1

RECEPTION AS A STATISTICAL DECISION PROBLEM

1.1 SIGNAL DETECTION AND ESTIMATION

As we have noted above, our aim in this chapter is to provide a concise review of Bayesian decision methods that are specifically adapted to the basic problems of signal detection (D) and estimation (E). From Fig. 1.1b we can express the reception situation concisely in a variety of equivalent ways through the following operational relations:

A. *Data Processing at the Receiver.*

$$\left(\hat{\mathbf{T}}_{\mathrm{D}} \text{ or } \hat{\mathbf{T}}_{\mathrm{E}}\right)X = Y, \quad \text{or} \quad \left(\hat{\mathbf{T}}_{\mathrm{D}}\hat{\mathbf{R}} \text{ or } \hat{\mathbf{T}}_{\mathrm{E}}\hat{\mathbf{R}}\right)\alpha = Y, \tag{1.1.1}$$

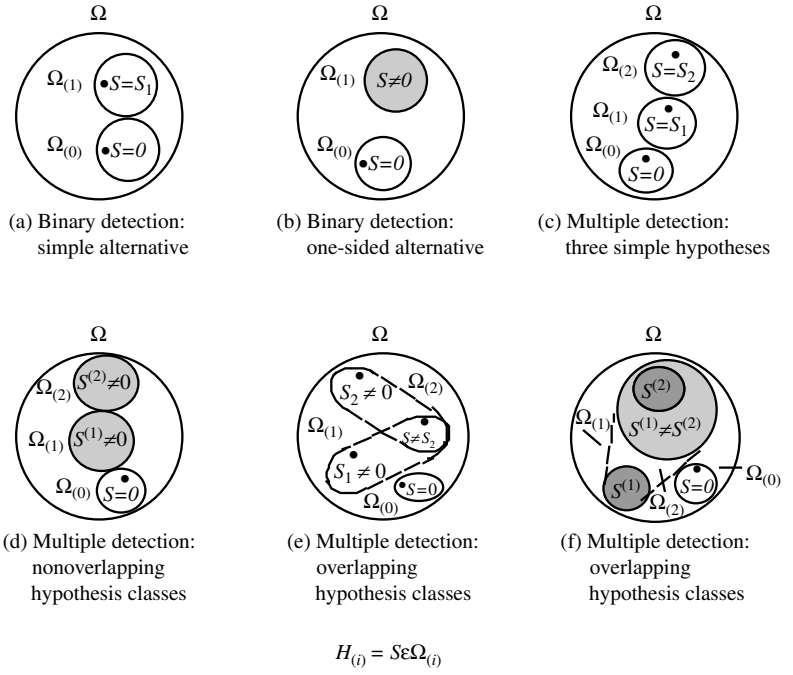
where X is the data input from the spatial processor $\hat{\mathbf{R}}$, that is, the receiving aperture $\hat{\mathbf{R}}\left(\equiv \hat{\mathbf{T}}_{\mathrm{AR}}\right)$, to the temporal data processing elements $\left(\hat{\mathbf{T}}_{\mathrm{D}}, \hat{\mathbf{T}}_{\mathrm{E}}\right)$; Y represents the output from these processors. In more detail from Fig. 1.1b we can also write the following.

B. *Data Input to Processors.*

$$X = \hat{\mathbf{R}}\alpha = \left(\hat{\mathbf{R}}\hat{\mathbf{T}}_{\mathrm{M}}^{(N)}\hat{\boldsymbol{T}}_{\mathrm{AT}}\right)S_{\mathrm{in}}, \tag{1.1.1a}$$

in which α is the propagating field in the medium, which contains ambient sources and scattering elements embodied in the operator $\hat{\mathbf{T}}_{\mathrm{M}}^{(N)}$.

Non-Gaussian Statistical Communication Theory, David Middleton.



(a) Binary detection: simple alternative

(b) Binary detection: one-sided alternative

(c) Multiple detection: three simple hypotheses

(d) Multiple detection: nonoverlapping hypothesis classes

(e) Multiple detection: overlapping hypothesis classes

(f) Multiple detection: overlapping hypothesis classes

FIGURE 1.1 Signal and hypothesis classes in detection.

C. *Field in the Medium.*

$$\alpha = \hat{\mathbf{T}}_{\mathrm{M}}^{(N)} \hat{\mathbf{T}}_{\mathrm{AT}} S_{\mathrm{in}}. \tag{1.1.1b}$$

The input or injected signal S_{in} and output "decisions" $\{v\}$ are described operationally by the following.

D. *Input Signals and Decision Outputs.*

$$S_{\mathrm{in}} = \hat{\mathbf{T}}_{\mathrm{mod}} \hat{\mathbf{T}}_{\mathrm{e}} \{u\}; \quad \{v\} = \hat{\mathbf{T}}_{\mathrm{d}} Y = \hat{\mathbf{T}}_{\mathrm{d}} \big(\hat{\mathbf{T}}_{\mathrm{D}} \text{ or } \hat{\mathbf{T}}_{\mathrm{E}} \big) X, \tag{1.1.1c}$$

where now $\{u\}$ is a set of "messages" to be transmitted and the "decisions" $\{v\}$ fall into two (not necessarily disjoint) classes: ("yes"/"no") for detection (D) and a set of numbers representing measurements, namely estimates of received signal properties or parameters. Comparing Fig. 1.1a and b we see that the "compact" channel operators $\hat{\mathbf{T}}_{\mathrm{T}}^{(N)}$, and so on, in Eq. (1.1.1d) are

E. *Components of the Compact Operators.*

$$\mathbf{T}_{\mathrm{T}}^{(N)} = \hat{\mathbf{T}}_{\mathrm{AT}} \hat{\mathbf{T}}_{\mathrm{mod}} \hat{\mathbf{T}}_{\mathrm{e}}; \quad \hat{\mathbf{T}}_{\mathrm{M}}^{(N)} = \hat{\mathbf{T}}_{\mathrm{M}}^{(N)}; \quad \hat{\mathbf{T}}_{\mathrm{R}}^{(N)} = \hat{\mathbf{T}}_{\mathrm{d}} \hat{\mathbf{T}}_{\mathrm{D/E}} \hat{\mathbf{T}}_{\mathrm{AR}}. \tag{1.1.1d}$$

We emphasize here and subsequently (unless otherwise indicated) that when the signal (if any) is present in X and therefore in the received data Y, Eq. (1.1.1), it is the *received signal* $S_{\mathrm{Rec}} (= S)$. The received signal is, or course, *not* the signal S_{in} originally injected into the medium. This dichotomy occurs because the medium and canonical channel as a

whole modify and generally contaminate S_{in}, with additive and signal-dependent noise (clutter and reverberation) as well as varieties of ambient noise and interference, in addition to such inherent phenomena as absorption and dispersion. All this, of course, is what makes achieving effective reception of the desired signals the challenging problem that we seek to resolve in subsequent chapters. (We remark that S_{Rec} may be generated either by the desired source, S_{in}, or by some undesired source, such as interference, or by a combination of both.)

