is the *same for all states* we have that

$$C_{\text{PSI}} = \int_s C(S = s) f_S(s)\, ds. \tag{4.6.15}$$