Detection, Estimation, and Modulation Theory, Part I: Detection, Estimation, and Linear Modulation Theory. Harry L. Van Trees

ISBNs: 0-471-09517-6 (Paperback); 0-471-22108-2 (Electronic)

4

Detection of Signals —Estimation of Signal Parameters

4.1 INTRODUCTION

In Chapter 2 we formulated the detection and estimation problems in the classical context. In order to provide background for several areas, we first examined a reasonably general problem. Then, in Section 2.6 of Chapter 2, we investigated the more precise results that were available in the general Gaussian case.

In Chapter 3 we developed techniques for representing continuous processes by sets of numbers. The particular representation that we considered in detail was appropriate primarily for Gaussian processes.

We now want to use these representations to extend the results of the classical theory to the case in which the observations consist of *continuous waveforms.*

4.1.1 Models

The problems of interest to us in this chapter may be divided into two categories. The first is the detection problem which arises in three broad areas: digital communications, radar/sonar, and pattern recognition and classification. The second is the signal parameter estimation problem which also arises in these three areas.

Detection. The conventional model of a simple digital communication system is shown in Fig. 4.1. The source puts out a binary digit (either 0 or 1) every T seconds. The most straightforward system would transmit either $\sqrt{E_{t_0}}\, s_0(t)$ or $\sqrt{E_{t_1}}\, s_1(t)$ during each interval. In a typical space communication system an attenuated version of the transmitted signal would be received with negligible distortion. The received signal consists of $\sqrt{E_0}\, s_0(t)$ or $\sqrt{E_1}\, s_1(t)$ plus an additive noise component.

The characterization of the noise depends on the particular application. One source, always present, is thermal noise in the receiver front end. This

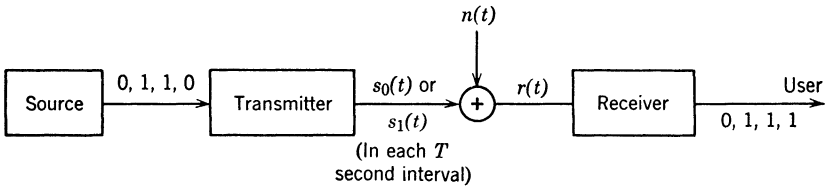

Fig. 4.1 A digital communication system.

noise can be modeled as a sample function from a Gaussian random process. As we proceed to more complicated models, we shall encounter other sources of interference that may turn out to be more important than the thermal noise. In many cases we can redesign the system to eliminate these other interference effects almost entirely. Then the thermal noise will be the disturbance that limits the system performance. In most systems the spectrum of the thermal noise is flat over the frequency range of interest, and we may characterize it in terms of a spectral height of $N_0/2$ joules. An alternate characterization commonly used is effective noise temperature T_e (e.g., Valley and Wallman [1] or Davenport and Root [2], Chapter 10). The two are related simply by

$$N_0 = kT_e, \tag{1}$$

where k is Boltzmann's constant, 1.38×10^{-23} joule/°K and T_e is the effective noise temperature, °K.

Thus in this particular case we could categorize the receiver design as a problem of detecting one of two *known signals in the presence of additive white Gaussian noise.*

If we look into a possible system in more detail, a typical transmitter could be as shown in Fig. 4.2. The transmitter has an oscillator with nominal center frequency of ω_c. It is biphase modulated according to whether the source output is 1 (0°) or 0 (180°). The oscillator's instantaneous phase varies slowly, and the receiver must include some auxiliary equipment to measure the oscillator phase. If the phase varies slowly enough, we shall see that accurate measurement is possible. If this is true, the problem may be modeled as above. If the measurement is not accurate, however, we must incorporate the phase uncertainty in our model.

A second type of communication system is the point-to-point ionospheric scatter system shown in Fig. 4.3 in which the transmitted signal is scattered by the layers in the ionosphere. In a typical system we can transmit a "one" by sending a sine wave of a given frequency and a "zero" by a sine wave of another frequency. The receiver signal may vary as shown in Fig. 4.4. Now, the receiver has a signal that fluctuates in amplitude and phase.

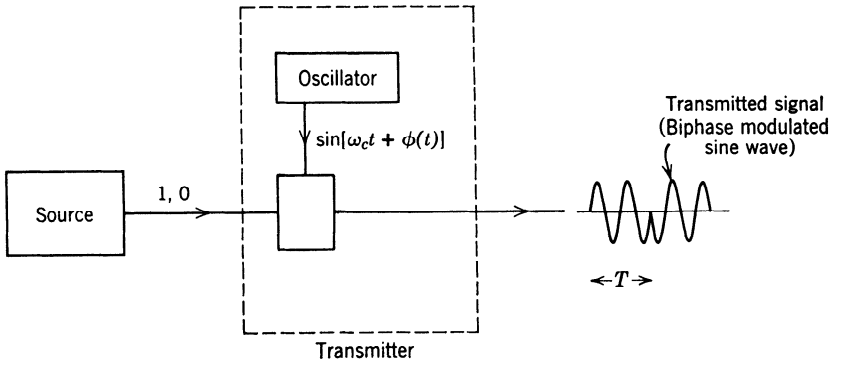

Fig. 4.2 Details of typical system.

In the commonly-used frequency range most of the additive noise is Gaussian.

Corresponding problems are present in the radar context. A conventional pulsed radar transmits a signal as shown in Fig. 4.5. If a target is present, the sequence of pulses is reflected. As the target fluctuates, the amplitude and phase of the reflected pulses change. The returned signal consists of a sequence of pulses whose amplitude and phase are unknown. The problem

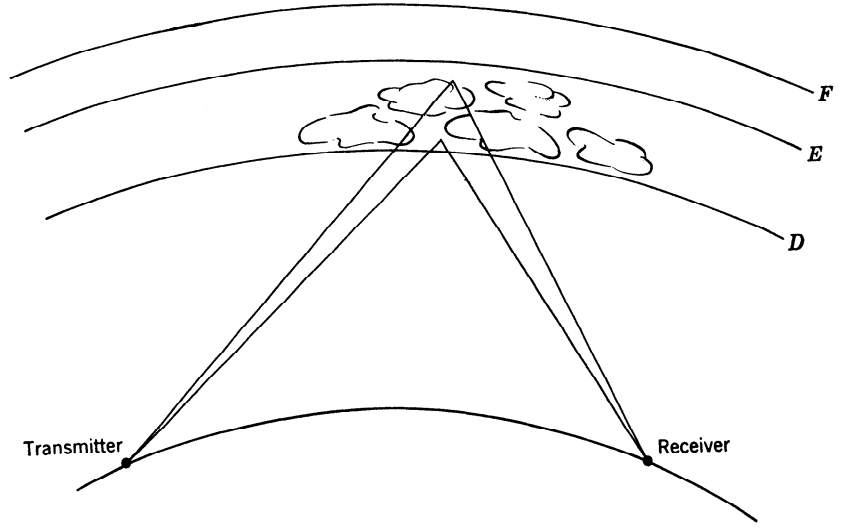

Fig. 4.3 Ionospheric scatter link.

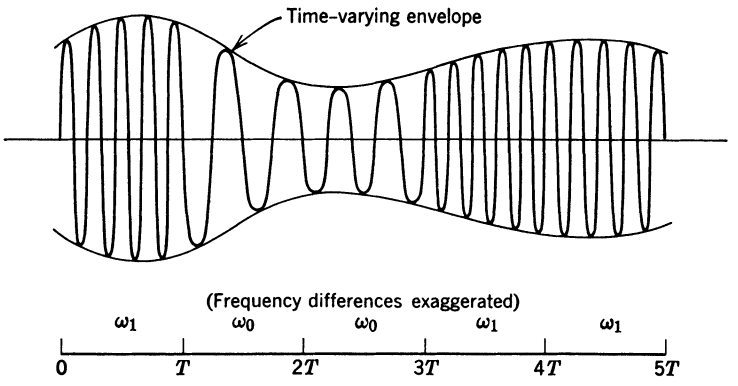

Fig. 4.4 Signal component in time-varying channel.

is to examine this sequence in the presence of receiver noise and decide whether a target is present.

There are obvious similarities between the two areas, but there are also some differences:

1. In a digital communication system the two types of error (say 1, when 0 was sent, and vice versa) are usually of equal importance. Furthermore, a signal may be present on both hypotheses. This gives a symmetry to the problem that can be exploited. In a radar/sonar system the two types of

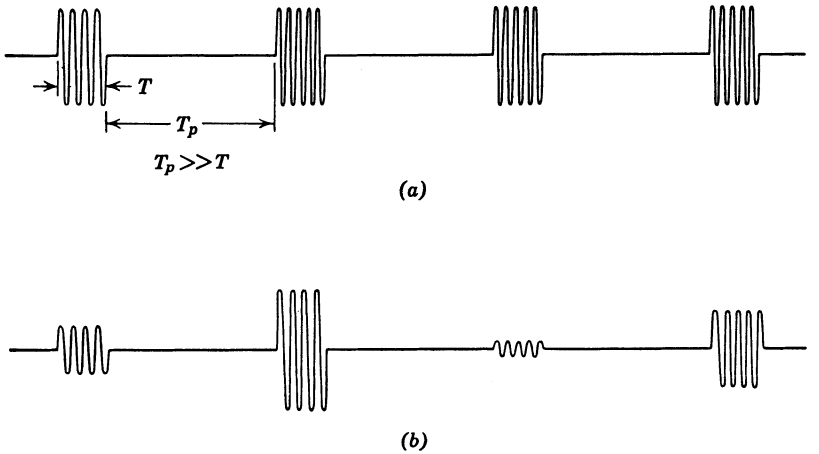

Fig. 4.5 Signals in radar model: (*a*) transmitted sequence of rf pulses; (*b*) received sequence [amplified (time-shift not shown)].

error are almost always of unequal importance. In addition, a signal is present only on one hypothesis. This means that the problem is generally nonsymmetric.

2. In a digital communication system the probability of error is usually an adequate measure of system performance. Normally, in radar/sonar a reasonably complete ROC is needed.

3. In a digital system we are sending a sequence of digits. Thus we can correct digit errors by putting some structure into the sequence. In the radar/sonar case this is not an available alternative.

In spite of these differences, a great many of the basic results will be useful for both areas.

Estimation. The second problem of interest is the estimation of signal parameters, which is encountered in both the communications and radar/sonar areas. We discuss a communication problem first.

Consider the analog message source shown in Fig. 4.6*a*. For simplicity we assume that it is a sample function from a bandlimited random process ($2W$ cps: double-sided). We could then sample it every $1/2W$ seconds without losing any information. In other words, given these samples at the

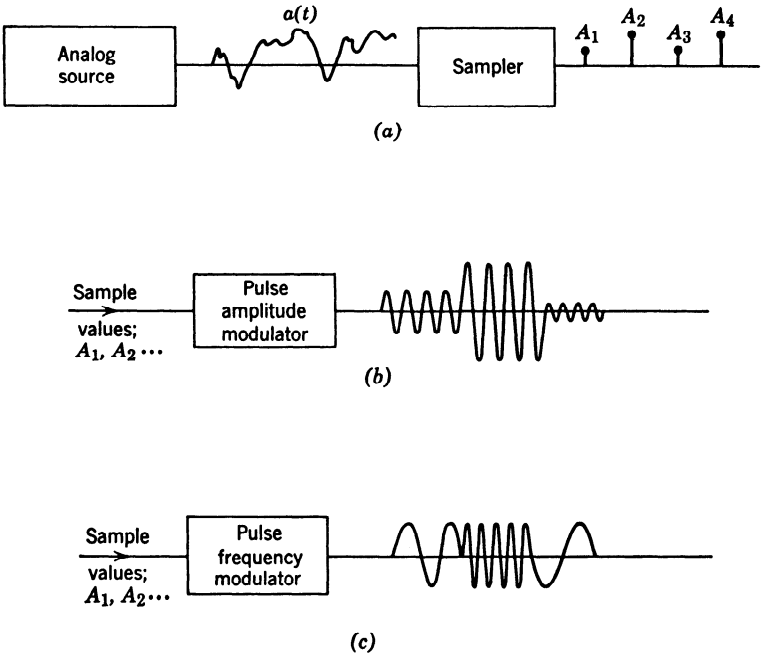

Fig. 4.6 Analog message transmission.

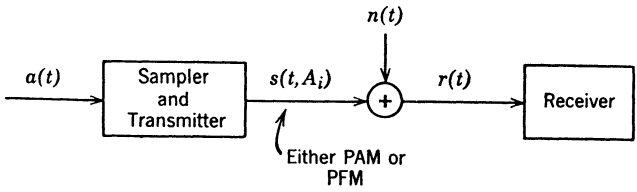

Fig. 4.7 A parameter transmission system.

receiver, we could reconstruct the message exactly (e.g. Nyquist, [4] or Problem 3.3.6). Every T seconds ($T = 1/2W$) we transmit a signal that depends on the particular value A_i at the last sampling time. In the system in Fig. 4.6*b* the amplitude of a sinusoid depends on the value of A_i. This system is referred to as a pulse-amplitude modulation system (PAM). In the system in Fig. 4.6*c* the frequency of the sinusoid depends on the sample value. This system is referred to as a pulse frequency modulation system (PFM). The signal is transmitted over a channel and is corrupted by noise (Fig. 4.7). The received signal in the ith interval is:

$$r(t) = s(t, A_i) + n(t), \qquad T_i \leq t \leq T_{i+1}. \tag{2}$$

The purpose of the receiver is to estimate the values of the successive A_i and use these estimates to reconstruct the message.

A typical radar system is shown in Fig. 4.8. In a conventional pulsed radar the transmitted signal is a sinusoid with a rectangular envelope.

$$\begin{aligned} s_t(t) &= \sqrt{2E_t} \sin \omega_c t, \qquad 0 \leq t \leq T, \\ &= 0, \qquad\qquad\quad \text{elsewhere.} \end{aligned} \tag{3a}$$

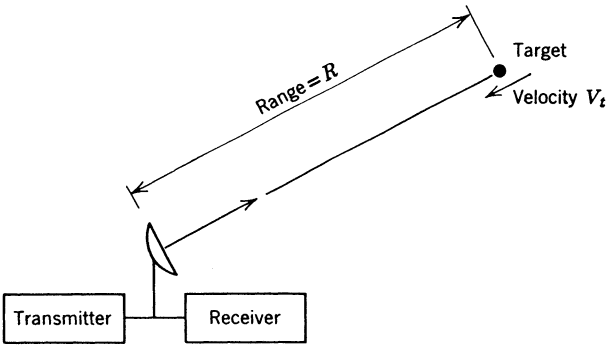

Fig. 4.8 Radar system diagram.

The returned signal is delayed by the round-trip time to the target. If the target is moving, there is Doppler shift. Finally there is a random amplitude and phase due to the target fluctuation. The received signal in the absence of noise is

$$\begin{aligned} s_r(t) &= v\sqrt{2E_t}\sin[(\omega_c + \omega_D)(t-\tau) + \phi], && \tau \le t \le \tau + T. \\ &= 0, && \text{elsewhere.} \end{aligned} \tag{3b}$$

Here we estimate τ and ω_D (or, equivalently, target range and velocity). Once again, there are obvious similarities between the communication and radar/sonar problem. The basic differences are the following:

1. In the communications context A_i is a random variable with a probability density that is usually known. In radar the range or velocity limits of interest are known. The parameters, however, are best treated as nonrandom variables (e.g., discussion in Section 2.4).
2. In the radar case the difficulty may be compounded by a lack of knowledge regarding the target's presence. Thus the detection and estimation problem may have to be combined.
3. In almost all radar problems a phase reference is not available.

Other models of interest will appear naturally in the course of our discussion.