With this in mind we see that our goals in Chapters 1–7 are first to establish explicit analytic connections between the received data X, the physical realities which affect them (via $\hat{\mathbf{T}}_{\text{M}}^{(N)}$), and the successful extraction of the desired signal, initially as $S(= S_{\text{Rec}})$, and eventually through attained knowledge of the medium (the inverse problem) to obtain acceptable reproduction of the original signal S_{in}. This chapter introduces the formal decision structure for achieving this, while Chapters 2–7 following provide the canonical algorithms (operations on the input data) and performance measures to be used subsequently for specific applications. We remark that the operations involving *coding* $[(\hat{\mathbf{T}}_{\text{d}}, \hat{\mathbf{T}}_{\text{e}})$ in (1.1.1c), and shown in Figure 1.1] belong to the domain of *Information Theory* per se [2], which is outside the scope of the present volume.[1]

1.2 SIGNAL DETECTION AND ESTIMATION

We begin our decision—theoretic formulation with a general description of the two principal reception problems, namely detection and estimation (sometimes called extraction) of signals in noise, expressed operationally above by Eq. (1.1.1). We first introduce some terminology, taken partly from the field of statistics, partly from communication engineering, and review the problems in these terms. We shall also point out some considerations that must be kept in mind concerning the given data of these problems. Later, in Sections 1.3–1.4, we generalize the reception problem, state it in mathematical language, and outline the nature of its solutions.[2]

1.2.1 Detection

The problem of the detection of a (received)[3] signal in noise is equivalent to one which, in statistical terminology, is called the problem of *testing hypotheses*: here, the hypothesis that noise alone is present is to be tested, on the basis of some received data, against the hypothesis (or hypotheses) that a signal (or one of several possible signals) is present.

Detection problems can be classified in a number of ways: by the number of possible signals that need to be distinguished, by the nature of the hypotheses, by the nature of the data and their processing, and by the characteristics of the signal and noise statistics. These will now be described in greater detail.

[1] However, through selected references we shall provide connections to these topics at appropriate points in this book.

[2] As an introduction to the methods of statistical inference, see, for example, the treatments of Kendall [3] and Cramér [4]; also Luce and Raiffa, [5].

[3] Note the comment following Eq. (1.1.1d) above.

1.2.1.1 The Number of Signal Classes to be Distinguished This is equal to the number of hypotheses to be tested but does not depend on their nature. A *binary detection system* can make but two decisions, corresponding to two hypotheses, while a *multiple alternative detection system* [6, 7] makes more than two decisions. For the time being, we deal only with the binary detection problem (the multiple alternative cases are discussed in Chapter 4 ff.).

1.2.1.2 The Nature of the Hypotheses Here the received signal is a desired system input during the interval available for observation of the mixture of signal and noise. Noise (homogeneous—*Hom-Stat* stationary or nonstationary inhomogeneous — *non-Hom-Stat*) is an undesired input, considered to enter the system independently[4] of the signal and to affect each observation according to an appropriate scheme whereby the two are combined.[5] The class of all possible (desired) system inputs is called the *signal class* and is conveniently represented as an abstract space (*signal space*) in which each point corresponds to an individual received signal.

A hypothesis, which asserts the presence of a single signal at the input is termed a simple *hypothesis*. A *class* (or *composite*) *hypothesis*, on the other hand, asserts the presence at the input of an unspecified member of a specified subclass of signals; that is, it reads "some member of subclass k (it does not matter which member) is present at the input." Such a subclass is called a *hypothesis class*. Hypothesis classes may or may not overlap (cf. Fig. 1.1e and f).

Usually, one hypothesis in detection asserts the presence of noise alone (or the absence of any signal) and is termed the *null hypothesis*. In binary detection, the other hypothesis is called the *alternative*. If the alternative is a class hypothesis and the class includes all nonzero signals involved in the problem, it is termed a *one-sided alternative*. It is a *simple alternative* if there is but one nonzero signal in the entire signal space (which signal must therefore contain nonrandom parameters only). Figure 1.1 illustrates some typical situations. In each case, the class of all possible system inputs is represented by signal space Ω.

The hypothesis classes are enclosed by dashed lines and denoted as $\Omega_{(k)}$, where the subscript refers to the hypothesis: that is, the kth hypothesis states that the signal is a member of $\Omega_{(k)}$, or, symbolically, $H_k : S \varepsilon \Omega_{(k)}$. In Fig. 1.1a and b are shown two binary cases corresponding to the simple alternative ($\Omega_{(1)}$ contains one point) and the one-sided alternative ($\Omega_{(1)}$ contains all nonzero system inputs). The latter would occur, for example, if all signals in a binary detection problem were the same except for a random amplitude scale factor, governed by, say, a Gaussian distribution. Figure 1.1c and d shows multiple hypothesis situations where the hypothesis classes do not overlap, while Fig. 1.1e and f represents situations where overlapping can occur, in (e) with single-point classes and in (f) when the classes are one - sided or composite. Many different combinations can be constructed, depending on the actual problems at hand. In the present treatment, we shall confine our attention to the nonoverlapping cases, although the general approach is in no way restricted by our so doing. [But see for example, Section 1.10.2 ff.]

[4] In most applications, but, of course, scattered radiation is signal dependent (cf. Chapters 7 and 8 ff.).

[5] There can be noiselike signals also, but these are not to be confused with the noise background. It is frequently convenient to speak of "noise alone" at the input, and this is to be interpreted as "no signal of any kind" present. In physical systems, which do not, of course, use ideal (i.e., noise-free) elements, noise may be introduced at various points in the system, so that care must be taken in accounting for the manner in which signal and noise are combined.

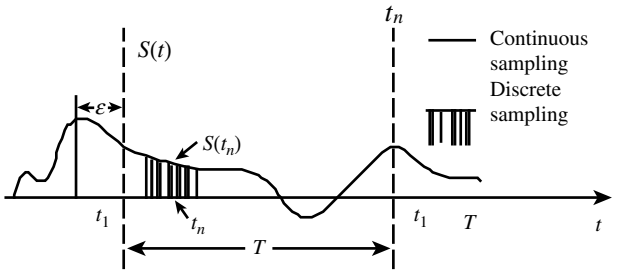


FIGURE 1.2 A temporal signal waveform, showing discrete and continuous time sampling, the epoch ε, and the observation period, T.

1.2.1.3 The Nature of the Data and Their Processing The observations made on the mixture of signal and noise during the observation period may consist of a discrete set of values (*discrete or digital sampling*) or may include a continuum of values throughout the interval (*continuous*, or *analogue sampling*) (cf. Fig. 1.2). Whether one procedure or the other is used is a datum of the problem. In radar, for example, detection may (to a first approximation) be based on a discrete set of successive observations, while, in certain communication cases, a continuous-wave signal may be sampled continuously.

Similarly, it is a datum of the problem whether or not the observation interval, that is, the interval over which the reception system can store the data for analysis, is fixed or variable. In the latter case, one can consider *sequential detection*. A sequential test proceeds in steps, deciding at each stage either to terminate the test or to postpone termination and repeat the test with additional data. In applications of decision theory, it turns out that the analysis divides conveniently at the choice between the sequential and nonsequential. The theory of each type is complete in a certain sense, and additional restrictions on the tests may not be imposed without compromising this completeness. It is, of course, true that since the class of sequential tests includes nonsequential tests as a special subclass, a higher grade of performance may be expected, on the average, under the wider assumption [8, 9].

1.2.1.4 The Signal and Noise Statistics The nature of these quantities is clearly of central importance, as it is upon them that specific calculations of performance depend. In general, individual sample values *cannot* be treated as statistically independent, and this inherent correlation between the sample values over the observation period, in both the continuous and discrete cases, is an essential feature of the problem.

We begin first with temporal waveforms, extending this signal class presently to space–time signals, in Section 1.3.1 ff. Temporal signals may be described in quite general terms involving both random and deterministic parameters. Thus, we write $S(t) = S(t, \varepsilon; a_0, \boldsymbol{\theta})$. Here, ε is an *epoch*, or time interval, measured between some selected point in the "history" of the signal S and, say, the beginning of the observation period $(t_1,\ t_1 + T)$, relating the observer's to the signal's timescale, as indicated in Fig. 1.2; a_0^2 is a scale factor, measuring (relative) intensity of the signal with respect to the noise background; and $\boldsymbol{\theta}$ denotes all other descriptive parameters, such as pulse duration, period, and so on, which may be needed to specify the signal; S itself gives the "shape," or functional *form*, here of the wave in time.

No restriction is placed on the received signal other than that it have finite energy in the observation interval. It may be entirely random, partly random (e.g., a "square wave" with random durations), or entirely causal or deterministic [e.g., a sinusoid, or a more complex

structure that is nevertheless uniquely specified by $S(t)$]. Signals for which the epoch ε assumes a fixed value are said to be *coherent* (with respect to the observer), while if ε is a random variable, such signals are called *incoherent*. Coherent signals may have random parameters and thus belong to subclasses of Ω containing more than one member. Coherent signals corresponding to subclasses containing but a single member will be called *completely coherent*. From these remarks, it is clear that an incoherent signal cannot belong to such an elementary class. The description of the noise is necessarily statistical, and here we distinguish between noise belonging to *stationary* and *nonstationary* processes [10, 11]. Generalizations of the noise structure to include partially deterministic waves offer no conceptual difficulties.

1.2.2 Types of Extraction

We use the term *extraction* here to describe a reception process that calls for an estimate of the received signal itself or one or more of its descriptive parameters.

Signal extraction, like detection, is a problem that in other areas has received considerable attention from statisticians and has been known under the name of *parameter estimation*. A certain terminology has become traditional in the field, which we shall mention presently. We can classify extraction problems under three headings: the nature of the estimate, the nature of the data processing, and the statistics of signal and noise. Much of what can be said under these headings has already been mentioned above. A few more comments may be helpful.

Information about the signal may be available in either of two forms: it may be given as an elementary random process in time, defined by the usual hierarchy of multidimensional distribution functions [12] or it may be a known function of time, containing one or more random parameters with specified distributions. In the latter case, the random parameters may be time independent, or, more generally, they may be themselves random processes (e.g., a noise-modulated sine wave). Clearly there is, as in detection, a wide variety of possible situations. They may be conveniently classified as follows:

1.2.2.1 The Nature of the Estimate A *point estimate*[6] is a decision that the signal or one or more of its parameters have a definite value. An *interval estimate*[6] is a decision that such a value lies within a certain interval with a given probability. Among point estimates, it is useful to make a further distinction between *one-dimensional* and *multidimensional estimates*. An illustration of the former is the estimate of an amplitude scale factor constant throughout the interval, while an estimate of the signal itself throughout the observation period is an example of the latter.

1.2.2.2 The Nature of the Data Processing When the value of a time-varying quantity $X(t)$, (1.1.1b) at a particular instant is being estimated, the relationship between the time t_λ for which the estimate is valid and the times at which data are collected becomes important (cf. Fig. 1.3). If t_λ coincides with one of the sampling instants, the estimation process is termed *simple estimation*, or *simple extraction*. If, on the other hand, t_λ does not coincide with any sampling instant, the process is called *interpolation*, or *smoothing*, when t_λ lies within the observation interval $(t_1,\ t_1 + T)$ and *extrapolation*, or *prediction*, when t_λ lies outside $(t_1,\ t_1 + T)$. Systems of these types may estimate the value of the signal itself or

[6] See Cramér [4] *op. cit.*, for a further discussion of conventional applications.

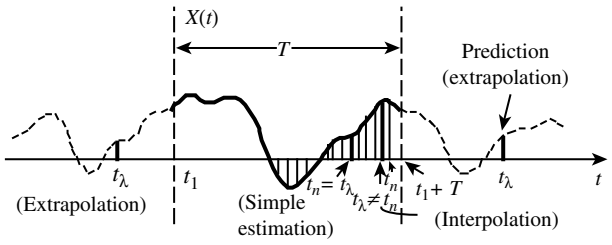


FIGURE 1.3 Simple estimation, interpolation (smoothing), and extrapolation (prediction).

alternatively that of a time-varying signal parameter or some functional of the signal, such as its derivative or integral.

Frequently, a requirement of linearity may be imposed on the optimum system (which is otherwise almost always nonlinear), so that its operations may be performed by a linear network, or sometimes certain specific classes of nonlinearity may be allowed. An important question, then, is the extent to which performance is degraded by such constraints.

1.2.2.3 The Signal and Noise Statistics As in detection, the finite sample upon which the estimate is based may be discrete or continuous, correlated or uncorrelated, and the random processes stationary or nonstationary, ergodic or nonergodic. In a similar way, we may speak of *coherent* and *incoherent extraction* according to whether the received signal's epoch ε is known exactly or is a random variable. The signals themselves may be structurally determinate, that is, the functions S have definite analytic forms; or they may be *structurally indeterminate*, when the S are described only in terms of a probability distribution. A sinusoid is a simple example of the former, while a purely random function is typical of the latter. The case where the signal is known completely does not arise in extraction.

1.2.3 Other Reception Problems

Reception itself may require a combination of detection and extraction operations. Extraction presupposes the presence of a signal at the input, and sometimes this cannot be assumed. We may then perform detection and extraction simultaneously and judge the acceptability of the estimate according to the outcome of the detection process. The problem here is that estimation is performed under uncertainty as to the signal's presence in the received data, which in turn leads to biased estimates that must be suitably accounted for. The analytic results for this new situation are developed and illustrated in detail in Chapters 5 and 6 following. The procedure is schematically illustrated in Figure 1.4, including possible coupling between the detector and extractor.