4.1.2 Format

Because this chapter is long, it is important to understand the over-all structure. The basic approach has three steps:

1. The observation consists of a waveform $r(t)$. Thus the observation space may be infinite dimensional. Our first step is to map the received signal into some convenient decision or estimation space. This will reduce the problem to one studied in Chapter 2.
2. In the detection problem we then select decision regions and compute the ROC or Pr (ϵ). In the estimation problem we evaluate the variance or mean-square error.
3. We examine the results to see what they imply about system design and performance.

We carry out these steps for a sequence of models of increasing complexity (Fig. 4.9) and develop the detection and estimation problem in parallel. By emphasizing their parallel nature for the simple cases, we can save appreciable effort in the more complex cases by considering only one problem in the text and leaving the other as an exercise. We start with the simple models and then proceed to the more involved.

Channel	Signal detection	Signal parameter estimation
Additive white noise	Simple binary	Single parameter, linear
Additive colored noise	General binary	Single parameter, nonlinear
Simple random	M-ary	Multiple parameter
Multiple channels		

Fig. 4.9 Sequence of models.

A logical question is: if the problem is so simple, why is the chapter so long? This is a result of our efforts to determine how the model and its parameters affect the design and performance of the system. We feel that only by examining some representative problems in detail can we acquire an appreciation for the implications of the theory.

Before proceeding to the solution, a brief historical comment is in order. The mathematical groundwork for our approach to this problem was developed by Grenander [5]. The detection problem relating to optimum radar systems was developed at the M.I.T. Radiation Laboratory, (e.g., Lawson and Uhlenbeck [6]) in the early 1940's. Somewhat later Woodward and Davies [7, 8] approached the radar problem in a different way. The detection problem was formulated at about the same time in a manner similar to ours by both Peterson, Birdsall, and Fox [9] and Middleton and Van Meter [10], whereas the estimation problem was first done by Slepian [11]. Parallel results with a communications emphasis were developed by Kotelnikov [12, 13] in Russia. Books that deal almost exclusively with radar include Helstrom [14] and Wainstein and Zubakov [15]. Books that deal almost exclusively with communication include Kotelnikov [13], Harman [16], Baghdady (ed.) [17], Wozencraft and Jacobs [18], and Golomb et al. [19]. The last two parts of Middleton [47] cover a number of topics in both areas. By presenting the problems side by side we hope to emphasize their inherent similarities and contrast their differences.

4.2 DETECTION AND ESTIMATION IN WHITE GAUSSIAN NOISE

In this section we formulate and solve the detection and estimation problems for the case in which the interference is additive white Gaussian noise.

We consider the detection problem first: the simple binary case, the general binary case, and the M-ary case are discussed in that order. By using the concept of a sufficient statistic the optimum receiver structures are simply derived and the performances for a number of important cases are evaluated. Finally, we study the sensitivity of the optimum receiver to the detailed assumptions of our model.

As we have seen in the classical context, the decision and estimation problems are closely related; linear estimation will turn out to be essentially the same as simple binary detection. When we proceed to the nonlinear estimation problem, new issues will develop, both in specifying the estimator structure and in evaluating its performance.

4.2.1 Detection of Signals in Additive White Gaussian Noise

Simple Binary Detection. In the simplest binary decision problem the received signal under one hypothesis consists of a completely known signal, $\sqrt{E}\,s(t)$, corrupted by an additive zero-mean white Gaussian noise $w(t)$ with spectral height $N_0/2$; the received signal under the other hypothesis consists of the noise $w(t)$ alone. Thus

$$\begin{aligned} r(t) &= \sqrt{E}\,s(t) + w(t), \qquad 0 \le t \le T: H_1, \\ &= w(t), \qquad\qquad\quad 0 \le t \le T: H_0. \end{aligned} \tag{4}$$

For convenience we assume that

$$\int_0^T s^2(t)\,dt = 1, \tag{5}$$

so that E represents the received signal energy. The problem is to observe $r(t)$ over the interval $[0, T]$ and decide whether H_0 or H_1 is true. The criterion may be either Bayes or Neyman-Pearson.

The following ideas will enable us to solve this problem easily:

1. Our observation is a time-continuous random waveform. The first step is to reduce it to a set of random variables (possibly a countably infinite set).
2. One method is the series expansion of Chapter 3:

$$r(t) = \underset{K\to\infty}{\text{l.i.m.}} \sum_{i=1}^{K} r_i\,\phi_i(t); \qquad 0 \le t \le T. \tag{6}$$

When $K = K'$, there are K' coefficients in the series, $r_1, \ldots, r_{K'}$ which we could denote by the vector $\mathbf{r}_{K'}$. In our subsequent discussion we suppress the K' subscript and denote the coefficients by $\mathbf{r}$.

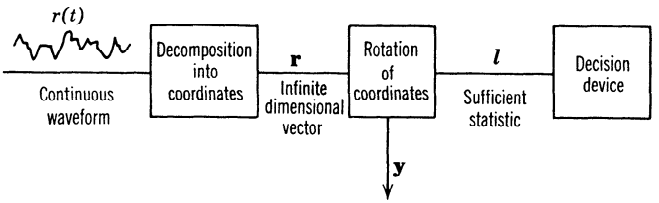

Fig. 4.10 Generation of sufficient statistics.

3. In Chapter 2 we saw that if we transformed $\mathbf{r}$ into two independent vectors, $\boldsymbol{l}$ (the sufficient statistic) and $\mathbf{y}$, as shown in Fig. 4.10, our decision could be based only on $\boldsymbol{l}$, because the values of $\mathbf{y}$ did not depend on the hypothesis. The advantage of this technique was that it reduced the dimension of the decision space to that of $\boldsymbol{l}$. Because this is a binary problem we know that $\boldsymbol{l}$ will be one-dimensional.

Here the method is straightforward. If we choose the first orthonormal function to be $s(t)$, the first coefficient in the decomposition is the Gaussian random variable,

$$r_1 = \begin{cases} \int_0^T s(t)\, w(t)\, dt \triangleq w_1 : H_0, \\ \int_0^T s(t)[\sqrt{E}\, s(t) + w(t)]\, dt = \sqrt{E} + w_1 : H_1. \end{cases} \tag{7}$$

The remaining r_i ($i > 1$) are Gaussian random variables which can be generated by using some arbitrary orthonormal set whose members are orthogonal to $s(t)$.

$$r_i = \begin{cases} \int_0^T \phi_i(t)\, w(t)\, dt \triangleq w_i : H_0, \\ \int_0^T \phi_i(t)[\sqrt{E}\, s(t) + w(t)]\, dt = w_i : H_1, \end{cases} \qquad i \neq 1. \tag{8}$$

From Chapter 3 (44) we know that

$$E(w_i w_j) = 0; \qquad i \neq j.$$

Because w_i and w_j are jointly Gaussian, they are statistically independent (see Property 3 on p. 184).

We see that *only* r_1 depends on which hypothesis is true. Further, all r_i ($i > 1$) are statistically independent of r_1. Thus r_1 is a sufficient statistic ($r_1 = l$). The other r_i correspond to $\mathbf{y}$. Because they will not affect the decision, there is no need to compute them.

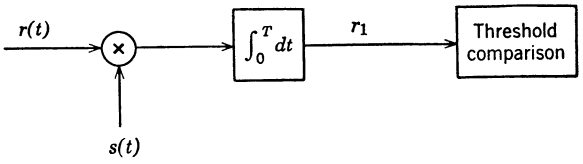

Fig. 4.11 Correlation receiver.

Several equivalent receiver structures follow immediately. The structure in Fig. 4.11 is called a *correlation receiver*. It correlates the input $r(t)$ with a stored replica of the signal $s(t)$. The output is r_1, which is a sufficient statistic ($r_1 = l$) and is a Gaussian random variable. Once we have obtained r_1, the decision problem will be identical to the classical problem in Chapter 2 (specifically, Example 1 on pp. 27–28). We compare l to a threshold in order to make a decision.

An equivalent realization is shown in Fig. 4.12. The impulse response of the linear system is simply the signal reversed in time and shifted,

$$h(\tau) = s(T - \tau). \tag{9}$$

The output at time T is the desired statistic l. This receiver is called a *matched filter receiver*. (It was first derived by North [20].) The two structures are mathematically identical; the choice of which structure to use depends solely on ease of realization.

Just as in Example 1 of Chapter 2, the sufficient statistic l is Gaussian under either hypothesis. Its mean and variance follow easily:

$$\begin{aligned} E(l|H_1) &= E(r_1|H_1) = \sqrt{E}, \\ E(l|H_0) &= E(r_1|H_0) = 0, \\ \text{Var}\,(l|H_0) &= \text{Var}\,(l|H_1) = \frac{N_0}{2}. \end{aligned} \tag{10}$$

Thus we can use the results of Chapter 2, (64)–(68), with

$$d = \left(\frac{2E}{N_0}\right)^{1/2}. \tag{11}$$

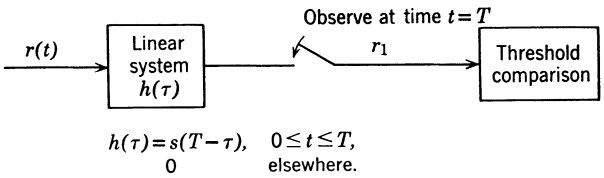

Fig. 4.12 Matched filter receiver.

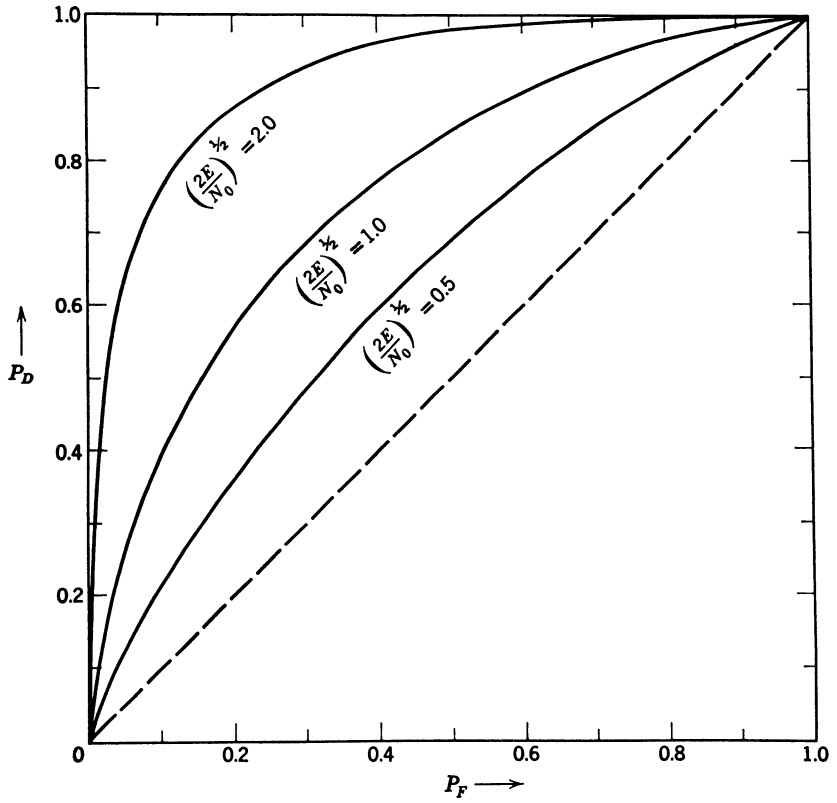

Fig. 4.13 Receiver operating characteristic: known signal in additive white Gaussian noise.

The curves in Figs. 2.9*a* and 2.9*b* of Chapter 2 are directly applicable and are reproduced as Figs. 4.13 and 4.14. We see that the performance depends only on the received signal energy E and the noise spectral height N_0—the signal shape is not important. This is intuitively logical because the noise is the same along any coordinate.

The key to the simplicity in the solution was our ability to reduce an infinite dimensional observation space to a one-dimensional decision space by exploiting the idea of a sufficient statistic. Clearly, we should end up with the same receiver even if we do not recognize that a sufficient statistic is available. To demonstrate this we construct the likelihood ratio directly. Three observations lead us easily to the solution.

1. If we approximate $r(t)$ in terms of some finite set of numbers,

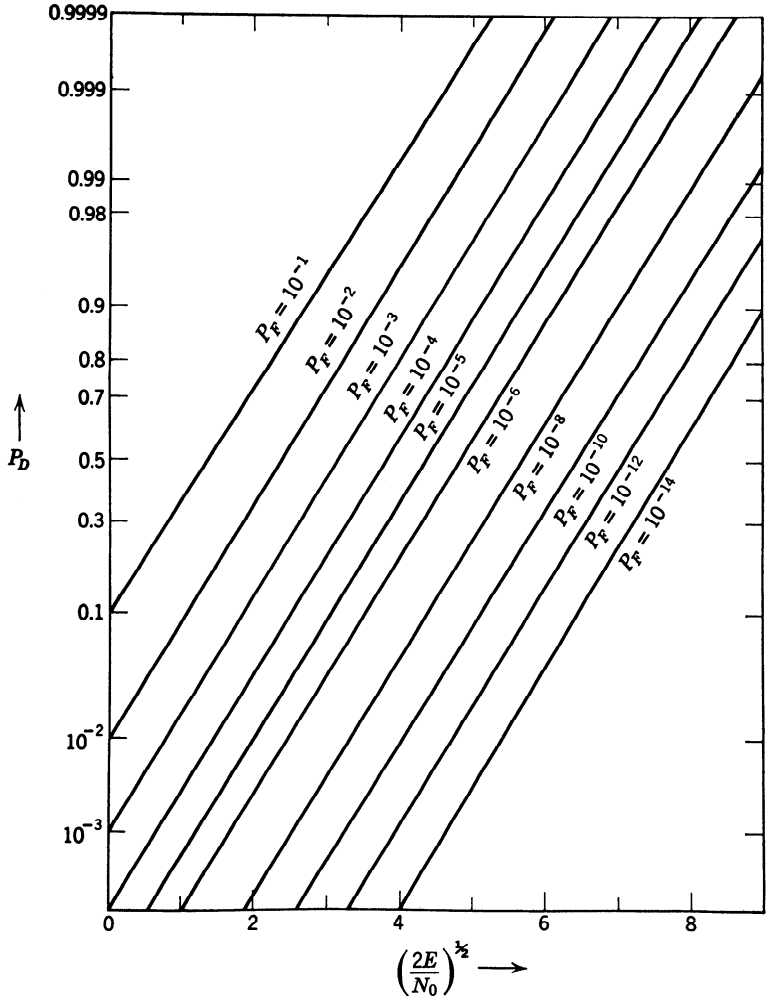

Fig. 4.14 Probability of detection vs $\left(\frac{2E}{N_0}\right)^{1/2}$.

$r_1, \ldots, r_K$, (**r**), we have a problem in classical detection theory that we can solve.

2. If we choose the set $r_1, r_2, \ldots, r_K$ so that

$$p_{r_1, r_2, \ldots, r_K | H_i}(R_1, R_2, \ldots, R_K | H_i) = \prod_{j=1}^{K} p_{r_j | H_i}(R_j | H_i), \qquad i = 0, 1, \quad (12)$$

that is, the observations are conditionally independent, we have an *easy* problem to solve.

3. Because we know that it requires an infinite set of numbers to represent $r(t)$ completely, we want to get the solution in a convenient form so that we can let $K \to \infty$.