In our reception problems here, the system designer usually has little control over the received signal, since the medium, embodied in $\hat{\mathbf{T}}_{\mathrm{T}}^{(N)}$, is specified *a priori*. The present definition of the problem states that each possible signal is prescribed, together with its probability of occurrence, and the designer cannot change these data. However, a different strategic situation confronts the designer of a system for transmitting messages from point to point through a noisy channel, since he is then permitted to control the way in which he matches the signal to the channel. The encoding process ($\hat{\mathbf{T}}_{\mathrm{e}}$), Fig. 1.1 is accordingly concerned with finding what class of signal is most effective against channel noise ($\hat{\mathbf{T}}_{\mathrm{M}}^{(N)}$) and

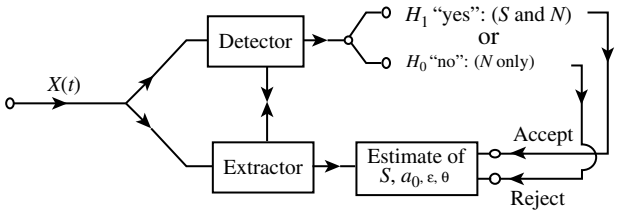


FIGURE 1.4 Reception involving joint signal detection and extraction.

how best to represent messages by such signals (Sections 6.1, 6.5.5 [1]). It is not directly a reception problem, except in the more general situation mentioned earlier (cf. also Section 23.2 [1]), where simultaneous adjustment of the transmission and reception operations $\hat{\mathbf{T}}_{\mathrm{T}}^{(N)}$, $\hat{\mathbf{T}}_{\mathrm{R}}^{(N)}$ is allowed. Decoding $\left(\hat{\mathbf{T}}_{\mathrm{d}}\right)$, of course, is a special form of reception in which the nature of the signals and their distributions are intimately related to the noise characteristics. Moreover, for the finite samples and finite delays available in practice, this is always a nontrivial problem, since it is impossible in physical cases[7] to extract messages (in finite time) from a noisy channel without the possibility of error.

1.3 THE RECEPTION SITUATION IN GENERAL TERMS

Let us now consider the main elements of a general reception problem. We have pointed out earlier that the reception problem can be formulated as a decision problem and that consequently certain information must be available concerning the statistics of signal and noise. We have also indicated that some assumptions are necessary concerning the nature of the data and of the sampling interval and procedures. Finally, we must prescribe a criterion of excellence by which to select an optimum system and must specify the set of alternatives among the decisions to be made.

In our present formulation, we shall make certain assumptions concerning these elements. For definiteness, these assumptions will not be the most general, but they will be sufficiently unrestrictive to exhibit the generality of the approach. Later, in Section 1.4.3, we shall discuss the reasoning by which some of these restrictions are removed.

1.3.1 Assumptions: Space–Time Sampling

Concerning the statistics of signal and noise, we shall assume for the present exposition that both are known *a priori* and as well as the discretely sampled received data $\mathbf{X}$. (In subsequent chapters we shall consider various techniques for handling the problem of unknown or unavailable priors.)

We further extend the sampling process here to space as well as time, since the array operators $\hat{\mathbf{T}}_{\mathrm{AR}}$ and $\hat{\mathbf{T}}_{\mathrm{AT}}$, cf. (1.1.1) et seq. sample the data *field* established in the medium by the signal and noise sources. We further assume that the sampling intervals, or sample size, in time are fixed and of finite duration T and similarly in space, that the *array* or aperture size is likewise finite. Thus, in time $n = 1, \ldots, N$ data elements can be acquired, at each of

[7] Strictly speaking, there is always some noise, although in certain limiting situations this may be a very small effect and hence to an excellent approximation ignorable vis-à-vis the signal.

$m = 1, \ldots, M$ spatial points[8]. Accordingly, we obtain a total of $J = MN$ data components in the received space–time sample.

We employ the following component designations: $j = mn = (\text{space} \times \text{time})$, so that $j = 11, 12, \ldots, 1N$ represents the N time samples at spatial point 1; $j = 21, 22, 23, \ldots, 2N$ similarly denotes the N time samples at spatial point 2, and so on. Thus, *j is a double index numeric*, obeying the convention that the first index (m) refers to the spatial point in question while the second (n) indicates the nth time sample point in T. Specifically, we write $X_{j=mn} = X(\mathbf{r}_m, t_n)$, $S_j = S(\mathbf{r}_m, t_n)$, $N_j = N(\mathbf{r}_m, t_n)$, respectively for the received data $\mathbf{X}$, the received signal $\mathbf{S}$, and noise $\mathbf{N}$, at point $\mathbf{r} = \mathbf{r}_m$ in space and at time $t = t_n$ (see 1.1.1). Furthermore, it is sometimes convenient to introduce a single index numeric, k. Thus, we write for j and k the following equivalent numbering systems:

$$\begin{pmatrix} j \\ k \end{pmatrix} = \begin{pmatrix} 1,1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1,2 \\ 2 \end{pmatrix}, \ldots, \begin{pmatrix} (1,N) \\ N \end{pmatrix}; \begin{pmatrix} 2,1 \\ N+1 \end{pmatrix}, \begin{pmatrix} 2,2 \\ N+2 \end{pmatrix}, \ldots, \begin{pmatrix} 2,N \\ 2N \end{pmatrix}; \ldots; \begin{pmatrix} M,1 \\ (M-1)N+1 \end{pmatrix}, \begin{pmatrix} M,2 \\ (M-1)N+2 \end{pmatrix}, \ldots, \begin{pmatrix} M,N \\ MN \end{pmatrix}. \tag{1.3.1}$$

The double index j is convenient when we need explicitly to distinguish the spatial from the temporal portion of the received field, in processing. It is also useful when we impose the constraint of space and time separability on operations at the receiver, such as array or aperture design, independent of optimization of the temporal processing, a usual although approximate procedure in practice. Of course, j may also be treated as a single index if we order it according to the equivalent scheme (1.3.1), that is, let $j \rightarrow k$. This alternative form is often required when quantitative, that is, numerical, results are desired. The formal structure of the sampling process itself is described in detail at the beginning of Section 1.6.1, cf. Eq. (1.6.2a).

At this point we make no special assumption concerning the criterion of optimality, but we do assume, for the sake of simplicity, that the decision to be made by the system is to select among a finite number L of alternatives. Figure 1.5a and b illustrates the problem. A set of decisions γ are to be made about a received signal $\mathbf{S}$, based on data $\mathbf{X}$, in accordance

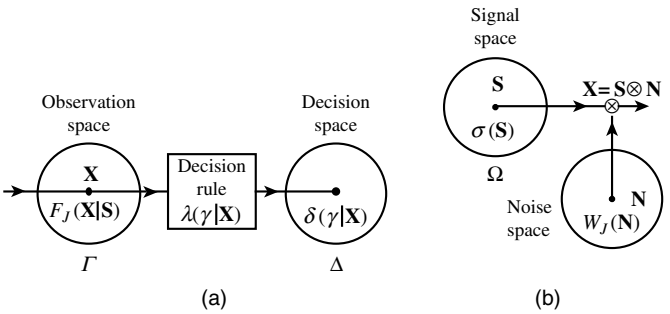


FIGURE 1.5 The reception situation. (a) Observation and decision space; (b) signal space and noise space. $\otimes$ indicates a combination, not necessarily additive, of received signal and noise.

[8] The spatiotemporal structure of both discrete element arrays and continuous apertures are discussed in "Chapter 8–9". See [13, 14].

with a decision rule $\delta(\boldsymbol{\gamma}|\mathbf{X})$, as shown at Fig. 1.5a. Here $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_L)$, $\mathbf{S} = [S_j]$, $\mathbf{X} = [X_j]$, and $\mathbf{N} = [N_j]$ are vectors, and the subscripts on the components of $\mathbf{S}$ and $\mathbf{X}$ are ordered in time so that $S_{mn} = S(\mathbf{r}_m, t_n)$, $X_{mn} = X(\mathbf{r}_m, t_n)$, and so on, with $0 \leq t_1 \leq t_2 \ldots \leq t_n \leq \ldots \leq T_N \leq T$. (Ordering the spatial indexes is arbitrary, essentially a convenience suggested by the structure of the array $(\hat{\mathbf{T}}_{\mathrm{RT}})$ sampling the input field α, cf. (1.1.1b) and Fig. 1.1.) Thus, the components of $\mathbf{X}$ form the *a posteriori* data of the sample upon which some decision γ_l is to be made.

Each of the quantities *received signal* $\mathbf{S}$, *noise* $\mathbf{N}$, *received data* $\mathbf{X}$, *decisions* $\boldsymbol{\gamma}$ can be represented by a point in an abstract space of the appropriate dimensionality. The occurrence of particular values is governed in each instance by an appropriate probability density function. These are multidimensional density functions, which are to be considered discrete or continuous depending on the discrete or continuous nature of the spaces and of corresponding dimensionality.

Here, we introduce $\sigma(\mathbf{S})$, $W_J(\mathbf{N})$, and $F_J(\mathbf{X}|\mathbf{S})$, respectively, as the probability–density functions for the received signal, for noise, and for the data $\mathbf{X}$ when $\mathbf{S}$ is given. (Note that $F_J(\mathbf{X}|\mathbf{0}) = W_J$ ($\mathbf{X}$, by definition.)) As mentioned earlier, the possible (received) signals $\mathbf{S}$ may be represented as points in a space Ω over which the *a priori* distribution $\sigma(\mathbf{S})$ is defined. Information about the signal and its distribution may be available in either of two forms: it may be given directly as an elementary random process, that is, the distribution $\sigma(\mathbf{S})$ is immediately available, as a datum of the problem (stationary, Gaussian, nonstationary, non-Gaussian, etc.). Or, as is more common, the signal $\mathbf{S}$ is a known function of one or more random parameters $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{\hat{M}})$, and it is the distribution[9] $\sigma(\boldsymbol{\theta})$ of these parameters which is given rather than $\sigma(\mathbf{S})$ itself. In fact, the reception problem may require decisions about the parameters instead of the signals.

We can also raise the question of what to do when $\sigma(\mathbf{S})$ is not known beforehand (or perhaps only partially known), contrary to our assumption here. Such situations are in fact encountered in practice, where it is considered risky or otherwise unreasonable to assume that complete knowledge of $\sigma(\mathbf{S})$ is available. This question is a difficult one, and a considerable portion of decision theory is devoted to providing a reasonable answer to it. [It is taken up initially in Sections 1.4.3 and 1.4.4 as well as discussed further in Section 23.4 of Ref. [1], and it is shown that even in this case the above formulation of the reception problem can be retained in its essentials.]

We can further compound the complexity of the reception problem and inquire into what to do when not only $\sigma(\mathbf{S})$, but also $W_J(\mathbf{N})$, the distribution of the noise, is partially or completely unknown *a priori*. In such a case, specification of $\mathbf{S}$ is not enough to determine $F_J(\mathbf{X}|\mathbf{S})$. In statistical terminology $F_J(\mathbf{X}|\mathbf{S})$ is then said to belong to a *nonparametric* family. Error probabilities, associated with possible incorrect decisions, cannot then be computed directly and the system can no longer be evaluated in such terms, so that the question of optimization is reopened. Nonparametric inference from the general point of view of decision theory has been discussed previously by several investigators [15–17][10] but will not be considered further here.

The discussion of possible generalizations, which has so far dealt with the decision space and with the statistics of signal and noise, can also be extended to one other topic, that is, the method of data acquisition. In the line of reasoning that led to the above formulation of the

[9] See Cramér [4] *op. cit.*

[10] See Gibson and Melsa [18] for more recent telecommunication applications.

reception problem, we assumed for convenience that the data were sampled discretely and that the sampling interval was fixed and finite. Actually, neither of these assumptions is strictly necessary. The sampling process can be continuous. Cases of this type are discussed in Chapters 19–23 of Ref. [1].

We observe also that the length of the sampling interval need not be kept fixed. In fact, the idea of a variable sampling interval leads to the notion of *sequential* decisions. A reception system that is based on sequential principles proceeds in steps, deciding, after each sample of data has been processed, whether or not to come to a conclusion or whether to extend the sampling interval and to take another reading. The class of sequential reception systems is broad and contains the nonsequential type discussed so far as a subclass.[11]

1.3.2 The Decision Rule

We begin by observing that the decision rule is represented as a probability. This may seem somewhat surprising. A reception system operating according to such a decision rule would not function like a conventional receiver, which generates a certain and definite output $\boldsymbol{\gamma}$ from a given set of inputs $\mathbf{X}$. Rather, it would contain a battery of chance mechanisms of a well-specified character. A given set of inputs would actuate the corresponding mechanism, which in turn would generate one of the L possible outputs $\boldsymbol{\gamma}$ with a certain probability, each mechanism in general with a different probability. Arrangements such as this will probably appear quite artificial but they are necessary concepts, at least in principle, for it can be shown that devices with chance mechanisms as their outputs can be superior in performance, under certain circumstances, to the conventional ones.

Accordingly, $\delta(\boldsymbol{\gamma}|\mathbf{X})$ is the conditional probability of deciding $\boldsymbol{\gamma}$ when $\mathbf{X}$ is given. More specifically, since the space Δ is here assumed to contain a finite number of decisions $\boldsymbol{\gamma}$, the decision rule $\delta(\boldsymbol{\gamma}|\mathbf{X})$ assigns a probability between (or equal to) 0 and 1 to each decision γ_l $(l = 1, \ \ldots, \ L)$, the distribution depending on $\mathbf{X}$. In most cases of practical interest, δ is either 0 or 1 for each $\mathbf{X}$ and $\boldsymbol{\gamma}$ in this case and is called a *nonrandomized decision rule*. The opposite case, a *randomized decision rule*, is not excluded from this general formulation, although, as we shall see in subsequent applications, the decision rules reduce to the nonrandom case for all the systems treated here.