We denote the approximation that uses K coefficients as $r_K(t)$. Thus

$$r_K(t) = \sum_{i=1}^{K} r_i \phi_i(t), \qquad 0 \le t \le T, \tag{13}$$

where

$$r_i = \int_0^T r(t)\, \phi_i(t)\, dt, \qquad i = 1,2,\ldots K, \tag{14}$$

and the $\phi_i(t)$ belong to an *arbitrary* complete orthonormal set of functions. Using (14), we see that under H_0

$$r_i = \int_0^T w(t)\, \phi_i(t)\, dt = w_i, \tag{15}$$

and under H_1

$$r_i = \int_0^T \sqrt{E}\, s(t)\, \phi_i(t)\, dt + \int_0^T w(t)\, \phi_i(t)\, dt = s_i + w_i. \tag{16}$$

The coefficients s_i correspond to an expansion of the signal

$$s_K(t) \triangleq \sum_{i=1}^{K} s_i \phi_i(t), \qquad 0 \le t \le T, \tag{17}$$

and

$$\sqrt{E}\, s(t) = \lim_{K \to \infty} s_K(t). \tag{18}$$

The r_i's are Gaussian with known statistics:

$$\begin{aligned} E(r_i|H_0) &= 0, \\ E(r_i|H_1) &= s_i, \\ \text{Var}\,(r_i|H_0) &= \text{Var}\,(r_i|H_1) = \frac{N_0}{2}. \end{aligned} \tag{19}$$

Because the noise is "white," these coefficients are independent along any set of coordinates. The likelihood ratio is

$$\Lambda[r_K(t)] = \frac{p_{\mathbf{r}|H_1}(\mathbf{R}|H_1)}{p_{\mathbf{r}|H_0}(\mathbf{R}|H_0)} = \frac{\prod_{i=1}^{K} \frac{1}{\sqrt{\pi N_0}} \exp\left(-\frac{1}{2}\frac{(R_i - s_i)^2}{N_0/2}\right)}{\prod_{i=1}^{K} \frac{1}{\sqrt{\pi N_0}} \exp\left(-\frac{1}{2}\frac{R_i^2}{N_0/2}\right)}. \tag{20}$$

Taking the logarithm and canceling common terms, we have

$$\ln \Lambda[r_K(t)] = \frac{2}{N_0} \sum_{i=1}^{K} R_i s_i - \frac{1}{N_0} \sum_{i=1}^{K} s_i^2. \tag{21}$$

The two sums are easily expressed as integrals. From Parseval's theorem,

$$\sum_{i=1}^{K} R_i s_i = \int_0^T r_K(t)\, s_K(t)\, dt$$

and

$$\sum_{i=1}^{K} s_i^2 = \int_0^T s_K^2(t)\, dt. \tag{22}$$

We now have the log likelihood ratio in a form in which it is convenient to pass to the limit:

$$\underset{K\to\infty}{\text{l.i.m.}}\ \ln \Lambda[r_K(t)] \triangleq \ln \Lambda[r(t)] = \frac{2\sqrt{E}}{N_0}\int_0^T r(t)\, s(t)\, dt - \frac{E}{N_0}. \tag{23}$$

The first term is just the sufficient statistic we obtained before. The second term is a bias. The resulting likelihood ratio test is

$$\frac{2\sqrt{E}}{N_0}\int_0^T r(t)\, s(t)\, dt \underset{H_0}{\overset{H_1}{\gtrless}} \ln \eta + \frac{E}{N_0}. \tag{24}$$

(Recall from Chapter 2 that η is a constant which depends on the costs and a priori probabilities in a Bayes test and the desired P_F in a Neyman-Pearson test.) It is important to observe that even though the probability density $p_{r(t)|H_i}(r(t)|H_i)$ is not well defined for either hypothesis, the likelihood ratio is.

Before going on to more general problems it is important to emphasize the two separate features of the signal detection problem:

1. First we reduce the received waveform to a single number which is a point in a decision space. This operation is performed physically by a correlation operation and is invariant to the decision criterion that we plan to use. This invariance is important because it enables us to construct the waveform processor without committing ourselves to a particular criterion.

2. Once we have transformed the received waveform into the decision space we have only the essential features of the problem left to consider. Once we get to the decision space the problem is the same as that studied in Chapter 2. The actual received waveform is no longer important and all physical situations that lead to the same picture in a decision space are identical for our purposes. In our simple example we saw that all signals of equal energy map into the same point in the decision space. It is therefore obvious that the signal shape is unimportant.

The separation of these two parts of the problem leads to a clearer understanding of the fundamental issues.

General Binary Detection in White Gaussian Noise. The results for the simple binary problem extend easily to the general binary problem. Let

$$\begin{aligned} r(t) &= \sqrt{E_1}\, s_1(t) + w(t), \qquad 0 \le t \le T{:}H_1, \\ &= \sqrt{E_0}\, s_0(t) + w(t), \qquad 0 \le t \le T{:}H_0, \end{aligned} \tag{25}$$

where $s_0(t)$ and $s_1(t)$ are normalized but are *not* necessarily orthogonal. We denote the correlation between the two signals as

$$\rho \triangleq \int_0^T s_0(t)\, s_1(t)\, dt.$$

(Note that $|\rho| \le 1$ because the signals are normalized.)

We choose our first two orthogonal functions as follows:

$$\phi_1(t) = s_1(t), \qquad 0 \le t \le T, \tag{26}$$

$$\phi_2(t) = \frac{1}{\sqrt{1-\rho^2}}\,[s_0(t) - \rho s_1(t)], \qquad 0 \le t \le T. \tag{27}$$

We see that $\phi_2(t)$ is obtained by subtracting out the component of $s_0(t)$ that is correlated with $\phi_1(t)$ and normalizing the result. The remaining $\phi_i(t)$ consist of an arbitrary orthonormal set whose members are orthogonal to $\phi_1(t)$ and $\phi_2(t)$ and are chosen so that the entire set is complete. The coefficients are

$$r_i = \int_0^T r(t)\, \phi_i(t)\, dt; \qquad i = 1, 2, \ldots. \tag{28}$$

All of the r_i except r_1 and r_2 do not depend on which hypothesis is true and are statistically independent of r_1 and r_2. Thus a two-dimensional decision region, shown in Fig. 4.15*a*, is adequate. The mean value of r_i along each coordinate is

$$E[r_i|H_0] = \sqrt{E_0} \int_0^T s_0(t)\, \phi_i(t)\, dt, \triangleq s_{0i}, \qquad i = 1, 2, \quad :H_0, \tag{29}$$

and

$$E[r_i|H_1] = \sqrt{E_1} \int_0^T s_1(t)\, \phi_i(t)\, dt, \triangleq s_{1i}, \qquad i = 1, 2, \quad :H_1. \tag{30}$$

The likelihood ratio test follows directly from Section 2.6 (2.327)

$$\ln \Lambda = -\frac{1}{N_0} \sum_{i=1}^{2} (R_i - s_{1i})^2 + \frac{1}{N_0} \sum_{i=1}^{2} (R_i - s_{0i})^2 \underset{H_0}{\overset{H_1}{\gtrless}} \ln \eta, \tag{31a}$$

$$\ln \Lambda = -\frac{1}{N_0} |\mathbf{R} - \mathbf{s}_1|^2 + \frac{1}{N_0} |\mathbf{R} - \mathbf{s}_0|^2 \underset{H_0}{\overset{H_1}{\gtrless}} \ln \eta, \tag{31b}$$

or, canceling common terms and rearranging the result,

$$\mathbf{R}^T(\mathbf{s}_1 - \mathbf{s}_0) \underset{H_0}{\overset{H_1}{\gtrless}} \frac{N_0}{2} \ln \eta + \frac{1}{2}(|\mathbf{s}_1|^2 - |\mathbf{s}_0|^2). \tag{31c}$$

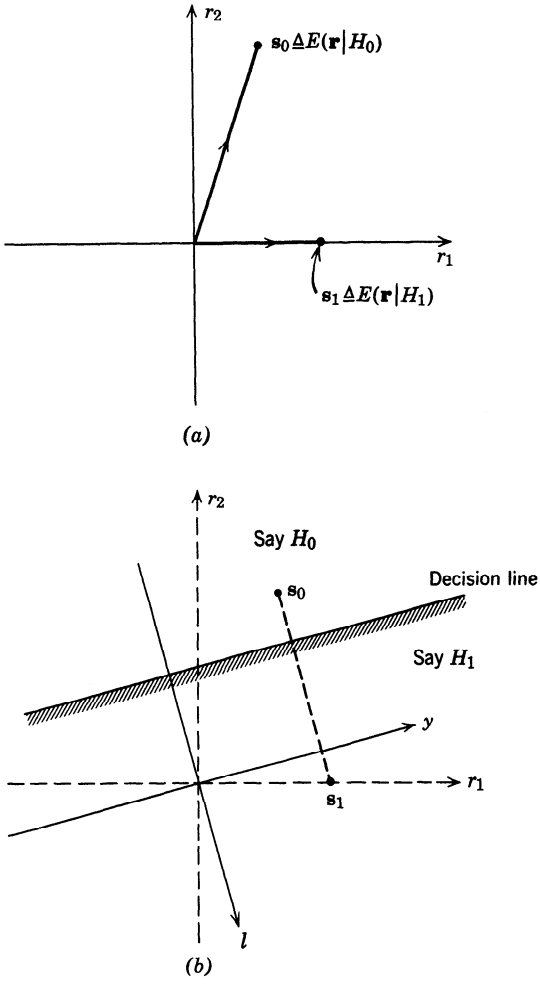

Fig. 4.15 Decision spaces.

Thus only the product of $\mathbf{R}^T$ with the difference vector $\mathbf{s}_1 - \mathbf{s}_0$ is used to make a decision. Therefore the decision space is divided into two parts by a line perpendicular to $\mathbf{s}_1$ -- $\mathbf{s}_0$ as shown in Fig. 4.15*b*. The noise components along the r_1 and r_2 axes are independent and identically distributed.

Now observe that we can transform the coordinates as shown in Fig. 4.15*b*. The noises along the new coordinates are still independent, but only the coefficient along the l coordinate depends on the hypothesis and the y coefficient may be disregarded. Therefore we can simplify our receiver by

generating l instead of r_1 and r_2. The function needed to generate the statistic is just the normalized version of the difference signal. Denote the difference signal by $s_\Delta(t)$:

$$s_\Delta(t) \triangleq \sqrt{E_1}\, s_1(t) - \sqrt{E_0}\, s_0(t). \tag{32}$$

The normalized function is

$$f_\Delta(t) = \frac{\sqrt{E_1}\, s_1(t) - \sqrt{E_0}\, s_0(t)}{(E_1 - 2\rho\sqrt{E_0E_1} + E_0)^{1/2}}. \tag{33}$$

The receiver is shown in Fig. 4.16. (Note that this result could have been obtained directly by choosing $f_\Delta(t)$ as the first orthonormal function.)

Thus once again the binary problem reduces to a one-dimensional decision space. The statistic l is Gaussian:

$$E(l|H_1) = \frac{E_1 - \sqrt{E_0E_1}\,\rho}{(E_1 - 2\rho\sqrt{E_0E_1} + E_0)^{1/2}}, \tag{34}$$

$$E(l|H_0) = \frac{\sqrt{E_0E_1}\,\rho - E_0}{(E_1 - 2\rho\sqrt{E_0E_1} + E_0)^{1/2}}. \tag{35}$$

The variance is $N_0/2$ as before. Thus

$$\boxed{d^2 = \frac{2}{N_0}(E_1 + E_0 - 2\rho\sqrt{E_0E_1}).} \tag{36}$$

Observe that if we normalized our coordinate system so that noise variance was unity then d would be the distance between the two signals. The resulting probabilities are

$$P_F = \operatorname{erfc}_*\left(\frac{\ln\eta}{d} + \frac{d}{2}\right), \tag{37}$$

$$P_D = \operatorname{erfc}_*\left(\frac{\ln\eta}{d} - \frac{d}{2}\right). \tag{38}$$

[These equations are just (2.67) and (2.68).]

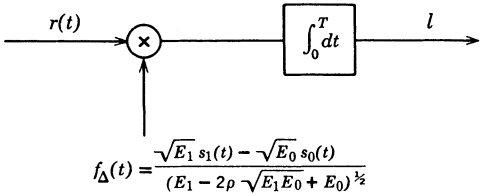

Fig. 4.16 Optimum correlation receiver, general binary problem.

The best choice of signals follows easily. The performance index d is monotonically related to the distance between the two signals in the decision space. For fixed energies the best performance is obtained by making $\rho = -1$. In other words,

$$s_0(t) = -s_1(t). \tag{39}$$

Once again the signal shape is not important.

When the criterion is minimum probability of error (as would be the logical choice in a binary communication system) *and* the a priori probabilities of the two hypotheses are equal, the decision region boundary has a simple interpretation. It is the perpendicular bisector of the line connecting the signal points (Fig. 4.17). Thus the receiver under these circumstances can be interpreted as a *minimum-distance* receiver and the error probability is

$$\Pr(\epsilon) = \int_{d/2}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx = \operatorname{erfc}_*\left(\frac{d}{2}\right). \tag{40}$$

If, *in addition*, the signals have *equal energy*, the bisector goes through the origin and we are simply choosing the signal that is most correlated with $r(t)$. This can be referred to as a "largest-of" receiver (Fig. 4.18).

The discussion can be extended to the M-ary problem in a straightforward manner.

M-ary Detection in White Gaussian Noise. Assume that there are M-hypotheses:

$$r(t) = \sqrt{E_i}\, s_i(t) + w(t); \qquad 0 \le t \le T: H_i. \tag{41}$$

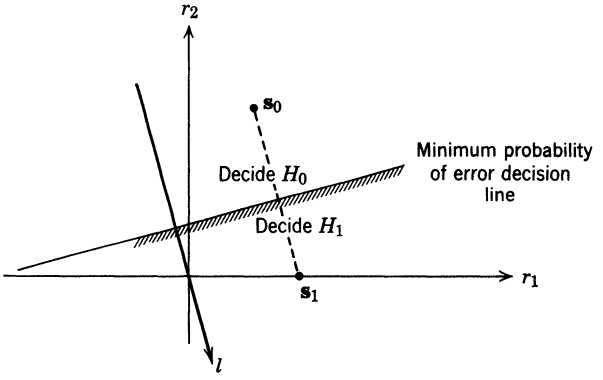

Fig. 4.17 Decision space.

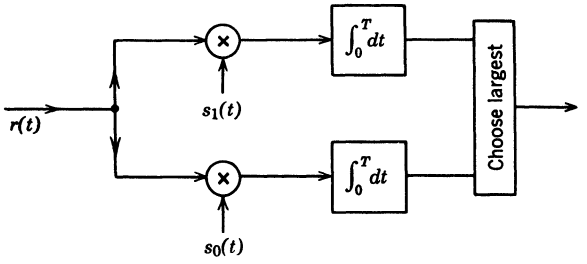

Fig. 4.18 "Largest of" receiver.

The $s_i(t)$ all have unit energy but may be correlated:

$$\int_0^T s_i(t)\, s_j(t)\, dt = \rho_{ij}. \qquad i, j = 1, 2, \ldots, M. \tag{42}$$

This problem is analogous to the M-hypothesis problem in Chapter 2. We saw that the main difficulty for a likelihood ratio test with arbitrary costs was the specification of the boundaries of the decision regions. We shall devote our efforts to finding a suitable set of sufficient statistics and evaluating the minimum probability of error for some interesting cases.

First we construct a suitable coordinate system to find a decision space with the minimum possible dimensionality. The procedure is a simple extension of the method used for two dimensions. The first coordinate function is just the first signal. The second coordinate function is that component of the second signal which is linearly independent of the first and so on. We let

$$\phi_1(t) = s_1(t), \tag{43a}$$

$$\phi_2(t) = (1 - \rho_{12}^2)^{-\frac{1}{2}}[s_2(t) - \rho_{12}\, s_1(t)]. \tag{43b}$$

To construct the third coordinate function we write

$$\phi_3(t) = c_3[s_3(t) - c_1\phi_1(t) - c_2\phi_2(t)], \tag{43c}$$

and find c_1 and c_2 by requiring orthogonality and c_3 by requiring $\phi_3(t)$ to be normalized. (This is called the Gram-Schmidt procedure and is developed in detail in Problem 4.2.7.) We proceed until one of two things happens:

1. M orthonormal functions are obtained.
2. N ($< M$) orthonormal functions are obtained and the remaining signals can be represented by linear combinations of these orthonormal functions. Thus the decision space will consist of *at most* M dimensions and fewer if the signals are linearly dependent.†

† Observe that we are talking about algebraic dependence.