We note now that the *key feature of the decision situation is that* $\delta(\boldsymbol{\gamma}|\mathbf{X})$ *is a rule for making the decisions* $\boldsymbol{\gamma}$ *from a posteriori data* $\mathbf{X}$ *alone*, that is, without knowledge of, or dependence upon, the particular $\mathbf{S}$ that results in the data $\mathbf{X}$. The *a priori* knowledge of the signal class and signal distribution, of course, is built into the optimum-decision rule, but the probability of deciding $\boldsymbol{\gamma}$, given $\mathbf{X}$, is independent of the particular $\mathbf{S}$; that is, $\boldsymbol{\gamma}$ is algebraically independent of $\mathbf{S}$, although statistically dependent upon it. This may be expressed as

$$\delta(\boldsymbol{\gamma}|\mathbf{X}) = \delta(\boldsymbol{\gamma}|\mathbf{X}, \mathbf{S}), \tag{1.3.1}$$

which states that the probability (density) of deciding $\boldsymbol{\gamma}$, given $\mathbf{X}$, is the same as the probability density of $\boldsymbol{\gamma}$, given both $\mathbf{X}$ and $\mathbf{S}$. Thus, *the decision rule* $\delta(\boldsymbol{\gamma}|\mathbf{X})$ *is the mathematical embodiment of the physical system used to process the data and yield decisions*.

[11] Earlier work on sequential detection is represented by Ref. [8] and [9] here, and more fully by Refs. [2, 32, 33, 36, 38–40] of Chapter 20 of [1]. For a recent, comprehensive treatment, see Chapter 9 of Helstrom [14].

Both fixed and sequential procedures are included in this formulation, and in both cases we deal with terminal decisions. We remark also that the Wald theory of sequential tests [8] introduces a further degree of freedom over the fixed-sample cases through the adoption of a second cost function, the "cost of experimentation" [8, 9]. In the general theory, we are free to limit the class of decision rules, in advance, to either of the above types without compromising the completeness of the theory of either type.

1.3.3 The Decision Problem

In order to give definite structure to the decision process, we must prescribe a criterion of excellence, in addition to *a priori* probabilities $\sigma(\mathbf{S})$ and $W_J(\mathbf{N})$. By this we mean the following: The decisions that are to be made by the reception system must be based on the given data $\mathbf{X}$, which, because of their contamination with noise, constitute only incomplete clues to the received signal $\mathbf{S}$. And, of course, as we have already noted at the beginning of the chapter, the received signal $\mathbf{S}$ itself is already modified by the medium through which it has been propagated, so that $\mathbf{S}_{\text{in}} \neq \mathbf{S}$, cf. Eqs. (1.1.1a–1.1.1c). Therefore, whatever the decision rule $\delta(\boldsymbol{\gamma}|\mathbf{X})$ that is finally adopted, the decisions to which it leads cannot always be correct (except possibly in the unrealizable limit $T \to \infty$). Thus, it is clear that whenever there is a nonzero probability of error some sort of value judgment is implied; in fact, the former always implies (1) a decision process and (2) a numerical cost assignment of some kind to the possible decisions. The units in which such a cost, or value, is measured are essentially irrelevant, but the relative amounts associated with the possible decisions are not.

In order to formulate the decision problem, a *loss* $\mathsf{F}(\mathbf{S}, \boldsymbol{\gamma})$ is assigned to each combination of decisions $\boldsymbol{\gamma}$ and signal $\mathbf{S}$ (the latter selecting a particular distribution function of $\mathbf{X}$, in accordance with some prior judgment of the relative importance of the various correct and incorrect decisions. Each decision rule may then be rated by adopting an *evaluation or risk function* $\mathsf{E}(\mathsf{F})$ (for example, the mathematical expectation of loss), which takes into consideration both the probabilities of correct and incorrect decisions and the losses associated with them. There are, of course, many ways of assigning loss, and hence many different risk functions. One example, which has been very common in statistics and in communication theory, is the squared-error loss. This type of loss is used in extraction problems in which the decision to be rendered is an estimation of a signal after it has been contaminated with noise. In this case, the loss is taken to be proportional to the square of the error in this estimation. Other examples are discussed in chapters 3–7.

We may now state the reception problem in the following general terms:

Given the family of distribution functions $F_J(\mathbf{X}|\mathbf{S})$, the *a priori* signal probability distribution $\sigma(\mathbf{S})$, the class of possible decisions, and the loss and evaluation functions F and $\mathsf{E}(\mathsf{F})$, the problem is to determine the best rule $\delta(\boldsymbol{\gamma}|X)$ for using the data to make decisions.

In arriving at this statement we have introduced a number of somewhat restrictive assumptions. We now give a brief heuristic discussion of what can be done to remove them. To begin with, the statement of the reception problem in these terms is actually more general than the argument that led up to it, a fact that requires some comment. A quick review of that argument shows, on the one hand, that the restriction of the decisions $\boldsymbol{\gamma}$ to a finite number L of alternatives $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_L)$ is irrelevant and that a denumerably infinite number may equally well be used. In fact, the extension to a continuum of possible alternatives is simply a matter of reinterpretation. The decision rule $\delta(\boldsymbol{\gamma}|\mathbf{X})$ that was introduced above as a discrete

probability distribution must in this case be interpreted as a probability–density functional; that is, $\delta(\boldsymbol{\gamma}|\mathbf{X})\,\mathbf{d}\boldsymbol{\gamma}$ is the probability that $\boldsymbol{\gamma}$ lies between $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma} + \mathbf{d}\boldsymbol{\gamma}$, given $\mathbf{X}$. To represent a nonrandomized decision rule in this case, we interpret $\delta(\boldsymbol{\gamma}|\mathbf{X})$ as a Dirac δ-function [see Eq. (1.4.14) ff., for example]. Usually, the family of distribution functions is not given directly and must be found from a given noise distribution $W_J(\mathbf{N})$ and the mode of combining signal and noise.

1.3.4 The Generic Similarity of Detection and Extraction

Figure 1.5 emphasizes that decision rules are essentially transformations that map observation space into decision space. In detection, each point of observation space Γ (or $\mathbf{X}$) is mapped into the various points constituting the space Δ of terminal decisions. For example, the simplest form of binary detection is the same as dividing Γ into two regions, one corresponding to "no signal" and the other to "signal and noise," and carrying out the operation of decision in one step, since only a single alternative is involved. The binary detection problem is then the problem of how best to make this division. The extension to multiple alternative detection situations is made in analogous fashion: one has now three or more alternative divisions of Γ, with a corresponding set of decisions leading to a final decision [7]. Similarly, in extraction each point of Γ is mapped into a point of the space Δ of terminal decisions, which in this instance has the same structure as the signal space Ω. If the dimensionality of Δ is smaller than that of Γ (as is usually the case in estimating signal parameters), the transformation is "irreversible"; that is, many points of Γ go into a single point of Δ. In this way, extraction may also be thought of as a division of Γ into regions, so that, basically, detection and extraction have this common and generic feature and are thus not ultimately different operations. It is merely necessary to group the points of Δ corresponding to $\mathbf{S} \neq 0$ into a single class labeled "signal and noise" to transform an extractor into a detector. Conversely, detection systems may be regarded as extractors followed by a threshold device that separates, say, $\mathbf{S} = 0$ from $\mathbf{S} \neq 0$. However, *a system optimized for the one function may not necessarily be optimized for the other*, and it is in this sense, that we consider detection and extraction as separate problems for analysis.