We then use this set of orthonormal functions to generate N coefficients ($N \leq M$)

$$r_i \triangleq \int_0^T r(t)\,\phi_i(t)\,dt, \qquad i = 1, 2, \ldots, N. \tag{44a}$$

These are statistically independent Gaussian random variables with variance $N_0/2$ whose means depend on which hypothesis is true.

$$E[r_i|H_j] \triangleq m_{ij}, \qquad \begin{matrix} i = 1, \ldots, N, \\ j = 1, \ldots, M. \end{matrix} \tag{44b}$$

The likelihood ratio test follows directly from our results in Chapter 2 (Problem No. 2.6.1). When the criterion is minimum Pr (ϵ), we compute

$$l_j = \ln P_j - \frac{1}{N_0}\sum_{i=1}^{N}(R_i - m_{ij})^2, \qquad j = 1, \ldots, M, \tag{45}$$

and choose the largest. (The modification for other cost assignments is given in Problem No. 2.3.2.)

Two examples illustrate these ideas.

Example 1. Let

$$s_i(t) = \left(\frac{2}{T}\right)^{1/2} \sin\left[\omega_c t + (i-1)\frac{\pi}{2}\right], \qquad 0 \leq t \leq T, \qquad i = 1, 2, 3, 4,$$

$$E_i = E, \qquad i = 1, 2, 3, 4, \tag{46}$$

and

$$\omega_c = \frac{2\pi n}{T}$$

(n is an arbitrary integer). We see that

$$\phi_1(t) = \left(\frac{2}{T}\right)^{1/2} \sin \omega_c t, \qquad 0 \leq t \leq T,$$

and (47)

$$\phi_2(t) = \left(\frac{2}{T}\right)^{1/2} \cos \omega_c t, \qquad 0 \leq t \leq T.$$

We see $s_3(t)$ and $s_4(t)$ are $-\phi_1(t)$ and $-\phi_2(t)$ respectively. Thus, in this case, $M = 4$ and $N = 2$. The decision space is shown in Fig. 4.19*a*. The decision regions follow easily when the criterion is minimum probability of error and the a priori probabilities are equal. Using the result in (45), we obtain the decision regions in Fig. 4.19*b*.

Example 2. Let

$$s_1(t) = \left(\frac{2}{T}\right)^{1/2} \sin \frac{2\pi n}{T} t, \qquad 0 \leq t \leq T,$$

$$s_2(t) = \left(\frac{2}{T}\right)^{1/2} \sin \frac{4\pi n}{T} t, \qquad 0 \leq t \leq T, \tag{48}$$

$$s_3(t) = \left(\frac{2}{T}\right)^{1/2} \sin \frac{6\pi n}{T} t, \qquad 0 \leq t \leq T,$$

(n is an arbitrary integer) and

$$E_i = E, \qquad i = 1, 2, 3.$$

Now

$$\phi_i(t) = s_i(t). \tag{49}$$

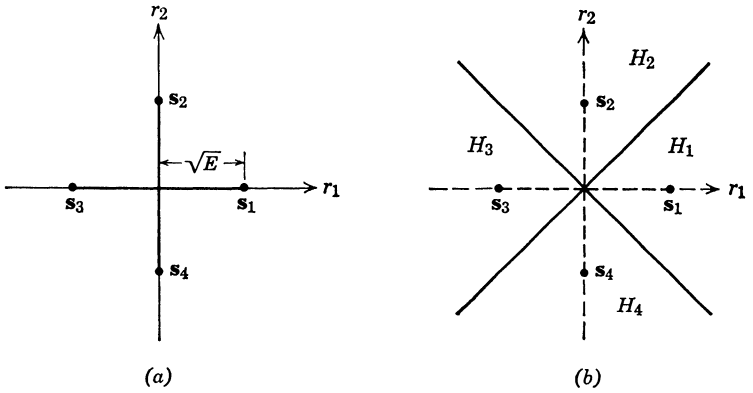

Fig. 4.19 Decision space.

In this case, $M = N = 3$ and the decision space is three-dimensional, as shown in Fig. 4.20*a*. For min Pr (ϵ) and equal a priori probabilities the decision regions follow easily from (45). The boundaries are planes perpendicular to the plane through $\mathbf{s}_1$, $\mathbf{s}_2$, and $\mathbf{s}_3$. Thus it is only the projection of $\mathbf{R}$ on this plane that is used to make a decision, and we can reduce the decision space to two dimensions as shown in Fig. 4.20*b*. (The coefficients r_1' and r_2' are along the two orthonormal coordinate functions used to define the plane.)

Note that in Examples 1 and 2 the signal sets were so simple that the Gram-Schmidt procedure was not needed.

It is clear that these results are directly analogous to the M hypothesis case in Chapter 2. As we have already seen, the calculation of the errors is conceptually simple but usually tedious for $M > 2$. To illustrate the procedure, we compute the Pr (ϵ) for Example 1.

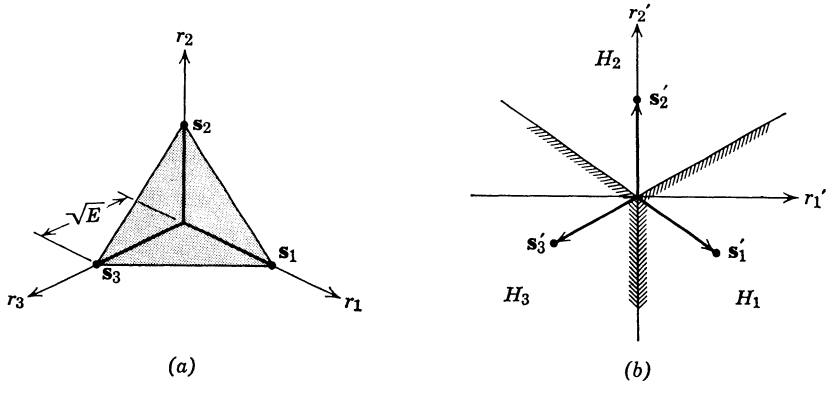

Fig. 4.20 Decision space: orthogonal signals.

Example 1 (*continued*). We assume that the hypotheses are equally likely. Now the problem is symmetrical. Thus it is sufficient to assume that $s_1(t)$ was transmitted and compute the resulting Pr (ϵ). (Clearly, Pr (ϵ) = Pr ($\epsilon|H_i$), $i = 1, \ldots, 4$.) We also observe that the answer would be invariant to a 45° rotation of the signal set because the noise is circularly symmetric.

Thus the problem of interest reduces to the simple diagram shown in Fig. 4.21.

The Pr (ϵ) is simply the probability that **r** lies outside the first quadrant when H_1 is true.

Now r_1 and r_2 are independent Gaussian variables with identical means and variances:

$$E(r_1|H_1) = E(r_2|H_1) = \left(\frac{E}{2}\right)^{1/2}$$

and

$$\text{Var}\,(r_1|H_1) = \text{Var}\,(r_2|H_1) = \frac{N_0}{2}. \tag{50}$$

The Pr (ϵ) can be obtained by integrating $p_{r_1,r_2|H_1}(R_1, R_2|H_1)$ over the area outside the first quadrant. Equivalently, Pr (ϵ) is the integral over the first quadrant subtracted from unity.

$$\text{Pr}\,(\epsilon) = 1 - \left[\int_0^\infty \left(2\pi \frac{N_0}{2}\right)^{-1/2} \exp\left(-\frac{(R_1 - \sqrt{E/2})^2}{N_0}\right) dR_1\right]^2 \tag{51}$$

Changing variables, we have

$$\text{Pr}\,(\epsilon) = 1 - \left(\int_{-\sqrt{E/N_0}}^\infty \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx\right)^2 = 1 - \left(\text{erfc}_*\left[-\left(\frac{E}{N_0}\right)^{1/2}\right]\right)^2, \tag{52}$$

which is the desired result.

Another example of interest is a generalization of Example 2.

Example 3. Let us assume that

$$r(t) = \sqrt{E}\, s_i(t) + w(t), \qquad 0 \le t \le T, \qquad H_i, \qquad i = 1, 2, \ldots, M \tag{53}$$

and

$$\rho_{ij} = \delta_{ij} \tag{54}$$

and the hypotheses are equally likely. Because the energies are equal, it is convenient to

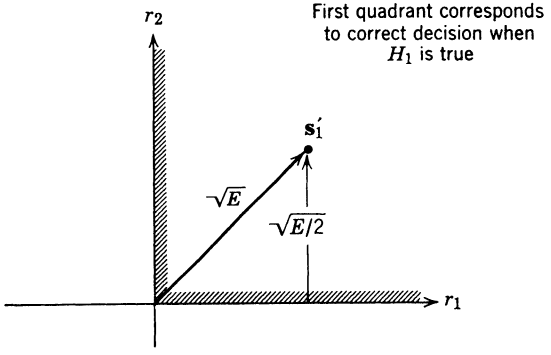

Fig. 4.21 Rotation of signal.

implement the LRT as a "greatest of" receiver as shown in Fig. 4.22. Once again the problem is symmetric, so we may assume H_1 is true. Then an error occurs if any $l_j > l_1 : j \neq 1$, where

$$l_j \triangleq \int_0^T r(t)s_j(t)\,dt, \qquad j = 1, 2, \ldots, M.$$

Thus

$$\Pr(\epsilon) = \Pr(\epsilon|H_1) = 1 - \Pr(\text{all } l_j < l_1 : j \neq 1|H_1) \tag{55}$$

or, noting that the $l_j (j \neq 1)$ have the same density on H_1,

$$\Pr(\epsilon) = 1 - \int_{-\infty}^{\infty} p_{l_1|H_1}(L_1|H_1)\left[\int_{-\infty}^{L_1} p_{l_2|H_1}(L_2|H_1)\,dL_2\right]^{M-1} dL_1. \tag{56}$$

In this particular case the densities are

$$p_{l_1|H_1}(L_1|H_1) = \frac{1}{\sqrt{\pi N_0}} \exp\left\{-\frac{1}{2}\frac{(L_1 - \sqrt{E})^2}{N_0/2}\right\} \tag{57}$$

and

$$p_{l_j|H_1}(L_j|H_1) = \frac{1}{\sqrt{\pi N_0}} \exp\left\{-\frac{1}{2}\frac{L_j^2}{N_0/2}\right\}, \qquad j \neq 1. \tag{58}$$

Substituting these densities into (56) and normalizing the variables, we obtain

$$\Pr(\epsilon) = 1 - \int_{-\infty}^{\infty} dx \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{[x - (2E/N_0)^{1/2}]^2}{2}\right\}\left(\int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{y^2}{2}\right] dy\right)^{M-1} \tag{59}$$

Unfortunately, we cannot integrate this analytically. Numerical results for certain

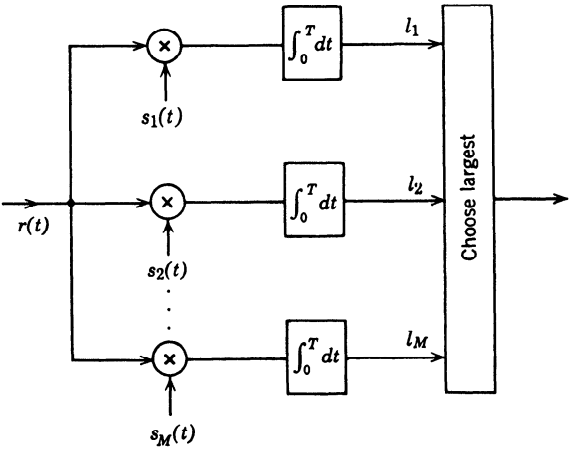

Fig. 4.22 "Largest of" receiver.

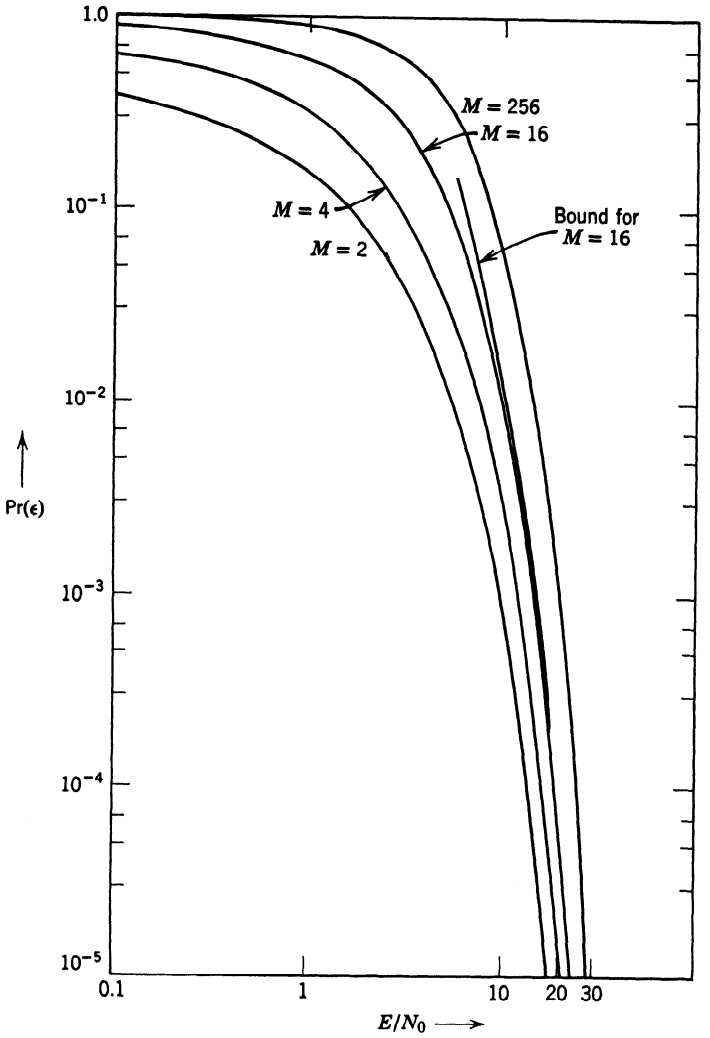

Fig. 4.23 Error probability: M orthogonal signals.

values of M and E/N_0 are tabulated in [21] and shown in Fig. 4.23. For some of our purposes an *approximate* analytic expression is more interesting. We derive a very simple bound. Some other useful bounds are derived in the problems. Looking at (55), we see that we could rewrite the Pr (ϵ) as

$$\Pr(\epsilon) = \Pr(\textit{any } l_j > l_1 : j \neq 1 \,|H_1), \tag{60}$$

$$\Pr(\epsilon) = \Pr(l_2 > l_1 \quad \text{or} \quad l_3 > l_1 \quad \text{or} \quad \cdots \quad \text{or} \quad l_M > l_1 \,|H_1). \tag{61}$$

Now, several l_j can be greater than l_1. (The events are not mutually exclusive.) Thus

$$\Pr(\epsilon) \leq \Pr(l_2 > l_1) + \Pr(l_3 > l_1) + \cdots + \Pr(l_M > l_1), \tag{62}$$

$$\Pr(\epsilon) \leq (M-1)\left\{\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{(x-\sqrt{2E/N_0})^2}{2}\right]\left(\int_x^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right) dy\right) dx\right\}; \tag{63}$$

but the term in the bracket is just the expression of the probability of error for two orthogonal signals. Using (36) with $\rho = 0$ and $E_1 = E_0 = E$ in (40), we have

$$\boxed{\Pr(\epsilon) \leq (M-1) \int_{\sqrt{E/N_0}}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right) dy.} \tag{64}$$

(Equation 64 also follows directly from (63) by a change of variables.) We can further simplify this equation by using (2.71):

$$\Pr(\epsilon) \leq \frac{(M-1)}{\sqrt{2\pi}\sqrt{E/N_0}} \exp\left(-\frac{E}{2N_0}\right). \tag{65}$$

We observe that the upper bound increases linearly with M. The bound on the $\Pr(\epsilon)$ given by this expression is plotted in Fig. 4.23 for $M = 16$.

A related problem in which M orthogonal signals arise is that of transmitting a sequence of binary digits.

Example 4. Sequence of Digits. Consider the simple digital system shown in Fig. 4.24, in which the source puts out a binary digit every T seconds. The outputs 0 and 1 are equally likely. The available transmitter power is P. For simplicity we assume that we are using orthogonal signals. The following choices are available:

1. Transmit one of two orthogonal signals every T seconds. The energy per signal is PT.
2. Transmit one of four orthogonal signals every $2T$ seconds. The energy per signal is $2PT$. For example, the encoder could use the mapping,

$$\begin{aligned} 00 &\rightarrow s_0(t), \\ 01 &\rightarrow s_1(t), \\ 10 &\rightarrow s_2(t), \\ 11 &\rightarrow s_3(t). \end{aligned}$$

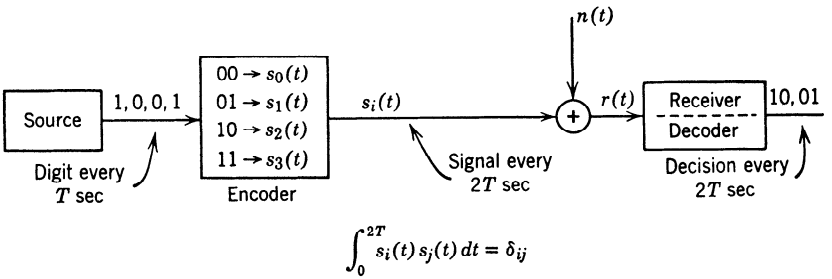

Fig. 4.24 Digital communication system.

3. In general, we could transmit one of M orthogonal signals every $T \log_2 M$ seconds. The energy per signal is $TP \log_2 M$. To compute the probability of error we use (59):

$$\Pr(\epsilon) = 1 - \int_{-\infty}^{\infty} dx \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left[x - \left(\frac{2PT \log_2 M}{N_0}\right)^{1/2}\right]^2\right\} \times \left[\int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right) dy\right]^{M-1}. \quad (66)$$

The results have been calculated numerically [19] and are plotted in Fig. 4.25. The behavior is quite interesting. Above a certain value of PT/N_0 the error probability decreases with increased M. Below this value the converse is true. It is instructive to investigate the behavior as $M \to \infty$. We obtain from (66), by a simple change of variables,

$$\lim_{M \to \infty} (1 - \Pr(\epsilon)) = \int_{-\infty}^{\infty} dy \frac{e^{-y^2/2}}{\sqrt{2\pi}} \lim_{M \to \infty} \left\{\operatorname{erf}_*^{M-1}\left[y + \left(\frac{2PT \log_2 M}{N_0}\right)^{1/2}\right]\right\}. \quad (67)$$

Now consider the limit of the logarithm of the expression in the brace:

$$\lim_{M \to \infty} \frac{\ln \operatorname{erf}_*\left[y + \left(\frac{2PT \log_2 M}{N_0}\right)^{1/2}\right]}{(M-1)^{-1}}. \quad (68)$$

Evaluating the limit by treating M as a continuous variable and using L'Hospital's rule, we find that (see Problem 4.2.15)

$$\lim_{M \to \infty} \ln\{\sim\} = \begin{cases} -\infty, & \dfrac{PT}{N_0} < \ln 2, \\ 0, & \dfrac{PT}{N_0} > \ln 2. \end{cases} \quad (69)$$

Thus, from the continuity of logarithm,

$$\lim_{M \to \infty} \Pr(\epsilon) = \begin{cases} 0, & \dfrac{PT}{N_0} > \ln 2, \\ 1, & \dfrac{PT}{N_0} < \ln 2. \end{cases} \quad (70)$$

Thus we see that there is a definite threshold effect. The value of T is determined by how fast the source produces digits. Specifically, the rate in binary digits per second is

$$R \triangleq \frac{1}{T} \text{ binary digits/sec.} \quad (71)$$

Using orthogonal signals, we see that if

$$R < \frac{1}{\ln 2} \frac{P}{N_0} \quad (72)$$

the probability of error will go to zero. The obvious disadvantage is the bandwidth requirement. As $M \to \infty$, the transmitted bandwidth goes to ∞.

The result in (72) was derived for a particular set of signals. Shannon has shown (e.g., [22] or [23]) that the right-hand side is the bound on the

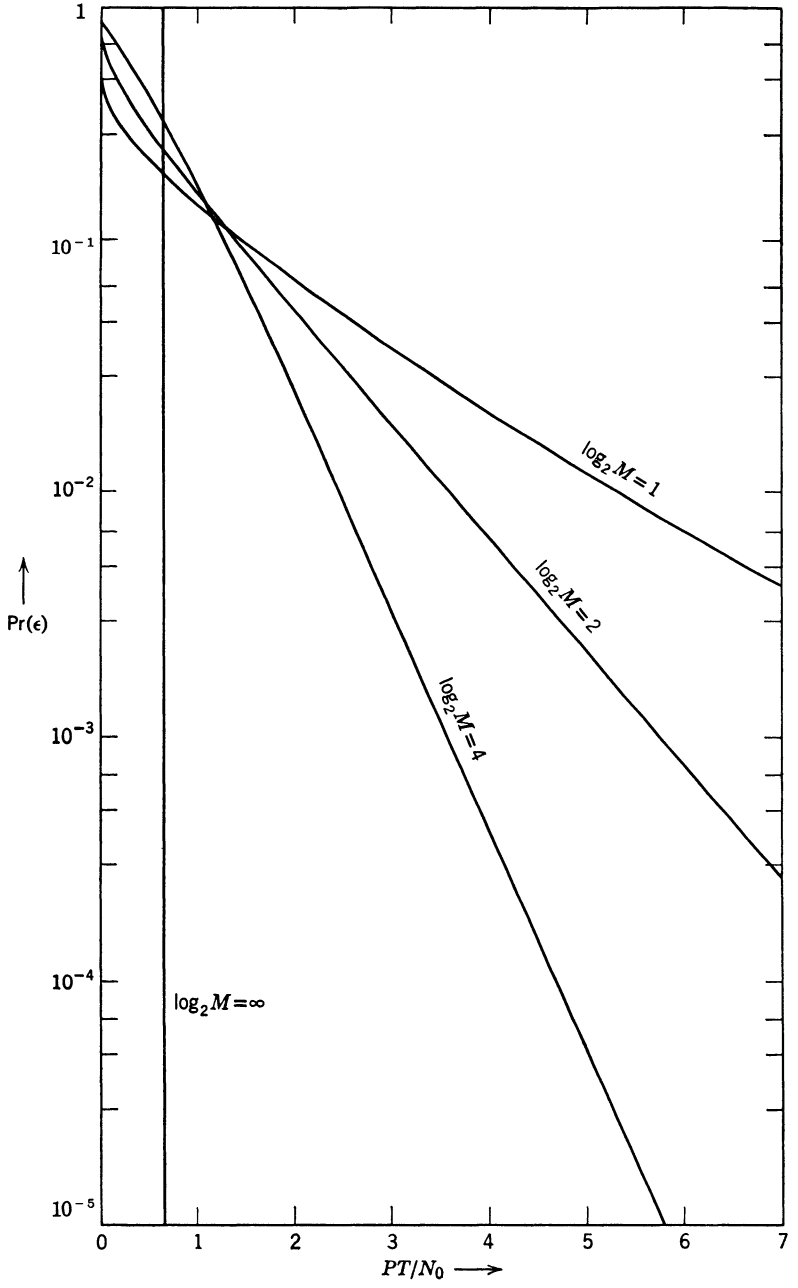

Fig. 4.25 Probability of decision error: M orthogonal signals, power constraint.

rate for error-free transmission for any communication scheme. This rate is referred to as the capacity of an infinite bandwidth, additive white Gaussian noise channel,

$$C_\infty = \frac{1}{\ln 2}\frac{P}{N_0} \text{ bits/sec.} \tag{73}$$

Shannon has also derived an expression for a bandlimited channel (W_{ch}: single-sided):

$$C = W_{ch} \log_2 \left(1 + \frac{P}{W_{ch} N_0}\right). \tag{74}$$

These capacity expressions are fundamental to the problem of sending sequences of digits. We shall not consider this problem, for an adequate discussion would take us too far afield. Suitable references are [18] and [66].

In this section we have derived the canonic receiver structures for the M-ary hypothesis problem in which the received signal under each hypothesis is a known signal plus additive white Gaussian noise. The simplicity resulted because we were always able to reduce an infinite dimensional observation space to a finite ($\leq M$) dimensional decision space.

In the problems we consider some of the implications of these results. Specific results derived in the problems include the following:

1. The probability of error for any set of M equally correlated signals can be expressed in terms of an equivalent set of M orthogonal signals (Problem 4.2.9).
2. The lowest value of uniform correlation is $-(M-1)^{-1}$. Signals with this property are optimum when there is no bandwidth restriction (Problems 4.2.9–4.2.12). They are referred to as Simplex signals.
3. For large M, orthogonal signals are essentially optimum.

Sensitivity. Before leaving the problem of detection in the presence of white noise we shall discuss an important issue that is frequently overlooked. We have been studying the mathematical model of a physical system and have assumed that we know the quantities of interest such as $s(t)$, E, and N_0 exactly. In an actual system these quantities will vary from their nominal values. It is important to determine how the performance of the optimum receiver will vary when the nominal values are perturbed. If the performance is highly sensitive to small perturbations, the validity of the nominal performance calculation is questionable. We shall discuss sensitivity in the context of the simple binary detection problem.

The model for this problem is

$$\begin{aligned} r(t) &= \sqrt{E}\, s(t) + w(t), \quad 0 \leq t \leq T: H_1 \\ r(t) &= \qquad\quad w(t), \quad 0 \leq t \leq T: H_0. \end{aligned} \tag{75}$$

The receiver consists of a matched filter followed by a decision device. The impulse response of the matched filter depends on the shape of $s(t)$. The energy and noise levels affect the decision level in the general Bayes case. (In the Neyman–Pearson case only the noise level affects the threshold setting). There are several possible sensitivity analyses. Two of these are the following:

1. Assume that the actual signal energy and signal shape are identical to those in the model. Calculate the change in P_D and P_F due to a change in the white noise level.
2. Assume that the signal energy and the noise level are identical to those in the model. Calculate the change in P_D and P_F due to a change in the signal.

In both cases we can approach the problem by first finding the change in d due to the changes in the model and then seeing how P_D and P_F are affected by a change in d. In this section we shall investigate the effect of an inaccurate knowledge of signal shape on the value of d. The other questions mentioned above are left as an exercise. We assume that we have designed a filter that is matched to the assumed signal $s(t)$,

$$h(T - t) = s(t), \qquad 0 \leq t \leq T, \tag{76}$$

and that the received waveform on H_1 is

$$r(t) = s_a(t) + w(t), \qquad 0 \leq t \leq T, \tag{77}$$

where $s_a(t)$ is the actual signal received. There are two general methods of relating $s_a(t)$ to $s(t)$. We call the first the function-variation method.

Function-Variation Method. Let

$$s_a(t) = \sqrt{E}\, s(t) + \sqrt{E_\epsilon}\, s_\epsilon(t), \qquad 0 \leq t \leq T, \tag{78}$$

where $s_\epsilon(t)$ is a normalized waveform representing the inaccuracy. The energy in the error signal is constrained to equal E_ϵ.

The effect can be most easily studied by examining the decision space (more precisely an augmented decision space). To include all of $s_\epsilon(t)$ in the decision space we *think* of adding another matched filter,

$$h_2(T - t) = \phi_2(t) = \frac{s_\epsilon(t) - \rho_\epsilon\, s(t)}{\sqrt{1 - \rho_\epsilon^2}}, \qquad 0 \leq t \leq T, \tag{79}$$

where ρ_ϵ is the correlation between $s_\epsilon(t)$ and $s(t)$. (Observe that we do not do this physically.) We now have a two-dimensional space. The effect of the constraint is clear. Any $s_a(t)$ will lead to a point on the circle surrounding $\mathbf{s}$, as shown in Fig. 4.26. Observe that the decision still uses only the

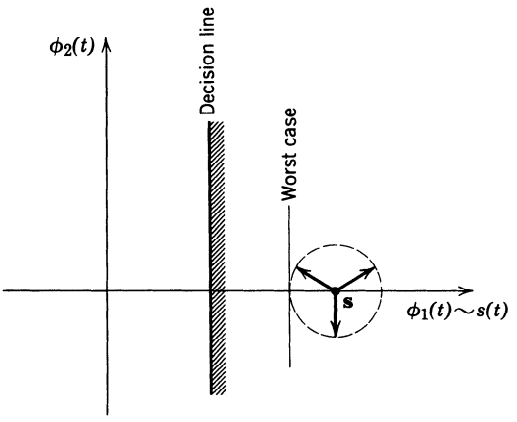

Fig. 4.26 Signal locus: fixed energy in error signal.

coordinate along $s(t)$. The effect is obvious. The error signal that causes the largest performance degradation is

$$s_\epsilon(t) = -s(t). \tag{80}$$

Then

$$d_a{}^2 = \frac{2}{N_0}(\sqrt{E} - \sqrt{E_\epsilon})^2. \tag{81}$$

To state the result another way,

$$\frac{\Delta d}{d} = -\frac{\sqrt{2E_\epsilon/N_0}}{\sqrt{2E/N_0}} = -\left(\frac{E_\epsilon}{E}\right)^{1/2} \tag{82}$$

where

$$\Delta d \triangleq d_a - d. \tag{83}$$

We see that small energy in the error signal implies a small change in performance. Thus the test is insensitive to small perturbations. The second method is called the parameter-variation method.

Parameter-Variation Method. This method can best be explained by an example. Let

$$s(t) = \left(\frac{2}{T}\right)^{1/2} \sin \omega_c t, \qquad 0 \le t \le T, \tag{84}$$

be the nominal signal. The actual signal is

$$s_a(t) = \left(\frac{2}{T}\right)^{1/2} \sin (\omega_c t + \theta), \qquad 0 \le t \le T, \tag{85}$$

which, for $\theta = 0$, corresponds to the nominal signal. The augmented

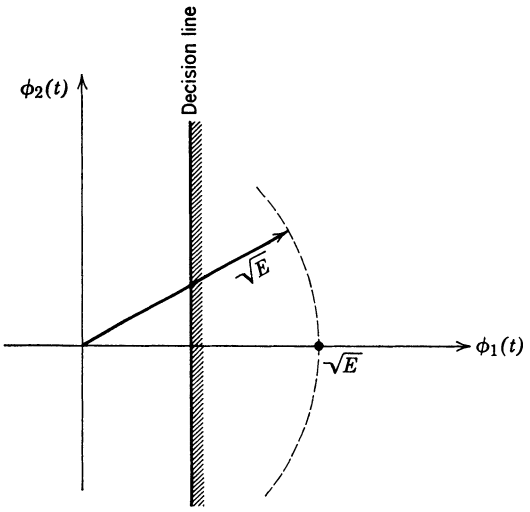

Fig. 4.27 Signal locus: fixed energy in total signal.

decision space is shown in Fig. 4.27. The vector corresponding to the actual signal moves on a circle around the origin.

$$d_a = \left(\frac{2E}{N_0}\right)^{1/2} \cos\theta \tag{86}$$

and

$$\frac{\Delta d}{d} = -(1 - \cos\theta). \tag{87}$$

Once again we see that the test is insensitive to small perturbations.