1.4 SYSTEM EVALUATION

In this section, we shall apply the concepts discussed above to a description of the problem of evaluating system performance, including that of both optimum and suboptimum types. It is necessary first to establish some reasonable method of evaluation, after which a number of criteria of excellence may be postulated, with respect to which optimization may then be specifically defined.

1.4.1 Evaluation Functions

As mentioned in Section 1.3.3, $\mathsf{F}(\mathbf{S}, \boldsymbol{\gamma})$ is a *generalized loss function*, adopted in advance of any optimization procedure, which assigns a *loss*, or *cost*, to every combination of system input and decision (system output) in a way which may or may not depend on the system's operation. Actual evaluation of system performance is now made as mentioned earlier, provided that we adopt an evaluation function $\mathsf{E}(\mathsf{F})$ that takes into account all possible

modes of system behavior and their relative frequencies of occurrence and assigns an overall loss rating to each system or decision rule. One obvious choice of E is the *mathematical expectation E*, or *average value*, of F, and it is on this reasonable but arbitrary choice which the present theory is based for the most part.[12]

At this point, it is convenient to define two different loss ratings for a system, one of which is used to rate performance when the signal input is fixed and the other to take account of *a priori* signal probabilities. For a given $\mathbf{S}$, we have first:

The Conditional Loss Rating.[13] $\mathsf{L}(\mathbf{S}, \delta)$ *of* δ *is defined as the conditional expectation of loss:*

$$\mathsf{L}(\mathbf{S}, \delta) = E_{\mathbf{X}|\mathbf{S}}\{\mathsf{F}[\mathbf{S}, \gamma(\mathbf{X})]\} = \int_{\Gamma} \mathbf{dX} \int_{\Delta} \mathbf{d\gamma} \mathsf{F}(\mathbf{S}, \boldsymbol{\gamma}) F_J(\mathbf{X}|\mathbf{S}) \delta(\boldsymbol{\gamma}|\mathbf{X}). \tag{1.4.1}$$

By this notation we include discrete as well as continuous spaces Δ; for the former, the integral over Δ is to be interpreted as a sum and $\delta(\boldsymbol{\gamma}|\mathbf{X})$ as a probability, rather than as a probability density. (See the remarks at the end of Section 1.3.3.)

Actually, as will be seen in Section 1.4.4, the conditional loss rating is most significant when the *a priori* probability $\sigma(\mathbf{S})$ is unknown. However, when $\sigma(\mathbf{S})$ is known, we use this information to rate the system by averaging the loss over both the sample and the signal distributions:

The Average Loss Rating. $\mathsf{L}(\sigma, \delta)$ of δ *is defined as the (unconditional) expectation of loss when the signal distribution is* $\sigma(\mathbf{S})$ *:*

$$\mathsf{L}(\sigma, \delta) = E_{\mathbf{X}|\mathbf{S}}\{\mathsf{F}(\mathbf{S}, \boldsymbol{\gamma})\} = \int_{\Omega} \mathbf{dS} \int_{\Gamma} \mathbf{dX} \int_{\Delta} \mathbf{d\gamma}\, \mathsf{F}(\mathbf{S}, \boldsymbol{\gamma}) \sigma(\mathbf{S}) F_J(\mathbf{X}|\mathbf{S}) \delta(\boldsymbol{\gamma}|\mathbf{X}). \tag{1.4.2}$$

Some remarks are appropriate concerning the loss function F. In the statistical literature, F is usually a function that assigns to each combination of signal and decision a certain loss, or *cost*, which is independent of δ:

$$\mathsf{F}_1 = C(\mathbf{S}, \boldsymbol{\gamma}). \tag{1.4.3}$$

In the present analysis, we restrict our discussion chiefly to systems whose performance is rated according to simple loss functions[14] of this nature. There exists a substantial body of theory for this case, and certain very general statements can be made about optimum systems derived under this restriction (cf. Wald's complete class theorem, admissibility [25], and so on; see Section 1.5 ff.

[12] Other linear or nonlinear operations for E are possible and should not be overlooked in subsequent generalizations (see the comments in Section 1.5.4).

[13] This quantity is called the *a priori risk* in Wald's terminology [25].

[14] We shall use the term *risk*, henceforth, as synonymous with this simple cost, or loss.

We point out, however, that a more general type of loss function can be constructed. In fact, one such function is suggested by information theory. For, if we let

$$F_2 = -\log p(\mathbf{S}|\boldsymbol{\gamma}), \tag{1.4.4}$$

where $p(\mathbf{S}|\boldsymbol{\gamma})$ is the *a posteriori* probability of $\mathbf{S}$ given $\boldsymbol{\gamma}$, the average loss rating [Eq. (1.4.2)] becomes the well-known equivocation of information theory [2, 26] (Section 6.5.2 of Ref. [1]). This loss function can be interpreted as a measure of the "uncertainty" (or "surprisal") about $\mathbf{S}$ when $\boldsymbol{\gamma}$ is known [26], (Section 6.2.1 of Ref. [1]). It is an example of a more general type than the simple cost function [Eq. (1.4.3)]. For, unlike $C(\mathbf{S}|\boldsymbol{\gamma})$, which depends on $\mathbf{S}$ and $\boldsymbol{\gamma}$ alone, Eq (1.4.4) depends also on the decision rule in use and cannot be preassigned independently of δ. Loss functions like Eq. (1.4.4) are more difficult to deal with, and some of the general statements (Section 1.5) that can be derived for Eq. (1.4.3) clearly do not hold true for Eq. (1.4.4). In Chapter 22 of Ref. [1], however, it is shown that close connections may exist between results based on the two types of loss function.