The general conclusion that we can infer from these two methods is that the results of detection in the presence of white noise are insensitive to the detailed assumptions. In other words, small perturbations from the design assumptions lead to small perturbations in performance. In almost all cases this type of insensitivity is necessary if the mathematical model is going to predict the actual system performance accurately.

Many statistical analyses tend to ignore this issue. The underlying reason is probably psychological. After we have gone through an involved mathematical optimization, it would be pleasant to demonstrate an order-of-magnitude improvement over the system designed by using an intuitive approach. Unfortunately, this does not always happen. When it does, we must determine whether the mathematical result is sensitive to some

detailed assumption. In the sequel we shall encounter several examples of this sensitivity.

We now turn to the problem of linear estimation.

4.2.2 Linear Estimation

In Section 4.1.1 we formulated the problem of estimating signal parameters in the presence of additive noise. For the case of additive white noise the received waveform is

$$r(t) = s(t, A) + w(t), \qquad 0 \leq t \leq T, \tag{88a}$$

where $w(t)$ is a sample function from a white Gaussian noise process with spectral height $N_0/2$. The parameter A is the quantity we wish to estimate. If it is a random parameter we will assume that the a priori density is known and use a Bayes estimation procedure. If it is a nonrandom variable we will use ML estimates. The function $s(t, A)$ is a deterministic mapping of A into a time function. If $s(t, A)$ is a linear mapping (in other words, superposition holds), we refer to the system using the signal as a *linear signaling* (or *linear modulation*) system. Furthermore, for the criterion of interest the estimator will turn out to be linear so we refer to the problem as a *linear estimation* problem. In this section we study linear estimation and in Section 4.2.3, nonlinear estimation. For linear modulation (88*a*) can always be written as

$$r(t) = A\sqrt{E}\, s(t) + w(t), \qquad 0 \leq t \leq T, \tag{88b}$$

where $s(t)$ has unit energy.

We can solve the linear estimation problem easily by exploiting its similarity to the detection problem that we just solved. From Section 2.4 we know that the likelihood function is needed. We recall, however, that the problem is greatly simplified if we can find a sufficient statistic and work with it instead of the received waveform. If we compare (88*b*) and (4)–(7), it is clear that a sufficient statistic is r_1, where

$$r_1 = \int_0^T r(t)\, s(t)\, dt. \tag{89}$$

Just as in Section 4.2.1, the probability density of r_1, given $a = A$, is Gaussian:

$$\begin{aligned} E(r_1|A) &= A\sqrt{E}, \\ \text{Var}\,(r_1|A) &= \frac{N_0}{2}. \end{aligned} \tag{90}$$

It is easy to verify that the coefficients along the other orthogonal functions

[see (8)] are independent of a. Thus the waveform problem reduces to the classical estimation problem (see pp. 58–59).

The logarithm of the likelihood function is

$$l(A) = -\frac{1}{2}\frac{(R_1 - A\sqrt{E})^2}{N_0/2}. \tag{91}$$

If A is a nonrandom variable, the ML estimate is the value of A at which this function is a maximum. Thus

$$\hat{a}_{ml}(R_1) = \frac{R_1}{\sqrt{E}}. \tag{92}$$

The receiver is shown in Fig. 4.28*a*. We see the estimate is unbiased.

If a is a random variable with a probability density $p_a(A)$, then the MAP estimate is the value of A where

$$l_p(A) = -\frac{1}{2}\frac{(R_1 - A\sqrt{E})^2}{N_0/2} + \ln p_a(A), \tag{93}$$

is a maximum. For the special case in which a is Gaussian, $N(0, \sigma_a)$, the

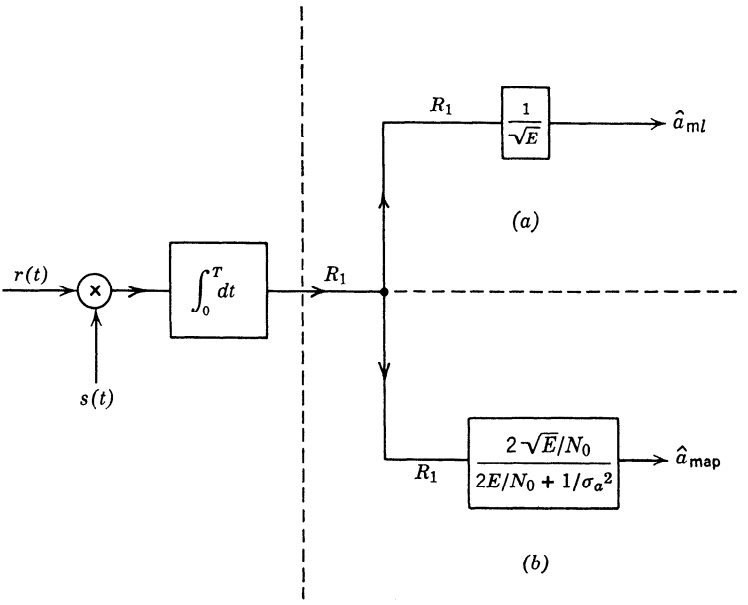

Fig. 4.28 Optimum receiver, linear estimation: (*a*) ML receiver; (*b*) MAP receiver.

MAP estimate is easily obtained by differentiating $l_p(A)$ and equating the result to zero:

$$\frac{\partial l_p(A)}{\partial A} = \frac{R_1 - A\sqrt{E}}{N_0/2}\sqrt{E} - \frac{A}{\sigma_a{}^2} \tag{94}$$

and

$$\hat{a}_{\text{map}}(R_1) = \frac{2E/N_0}{2E/N_0 + 1/\sigma_a{}^2}\frac{R_1}{\sqrt{E}}. \tag{95}$$

In both the ML and MAP cases it is easy to show that the result is the absolute maximum.

The MAP receiver is shown in Fig. 4.28*b*. Observe that the only difference between the two receivers is a gain. The normalized error variances follow easily: for MAP

$$\frac{E[a_\epsilon{}^2]}{\sigma_a{}^2} = \sigma_{a\epsilon n}^2 \triangleq \frac{\sigma_{a_\epsilon}{}^2}{\sigma_a{}^2} = \left(1 + \frac{2\sigma_a{}^2 E}{N_0}\right)^{-1} \qquad \text{(MAP)}. \tag{96}$$

The quantity $\sigma_a{}^2 E$ is the expected value of the received energy. For ML

$$\sigma_{a_\epsilon n}^2 \triangleq \frac{\sigma_{a_\epsilon}{}^2}{A^2} = \left(\frac{2A^2E}{N_0}\right)^{-1} \qquad \text{(ML)}. \tag{97}$$

Here A^2E is the actual value of the received energy. We see that the variance of the maximum likelihood estimate is the reciprocal of d^2, the performance index of the simple binary problem. In both cases we see that the only way to decrease the mean-square error is to increase the energy-to-noise ratio. In many situations the available energy-to-noise ratio is not adequate to provide the desired accuracy. In these situations we try a nonlinear signaling scheme in an effort to achieve the desired accuracy. In the next section we discuss the nonlinear estimation.

Before leaving linear estimation, we should point out that the MAP estimate is also the Bayes estimate for a large class of criteria. Whenever a is Gaussian the a posteriori density is Gaussian and Properties 1 and 2 on pp. 60–61 are applicable. This invariance to criterion depends directly on the linear signaling model.

4.2.3 Nonlinear Estimation

The system in Fig. 4.7 illustrates a typical nonlinear estimation problem. The received signal is

$$r(t) = s(t, A) + w(t), \qquad 0 \leq t \leq T. \tag{98}$$

From our results in the classical case we know that a sufficient statistic does not exist in general. As before, we can construct the likelihood

function. We approach the problem by making a K-coefficient approximation to $r(t)$. By proceeding as on p. 252 with obvious notation we have

$$\Lambda[r_K(t), A] = p_{\mathbf{r}|a}(\mathbf{R}|A) = \prod_{i=1}^{K} \frac{1}{\sqrt{\pi N_0}} \exp\left(-\frac{1}{2}\frac{[R_i - s_i(A)]^2}{N_0/2}\right). \quad (99)$$

where

$$s_i(A) \triangleq \int_0^T s(t, A)\, \phi_i(t)\, dt.$$

Now, if we let $K \to \infty$, $\Lambda[r_K(t), A]$ is not well defined. We recall from Chapter 2 that we can divide a likelihood function by anything that does not depend on A and still have a likelihood function. On p. 252 we avoided the convergence problem by dividing by

$$p_{r_K(t)|H_0}[r_K(t)|H_0] = \prod_{i=1}^{K} \frac{1}{\sqrt{\pi N_0}} \exp\left(-\frac{1}{2}\frac{R_i^2}{N_0/2}\right),$$

before letting $K \to \infty$. Because this function does not depend on A, it is legitimate to divide by it here. Define

$$\Lambda_1[r_K(t), A] = \frac{\Lambda[r_K(t), A]}{p_{r_K(t)|H_0}[r_K(t)|H_0]}. \quad (100)$$

Substituting into this expression, canceling common terms, letting $K \to \infty$, and taking the logarithm we obtain

$$\ln \Lambda_1[r(t), A] = \frac{2}{N_0}\int_0^T r(t)\, s(t, A)\, dt - \frac{1}{N_0}\int_0^T s^2(t, A)\, dt. \quad (101)$$

To find $\hat{a}_{ml}$ we must find the absolute maximum of this function. To find $\hat{a}_{\text{map}}$ we add $\ln p_a(A)$ to (101) and find the absolute maximum. The basic operation on the received data consists of generating the first term in (101) as a function of A. The physical device that we actually use to accomplish it will depend on the functional form of $s(t, A)$. We shall consider some specific cases and find the actual structure.

Before doing so we shall derive a result for the general case that will be useful in the sequel. Observe that if the maximum is interior and $\ln \Lambda_1(A)$ is differentiable at the maximum, then a necessary, but not sufficient, condition is obtained by first differentiating (101):

$$\frac{\partial \ln \Lambda_1(A)}{\partial A} = \frac{2}{N_0}\int_0^T [r(t) - s(t, A)]\frac{\partial s(t, A)}{\partial A}\, dt \quad (102)$$

(assuming that $s(t, A)$ is differentiable with respect to A). For $\hat{a}_{ml}$, a necessary condition is obtained by setting the right-hand side of (102) equal to zero. For $\hat{a}_{\text{map}}$ we add $d \ln p_a(A)/dA$ to the right-hand side of (102)

and set the sum equal to zero. In the special case in which $p_a(A)$ is Gaussian, $N(0, \sigma_a)$, we obtain

$$\hat{a}_{\text{map}} = \frac{2\sigma_a^2}{N_0} \int_0^T [r(t) - s(t, A)] \frac{\partial s(t, A)}{\partial A} \, dt|_{A=\hat{a}_{\text{map}}}. \tag{103}$$

In the linear case (103) reduces to (95) and gives a unique solution. A number of solutions may exist in the nonlinear case and we must examine the sum of (101) and $\ln p_a(A)$ to guarantee an absolute maximum.

However, just as in Chapter 2, (102) enables us to find a bound on the variance of any unbiased estimate of a nonrandom variable and the addition of $d^2 \ln p_a(A)/dA^2$ leads to a bound on the mean-square error in estimating a random variable. For nonrandom variables we differentiate (102) and take the expectation

$$E\left[\frac{\partial^2 \ln \Lambda_1(A)}{\partial A^2}\right] = \frac{2}{N_0}\left\{E\int_0^T [r(t) - s(t, A)] \frac{\partial^2 s(t, A)}{\partial A^2} \, dt - E\int_0^T \left[\frac{\partial s(t, A)}{\partial A}\right]^2 dt\right\}, \tag{104}$$

where we assume the derivatives exist. In the first term we observe that

$$E[r(t) - s(t, A)] = E[w(t)] = 0. \tag{105}$$

In the second term there are no random quantities; therefore the expectation operation gives the integral itself.

Substituting into (2.179), we have

$$\text{Var}\,(\hat{a} - A) \geq \frac{N_0}{2\displaystyle\int_0^T \left[\frac{\partial s(t, A)}{\partial A}\right]^2 dt} \tag{106}$$

for any unbiased estimate $\hat{a}$. Equality holds in (106) if and only if

$$\frac{\partial \ln \Lambda_1(A)}{\partial A} = k(A)\{\hat{a}[r(t)] - A\} \tag{107}$$

for all A and $r(t)$. Comparing (102) and (107), we see that this will hold only for linear modulation. Then $\hat{a}_{ml}$ is the minimum variance estimate.

Similarly, for random variables

$$E[\hat{a} - a]^2 \geq \left(E_a\left\{\frac{2}{N_0}\int_0^T \left[\frac{\partial s(t, A)}{\partial A}\right]^2 dt - \frac{d^2 \ln p_a(A)}{dA^2}\right\}\right)^{-1}, \tag{108}$$

where E_a denotes an expectation over the random variable a. Defining

$$\gamma_a^2 \triangleq E_a \int_0^T \left[\frac{\partial s(t, A)}{\partial A}\right]^2 dt, \tag{109}$$

we have

$$E(\hat{a} - a)^2 \geq \left(\frac{2}{N_0}\gamma_a^2 - E_a\left[\frac{d^2 \ln p_a(A)}{dA^2}\right]\right)^{-1} \tag{110}$$

Equality will hold if and only if (see 2.226)

$$\frac{\partial^2 \ln \Lambda_1(A)}{\partial A^2} + \frac{d^2 \ln p_a(A)}{dA^2} = \text{constant.} \tag{111}$$

Just as in Chapter 2 (p. 73), in order for (111) to hold it is necessary and sufficient that $p_{a|r(t)}[A|r(t) : 0 \leq t \leq T]$ be a Gaussian probability density. This requires a linear signaling scheme and a Gaussian a priori density.

What is the value of the bound if it is satisfied only for linear signaling schemes?

As in the classical case, it has two principal uses:

1. It always provides a lower bound.
2. In many cases, the actual variance (or mean-square error) in a nonlinear signaling scheme will approach this bound under certain conditions. These cases are the analogs to the *asymptotically efficient* estimates in the classical problem. We shall see that they correspond to large E/N_0 values.

To illustrate some of the concepts in the nonlinear case we consider two simple examples.

Example 1. Let $s(t)$ be the pulse shown in Fig. 4.29*a*. The parameter a is the arrival

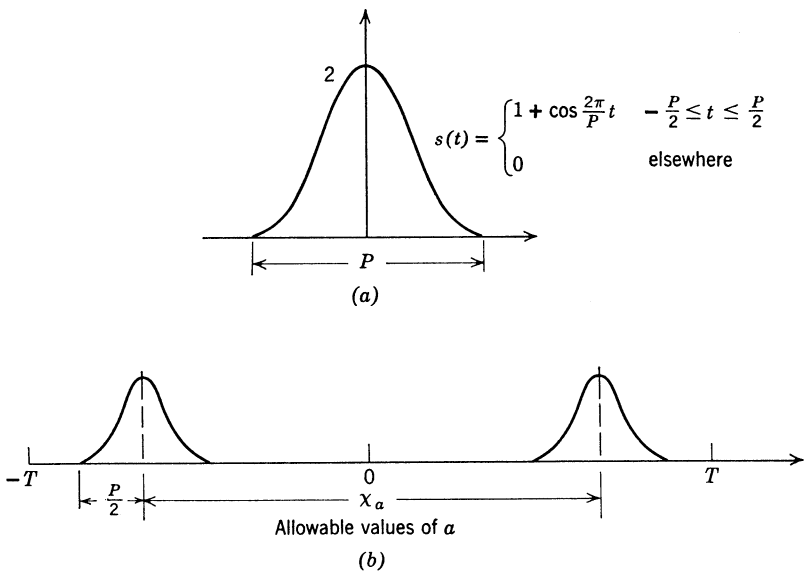

Fig. 4.29 (*a*) Pulse shape; (*b*) Allowable parameter range.

time of the pulse. We want to find a MAP estimate of a. We know a range of values χ_a that a may assume (Fig. 4.29*b*). Inside this range the probability density is uniform. For simplicity we let the observation interval $[-T, T]$ be long enough to completely contain the pulse.

From (101) we know that the operation on the received waveform consists of finding $\ln \Lambda_1[r(t), A]$. Here

$$\ln \Lambda_1[r(t), A] = \frac{2}{N_0}\int_{-T}^{T} r(u)\, s(u - A)\, du - \frac{1}{N_0}\int_{-T}^{T} s^2(u - A)\, du. \tag{112}$$

For this particular case the second term does not depend on A, for the entire pulse is always in the interval. The first term is a convolution operation. The output of a linear filter with impulse response $h(\tau)$ and input $r(u)$ over the interval $[-T, T]$ is

$$y(t) = \int_{-T}^{T} r(u)\, h(t - u)\, du, \qquad -T \leq t \leq T \tag{113}$$

Clearly, if we let

$$h(\tau) = s(-\tau), \tag{114}$$

the output as a function of time over the range x_a will be identical to the likelihood function as a function of A. We simply pick the peak of the filter output as a function of time. The time at which the peak occurs is $\hat{a}_{\text{map}}$ .The filter is the matched filter that we have already encountered in the detection problem.

In Fig. 4.30 we indicate the receiver structure. The output due to the signal component is shown in line (*a*). Typical total outputs for three noise levels are shown in lines (*b*), (*c*), and (*d*). In line (*b*) we see that the peak of $\ln \Lambda(A)$ is large compared to

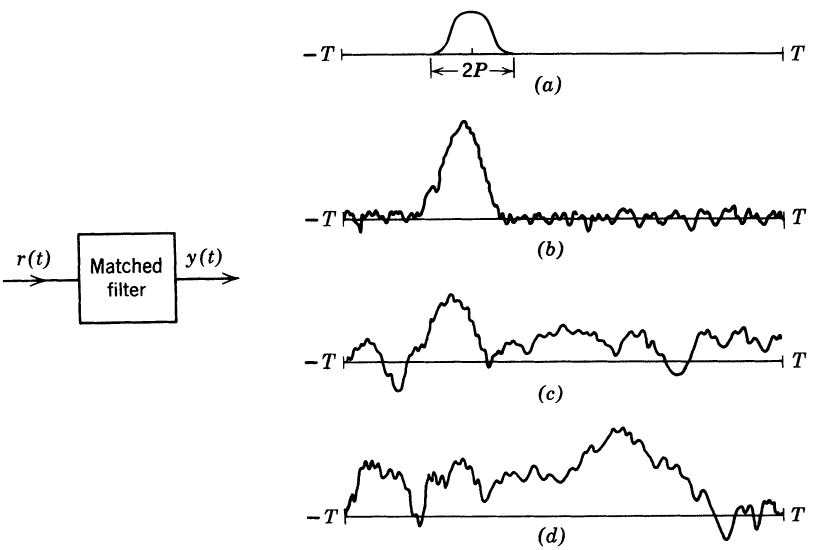

Fig. 4.30 Receiver outputs [arrival time estimation]: (*a*) signal component; (*b*) low noise level; (*c*) moderate noise level; (*d*) high noise level.

the noise background. The actual peak is near the correct peak, and we can expect that the error will be accurately predicted by using the expression in (110). In line (*c*) the noise has increased and large subsidiary peaks which have no relation to the correct value of A are starting to appear. Finally, in line (*d*) the noise has reached the point at which the maximum bears no relation to the correct value. Thus two questions are posed:

1. Under what conditions does the lower bound given by (110) accurately predict the error?
2. How can one predict performance when (110) is not useful?

Before answering these questions, we consider a second example to see if similar questions arise.

Example 2. Another common example of a nonlinear signaling technique is discrete frequency modulation (also referred to as pulse frequency modulation, PFM). Every T seconds the source generates a new value of the parameter a. The transmitted signal is $s(t, A)$, where

$$s(t, A) = \left(\frac{2E}{T}\right)^{\frac{1}{2}} \sin(\omega_c + \beta A)t, \qquad -\frac{T}{2} \leq t \leq \frac{T}{2}. \tag{115}$$

Here ω_c is a known carrier frequency, β is a known constant, and E is the transmitted energy (also the received signal energy). We assume that $p_a(A)$ is a uniform variable over the interval $(-\sqrt{3}\,\sigma_a, \sqrt{3}\,\sigma_a)$.

To find $\hat{a}_{\text{map}}$ we construct the function indicated by the first term in (101): (The second term in (101) and the a priori density are constant and may be discarded.)

$$\begin{aligned} l_1(A) &= \int_{-T/2}^{T/2} r(t) \sin(\omega_c t + \beta A t)\, dt, \qquad -\sqrt{3}\,\sigma_a \leq A \leq \sqrt{3}\,\sigma_a \\ &= 0, \qquad\qquad \text{elsewhere.} \end{aligned} \tag{116}$$

One way to construct $l_1(A)$ would be to record $r(t)$ and perform the multiplication and integration indicated by (116) for successive values of A over the range. This is obviously a time-consuming process. An alternate approach† is to divide the range into increments of length Δ and perform the parallel processing operation shown in Fig. 4.31 for discrete values of A:

$$\begin{aligned} A_1 &= -\sqrt{3}\,\sigma_a + \frac{\Delta}{2}, \\ A_2 &= -\sqrt{3}\,\sigma_a + \frac{3\Delta}{2}, \\ &\vdots \qquad\qquad M = \left\lfloor \frac{2\sqrt{3}\,\sigma_a}{\Delta} + \frac{1}{2} \right\rfloor, \\ A_M &= -\sqrt{3}\,\sigma_a + (M - \tfrac{1}{2})\,\Delta, \end{aligned} \tag{117}$$

† This particular type of approach and the resulting analysis were first done by Woodward (radar range measurement [8]) and Kotelnikov (PPM and PFM, [13]). Subsequently, they have been used with various modifications by a number of authors (e.g., Darlington [87], Akima [88], Wozencraft and Jacobs [18], Wainstein and Zubakov [15]). Our approach is similar to [18]. A third way to estimate A is discussed in [87]. (See also Chapter II.4.)

where $[\cdot]$ denotes the largest integer smaller than or equal to the argument. The output of this preliminary processing is M numbers. We choose the largest and assume that the correct value of A is in that region.

To get the final estimate we conduct a local maximization by using the condition

$$\int_{-T/2}^{T/2} [r(t) - s(t, A)] \frac{\partial s(t, A)}{\partial A} dt|_{A=\hat{a}_{map}} = 0. \tag{118}$$

(This assumes the maximum is interior.)

A possible way to accomplish this maximization is given in the block diagram in Fig. 4.32. We expect that if we chose the correct interval in our preliminary processing the final accuracy would be closely approximated by the bound in (108). This bound can be evaluated easily. The partial derivative of the signal is

$$\frac{\partial s(t, A)}{\partial A} = \left(\frac{2E}{T}\right)^{1/2} \beta t \cos(\omega_c t + \beta A t), \qquad -T/2 \leq t \leq T/2, \tag{119}$$

and

$$\gamma_a^2 = \frac{2E}{T} \beta^2 \int_{-T/2}^{T/2} \overline{t^2 \cos^2(\omega_c t + \beta A t)}\, dt \simeq \frac{ET^2}{12} \beta^2, \tag{120}$$

when

$$T \gg \frac{1}{\omega_c}.$$

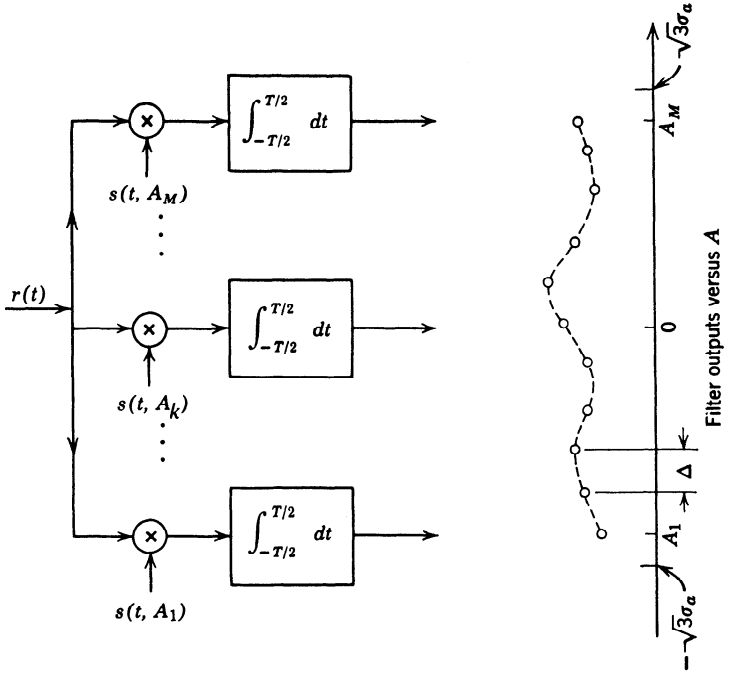

Fig. 4.31 Receiver structure [frequency estimation].

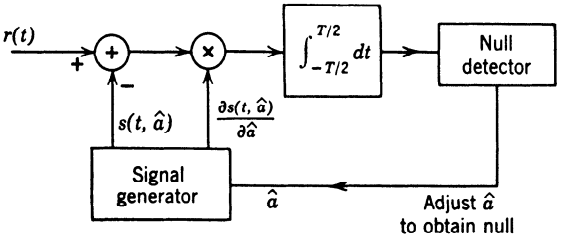

Fig. 4.32 Local estimator.

Then the normalized mean-square error of any estimate is bounded by

$$\sigma_{a_\epsilon n}^2 \triangleq \frac{\sigma_\epsilon^2}{\sigma_a^2} \geq \frac{N_0}{2\gamma_a^2 \sigma_a^2} = \frac{12}{T^2} \frac{N_0}{2E} \frac{1}{\beta^2} \frac{1}{\sigma_a^2}, \tag{121}$$

which seems to indicate that, regardless of how small E/N_0 is, we can make the mean-square error arbitrarily small by increasing β. Unfortunately, this method neglects an important part of the problem. How is the probability of an initial interval error affected by the value of β?

With a few simplifying assumptions we can obtain an approximate expression for this probability. We denote the actual value of A as A_a. (This subscript is necessary because A is the argument in our likelihood function.) A plot of

$$\frac{1}{E} \int_{-T/2}^{T/2} s(t, A_a)\, s(t, A)\, dt$$

for the signal in (115) as a function of $A_x \triangleq A - A_a$ is given in Fig. 4.33. (The double frequency term is neglected.) We see that the signal component of $l_1(A)$ passes through zero every $2\pi/\beta T$ units. This suggests that a logical value of Δ is $2\pi/\beta T$.

To calculate the probability of choosing the wrong interval we use the approximation that we can replace all A in the first interval by A_1 and so forth. We denote the probability of choosing the wrong interval as $\Pr(\epsilon_I)$. With this approximation the problem is reduced to detecting which of M orthogonal, equal energy signals is present. For large M we neglect the residual at the end of the interval and let

$$M \cong \sqrt{3}\, \sigma_a \beta \frac{T}{\pi}; \tag{122}$$

but this is a problem we have already solved (64). Because large βT is the case of interest, we may use the approximate expression in (65):

$$\Pr(\epsilon_I) \leq \frac{(\sqrt{3}\, \sigma_a \beta T/\pi - 1)}{\sqrt{2\pi E/N_0}} \exp\left(-\frac{E}{2N_0}\right). \tag{123}$$

We see that as $\sigma_a \beta T$ increases the probability that we will choose the wrong interval also increases.† The conclusion that can be inferred from this result is of fundamental importance in the nonlinear estimation problem.

For a fixed E/N_0 and T we can increase β so that the local error will be arbitrarily small *if* the receiver has chosen the correct interval. As β increases, however, the

† This probability is sometimes referred to as the probability of anomaly or ambiguity.

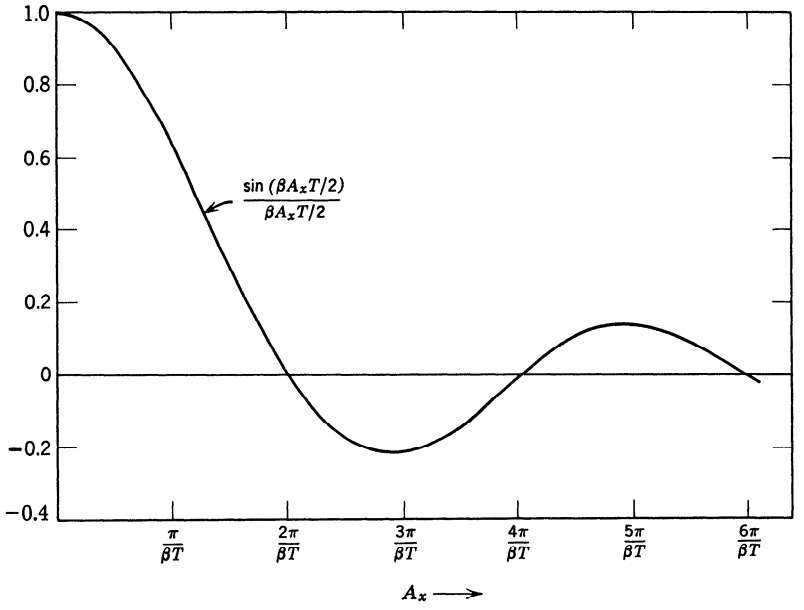

Fig. 4.33 Signal component vs. A_x.

probability that we will be in the correct interval goes to zero. Thus, for a particular β, we must have some minimum E/N_0 to ensure that the probability of being in the wrong interval is adequately small.

The expression in (123) suggests the following design procedure. We decide that a certain Pr (ϵ_I) (say p_0) is acceptable. In order to minimize the mean-square error subject to this constraint we choose β such that (123) is satisfied with equality. Substituting p_0 into the left side of (123), solving for $\sigma_a \beta T$, and substituting the result in (121), we obtain

$$\frac{E[a_\epsilon^2]}{\sigma_a^2} = \sigma_{a_\epsilon n}^2 \simeq \frac{1}{p_0^2}\frac{9}{\pi^3}\left(\frac{N_0}{E}\right)^2 e^{-E/N_0}. \tag{124}\dagger$$

The reciprocal of the normalized mean-square error as a function of E/N_0 for typical values of p_0 is shown in Fig. 4.34. For reasons that will become obvious shortly, we refer to the constraint imposed by (123) as a threshold constraint.

The result in (123) indicates one effect of increasing β. A second effect can be seen directly from (115). Each value of A shifts the frequency of the transmitted signal from ω_c to $\omega_c + \beta A$. Therefore we must have enough bandwidth available in the channel to accommodate the maximum possible frequency excursion. The pulse

† Equation 124 is an approximation, for (123) is a bound and we neglected the 1 in the parentheses because large βT is the case of interest.

bandwidth is approximately $2\pi/T$ rad/sec. The maximum frequency shift is $\pm\sqrt{3}\,\beta\sigma_a$. Therefore the required channel bandwidth centered at ω_c is approximately

$$2\pi W_{ch} \cong 2\sqrt{3}\,\beta\sigma_a + \frac{2\pi}{T} = \frac{1}{T}(2\sqrt{3}\,\beta\sigma_a T + 2\pi) \tag{125a}$$

When $\sigma_a\beta T$ is large we can neglect the 2π and use the approximation

$$2\pi W_{ch} \simeq 2\sqrt{3}\,\beta\sigma_a. \tag{125b}$$

In many systems of interest we have only a certain bandwidth available. (This bandwidth limitation may be a legal restriction or may be caused by the physical nature of the channel.) If we assume that E/N_0 is large enough to guarantee an acceptable Pr (ϵ_I), then (125*b*) provides the constraint of the system design. We simply increase β until the available bandwidth is occupied. To find the mean-square error using this design procedure we substitute the expression for $\beta\sigma_a$ in (125*b*) into (121) and obtain

$$\frac{E[a^2\epsilon]}{\sigma_a{}^2} = \sigma^2_{a\epsilon n} = \frac{18}{\pi^2}\frac{N_0}{E}\frac{1}{(W_{ch}T)^2} \qquad \text{(bandwidth constraint).} \tag{126}$$

We see that the two quantities that determine the mean-square error are E/N_0, the energy-to-noise ratio, and $W_{ch}T$, which is proportional to the time-bandwidth product of the transmitted pulse. The reciprocal of the normalized mean-square error is plotted in Fig. 4.34 for typical values of $W_{ch}T$.

The two families of constraint lines provide us with a complete design procedure for a PFM system. For low values of E/N_0 the threshold constraint dominates. As E/N_0 increases, the MMSE moves along a fixed p_0 line until it reaches a point where the available bandwidth is a constraint. Any further increase in E/N_0 moves the MMSE along a fixed β line.

The approach in which we consider two types of error separately is useful and contributes to our understanding of the problem. To compare the results with other systems it is frequently convenient to express them as a single number, the over-all mean-square error.

We can write the mean-square error as

$$\begin{aligned} E(a_\epsilon{}^2) = \sigma_{\epsilon_T}{}^2 = {} & E[a_\epsilon{}^2|\text{interval error}]\ \Pr\ [\text{interval error}] \\ & + E[a_\epsilon{}^2|\text{no interval error}]\ \Pr\ [\text{no interval error}] \end{aligned} \tag{127}$$

We obtained an approximation to Pr (ϵ_I) by collecting each incremental range of A at a single value A_i. With this approximation there is no signal component at the other correlator outputs in Fig. 4.31. Thus, if an interval error is made, it is equally likely to occur in any one of the wrong intervals. Therefore the resulting estimate $\hat{a}$ will be uncorrelated with a.

$$\begin{aligned} E[a_\epsilon{}^2|\text{interval error}] &= E[(\hat{a} - a)^2|\text{interval error}] \\ &= E[\hat{a}^2|\text{interval error}] + E[a^2|\text{interval error}] \\ &\quad - 2E[\hat{a}a|\text{interval error}]. \end{aligned} \tag{128}$$

Our approximation makes the last term zero. The first two terms both equal $\sigma_a{}^2$. Therefore

$$E[a_\epsilon{}^2|\text{interval error}] = 2\sigma_a{}^2. \tag{129}$$

If we assume that p_0 is fixed, we then obtain by using (124) and (129) in (127)

$$\sigma^2_{\epsilon_T n} = \frac{E(a_\epsilon{}^2)}{\sigma_a{}^2} = 2p_0 + (1 - p_0)\frac{9}{\pi^3 p_0^2}\left(\frac{N_0}{E}\right)^2 e^{-E/N_0}. \tag{130}$$

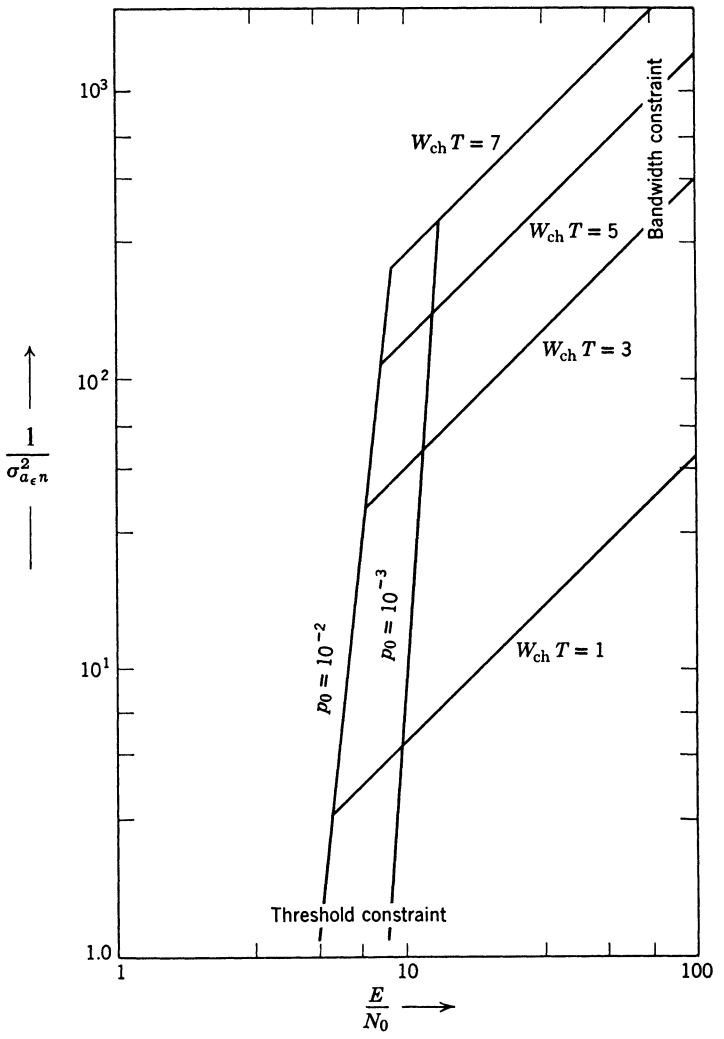

Fig. 4.34 Reciprocal of the mean-square error under threshold and bandwidth constraints.

In this case the modulation index β must be changed as E/N_0 is changed. For a fixed β we use (121) and (123) to obtain

$$\sigma^2_{\epsilon_T n} = \frac{12}{(\sigma_a \beta T)^2} \frac{N_0}{2E} \left[1 - \frac{\sqrt{3}\,\sigma_a \beta T/\pi}{\sqrt{2\pi E/N_0}} e^{-E/2N_0} \right] + 2 \frac{\sqrt{3}\,\sigma_a \beta T/\pi}{\sqrt{2\pi E/N_0}} e^{-E/2N_0}. \tag{131}$$

The result in (131) is plotted in Fig. 4.35, and we see that the mean-square error

exhibits a definite threshold. The reciprocal of the normalized mean-square error for a PAM system is also shown in Fig. 4.35 (from 96). The magnitude of this improvement can be obtained by dividing (121) by (96).

$$\frac{\sigma^2_{a\epsilon n|\text{PFM}}}{\sigma^2_{a\epsilon n|\text{PAM}}} \cong \frac{12}{\beta^2 T^2}, \qquad \frac{2\sigma_a{}^2 E}{N_0} \gg 1.$$

Thus the improvement obtained from PFM is proportional to the square of βT. It is important to re-emphasize that this result assumes E/N_0 is such that the system is above threshold. If the noise level should increase, the performance of the PFM system can decrease drastically.

Our approach in this particular example is certainly plausible. We see, however, that it relies on a two-step estimation procedure. In discrete frequency modulation this procedure was a natural choice because it was also a logical practical implementation. In the first example there was no need for the two-step procedure. However, in order to obtain a parallel set of results for Example 1 we can carry out an analogous two-step analysis and similar results. Experimental studies of both types of systems indicate that the analytic results correctly describe system performance. It would still be desirable to have a more rigorous analysis.

We shall discuss briefly in the context of Example 2 an alternate approach in which we bound the mean-square error directly. From (115) we see that $s(t, A)$ is an analytic function with respect to the parameter A. Thus all derivatives will exist and can be expressed in a simple form:

$$\frac{\partial^n s(t, A)}{\partial A^n} = \begin{cases} \left(\frac{2E}{T}\right)^{\frac{1}{2}} (\beta t)^n (-1)^{(n-1)/2} \cos(\omega_c t + \beta A t), & n \text{ odd}, \\ \left(\frac{2E}{T}\right)^{\frac{1}{2}} (\beta t)^n (-1)^{n/2} \sin(\omega_c t + \beta A t), & n \text{ even}. \end{cases} \tag{132}$$

This suggests that a generalization of the Bhattacharyya bound that we developed in the problem section of Chapter 2 would enable us to get as good an estimate of the error as desired. This extension is conceptually straightforward (Van Trees [24]). For $n = 2$ the answer is still simple. For $n \geq 3$, however, the required matrix inversion is tedious and it is easier to proceed numerically. The detailed calculations have been carried out [29]. In this particular case the series does not converge fast enough to give a good approximation to the actual error in the high noise region.

One final comment is necessary. There are some cases of interest in which the signal is not differentiable with respect to the parameter. A simple example of this type of signal arises when we approximate the transmitted signal in a radar system by an ideal rectangular pulse and want to estimate the time of arrival of the returning pulse. When the noise is weak, formulas for these cases can be developed easily (e.g., Mallinckrodt

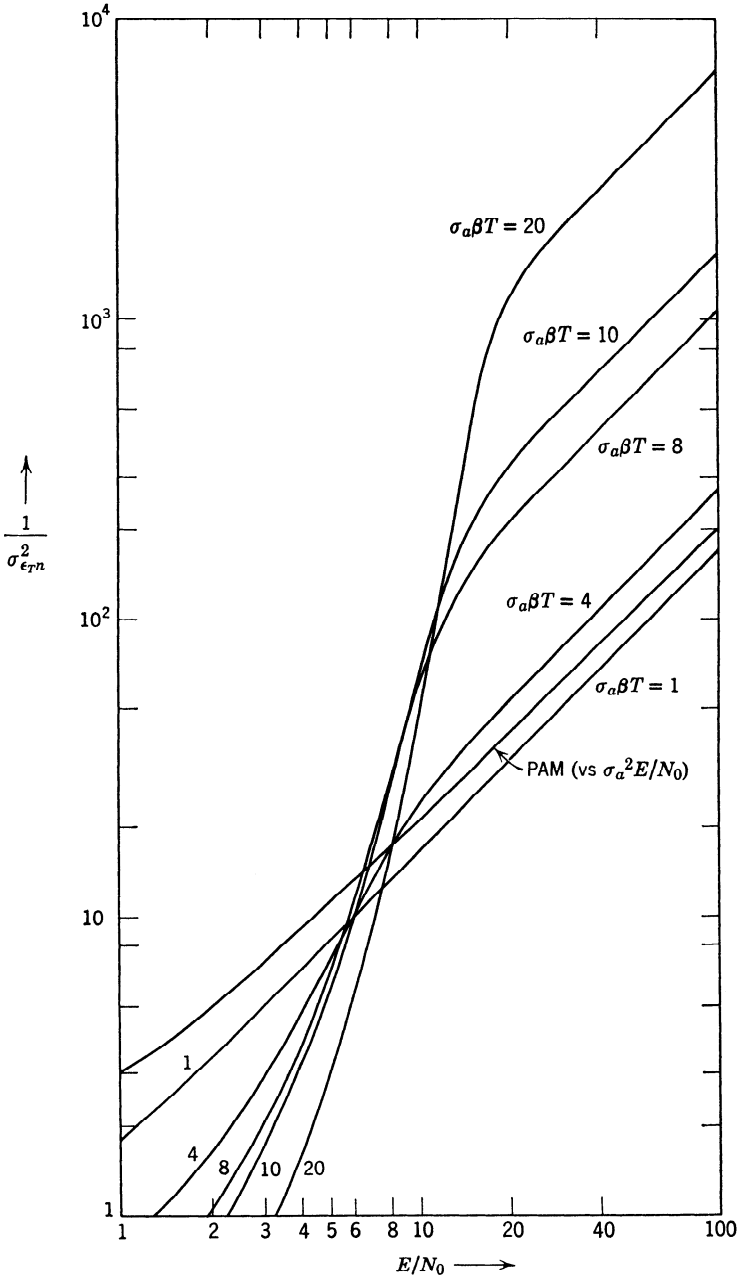

Fig. 4.35 Reciprocal of the total mean-square error.

and Sollenberger [25], Kotelnikov [13], Manasse [26], Skolnik [27]). For arbitrary noise levels we can use the approach in Example 2 or the Barankin bound (e.g., Swerling [28]) which does not require differentiability.

4.2.4 Summary: Known Signals in White Gaussian Noise

It is appropriate to summarize some of the important results that have been derived for the problem of detection and estimation in the presence of additive white Gaussian noise.

Detection. 1. For the simple binary case the optimum receiver can be realized as a matched filter or a correlation receiver, as shown in Figs. 4.11 and 4.12, respectively.

2. For the general binary case the optimum receiver can be realized by using a single matched filter or a pair of filters.

3. In both cases the performance is determined completely by the normalized distance d between the signal points in the decision space,

$$d^2 = \frac{2}{N_0}(E_1 + E_0 - 2\rho\sqrt{E_1E_0}). \tag{133}$$

The resulting errors are

$$P_F = \text{erfc}_*\left(\frac{\ln \eta}{d} + \frac{d}{2}\right), \tag{134}$$

$$P_M = \text{erf}_*\left(\frac{\ln \eta}{d} - \frac{d}{2}\right). \tag{135}$$

For equally-likely hypotheses and a minimum Pr (ϵ) criterion, the total error probability is

$$\Pr(\epsilon) = \text{erfc}_*\left(\frac{d}{2}\right) \leq \left(\frac{2}{\pi d^2}\right)^{1/2} e^{-d^2/8}. \tag{136}$$

4. The performance of the optimum receiver is insensitive to small signal variations.

5. For the M-ary case the optimum receiver requires at most $M - 1$ matched filters, although frequently M matched filters give a simpler implementation. For M orthogonal signals a simple bound on the error probability is

$$\Pr(\epsilon) \leq \frac{M-1}{\sqrt{2\pi(E/N_0)}} \exp\left(-\frac{E}{2N_0}\right). \tag{137}$$

6. A simple example of transmitting a sequence of digits illustrated the idea of a channel capacity. At transmission rates below this capacity the

$\Pr(\epsilon)$ approaches zero as the length of encoded sequence approaches infinity. Because of the bandwidth requirement, the orthogonal signal technique is not efficient.

Estimation. 1. Linear estimation is a trivial modification of the detection problem. The optimum estimator is a simple correlator or matched filter followed by a gain.

2. The nonlinear estimation problem introduced several new ideas. The optimum receiver is sometimes difficult to realize exactly and an approximation is necessary. Above a certain energy-to-noise level we found that we could make the estimation error appreciably smaller than in the linear estimation case which used the same amount of energy. Specifically,

$$\text{Var}\,[\hat{a} - A] \approx \frac{N_0/2}{\int_0^T \left[\frac{\partial s(t, A)}{\partial A}\right]^2 dt}. \tag{138}$$

As the noise level increased however, the receiver exhibited a *threshold* phenomenon and the error variance increased rapidly. Above the threshold we found that we had to consider the problem of a bandwidth constraint when we designed the system.

We now want to extend our model to a more general case. The next step in the direction of generality is to consider known signals in the presence of nonwhite additive Gaussian noise.

4.3 DETECTION AND ESTIMATION IN NONWHITE GAUSSIAN NOISE

Several situations in which nonwhite Gaussian interference can occur are of interest:

1. Between the actual noise source and the data-processing part of the receiver are elements (such as an antenna and RF filters) which shape the noise spectrum.
2. In addition to the desired signal at the receiver, there may be an interfering signal that can be characterized as a Gaussian process. In radar/sonar it is frequently an interfering target.

With this motivation we now formulate and solve the detection and estimation problem. As we have seen in the preceding section, a close coupling exists between detection and estimation. In fact, the development through construction of the likelihood ratio (or function) is identical. We derive the simple binary case in detail and then indicate how the results extend to other cases of interest. The first step is to specify the model.