The conditional and average loss ratings of δ may now be written, from Eqs. (1.4.1)–(1.4.4), as

I. *Conditional Risk:*

$$r(\mathbf{S}, \delta) = \int_{\Gamma} \mathbf{dX} F_J(\mathbf{X}|\mathbf{S}) \int_{\Delta} \mathbf{d}\boldsymbol{\gamma} C(\mathbf{S}|\gamma)\delta(\gamma|\mathbf{X}). \tag{1.4.5}$$

II. *Average Risk:*

$$R(\sigma, \delta) = E\{r(\mathbf{S}, \delta)\} = \int_{\Omega} r(\mathbf{S}, \delta)\sigma(\mathbf{S})\mathbf{dS}, \tag{1.4.6a}$$

or

$$R(\sigma, \delta) = \int_{\Omega} \sigma(\mathbf{S})\mathbf{dS} \int_{\Gamma} \mathbf{dX} F_J(\mathbf{X}|\mathbf{S}) \int_{\Delta} \mathbf{d}\boldsymbol{\gamma} C(\mathbf{S}|\boldsymbol{\gamma})\delta(\boldsymbol{\gamma}|\mathbf{X}). \tag{1.4.6b}$$

III. *Conditional Information Loss:*

$$h(\mathbf{S}, \delta) = -\int_{\Gamma} \mathbf{dX} F_J(\mathbf{X}|\mathbf{S}) \int_{\Delta} \mathbf{d}\boldsymbol{\gamma} [\log p(\mathbf{S}|\boldsymbol{\gamma})]\delta(\boldsymbol{\gamma}|\mathbf{X}). \tag{1.4.7}$$

IV. *Average Information Loss:*

$$H(\sigma, \delta) = E\{h(\mathbf{S}, \delta)\} = \int_{\Omega} h(\mathbf{S}, \delta)\sigma(\mathbf{S})\mathbf{dS}, \tag{1.4.8a}$$

or

$$H(\sigma, \delta) = -\int_{\Omega} \sigma(\mathbf{S})\mathbf{dS} \int_{\Gamma} \mathbf{dX} F_J(\mathbf{X}|\mathbf{S}) \int_{\Delta} \mathbf{d}\boldsymbol{\gamma} [\log p(\mathbf{S}|\boldsymbol{\gamma})]\delta(\boldsymbol{\gamma}|\mathbf{X}). \tag{1.4.8b}$$

The last of these is the well-known "equivocation" of information theory[15] (cf. Sections 6.5.2 and 6.5.3 of Ref. [1].

As we have already mentioned in Section 1.1, $\mathbf{S}$, when deterministic, is a function of a set of random parameters[16] $\boldsymbol{\theta}$, and frequently it is the parameters $\boldsymbol{\theta}$ about which decisions are to be made, rather than about $\mathbf{S}$ itself (see, e.g., Section 1.4.2). Similar to Eqs. (1.4.5) and (1.4.6), the conditional and average risks for this situation may be expressed as[17]

$$r(\boldsymbol{\theta}, \delta) = \int_{\Gamma} \mathbf{dX} F_J(\mathbf{X}|\mathbf{S}(\boldsymbol{\theta})) \int_{\Delta} \mathbf{d}\boldsymbol{\gamma} C(\boldsymbol{\theta}|\gamma) \delta(\gamma|\mathbf{X}), \tag{1.4.9}$$

and

$$R(\sigma, \delta)_{\boldsymbol{\theta}} = \int_{\Omega_{\boldsymbol{\theta}}} r(\boldsymbol{\theta}, \delta) \sigma(\boldsymbol{\theta}) \mathbf{d}\boldsymbol{\theta}. \tag{1.4.10}$$

Here, of course, $r(\boldsymbol{\theta}, \delta)$ and $R(\sigma, \delta)_{\boldsymbol{\theta}}$ are not necessarily the same as $r(\mathbf{s}, \delta)$, $R(\sigma, \delta)$ above, nor is the form of $\boldsymbol{\gamma}$ either. Notice that the cost function $C(\boldsymbol{\theta}|\boldsymbol{\gamma})$ is usually a different function of $\boldsymbol{\theta}$ from $C[\mathbf{S}(\boldsymbol{\theta}), \boldsymbol{\gamma}]$ also. Considerable freedom of choice as to the particular conditional and average risks is thus frequently available to the system analyst, although the appropriate choice is often dictated by the problem in question. Finally, observe that $r(\mathbf{S}, \delta)$ and $R(\sigma, \delta)$ for decisions about $\mathbf{S} = \mathbf{S}(\boldsymbol{\theta})$ are still given by Eqs. (1.4.5), (1.4.6) where $\sigma(\mathbf{S})\mathbf{dS}$ is replaced by its equivalent $\sigma(\boldsymbol{\theta})\mathbf{d}\boldsymbol{\theta}$ in Eqs. (1.4.6a) and (1.4.6b), with a corresponding change from Ω-space (for $\mathbf{S}$) to $\Omega_{\boldsymbol{\theta}}$-space (for $\boldsymbol{\theta}$) according to the transformations implied by $\mathbf{S} = \mathbf{S}(\boldsymbol{\theta})$. Similar remarks apply for the conditional and average information losses, Eqs. (1.4.7) and (1.4.8), as well.

1.4.2 System Comparisons and Error Probabilities

The expressions (1.4.1) and (1.4.2) for the loss ratings can be put into another and often more revealing form, which exhibits directly the rôle of the error probabilities associated with the various possible decisions. Let $p(\boldsymbol{\gamma}|\mathbf{S})$ be the conditional probability[18] that the system in question makes decisions $\boldsymbol{\gamma}$ when the signal is $\mathbf{S}$ and a decision rule $\delta(\boldsymbol{\gamma}|\mathbf{X})$ is adopted, so that

$$p(\boldsymbol{\gamma}|\mathbf{S}) = \int_{\Gamma} F_J(\mathbf{X}|\mathbf{S}) \delta(\boldsymbol{\gamma}|\mathbf{X}) \mathbf{dX}. \tag{1.4.11}$$

Comparison with the conditional risk (1.4.5) shows that the latter may be written

$$r(\mathbf{S}, \delta) = \int_{\Delta} \mathbf{d}\boldsymbol{\gamma} p(\gamma|\mathbf{S}) C(\mathbf{S}, \gamma), \tag{1.4.12}$$

[15] Note that when $\mathbf{S}$ can assume a continuum of values (as is usually the case), we must replace the probability $p(\mathbf{S}|\gamma)$ in Eq. (1.4.4) by the corresponding probability *density* $w(\mathbf{S}|\boldsymbol{\gamma})$ and include in Eqs. (1.4.7) and (1.4.8) the absolute entropy (cf. Section 6.4.1 of Ref. [1]).

[16] For simplicity, these are assumed to be time-invariant here; the generalization to include time variations $\boldsymbol{\theta} = \boldsymbol{\theta}(t)$ is straightforward.

[17] Here and henceforth, unless otherwise indicated, we adopt the notational convention that the principal argument of a function distinguishes that function from other functions: thus, $\sigma(\mathbf{S}) \neq \sigma(\boldsymbol{\theta})$, $p(\mathbf{X}) \neq p(\boldsymbol{\theta})$, and so on; however, $\sigma[\mathbf{S}(\boldsymbol{\theta})] = \sigma(\mathbf{S})$ and so on.

[18] Or probability density, when $\boldsymbol{\gamma}$ represents a continuum of decisions (as in extraction, cf. Chapter 5).

which is simply the sum of the costs associated with all possible decisions for the given **S**, weighted according to their probability of occurrence. In a similar way, we can obtain the probability (density) of the decisions $\boldsymbol{\gamma}$ by averaging Eq. (1.4.11) with respect to **S**, for example,

$$p(\boldsymbol{\gamma}) = \langle p(\gamma|\mathbf{S})\rangle_S = \int_\Omega \sigma(\mathbf{S})\mathbf{dS}\int_\Gamma \mathbf{dX}F_j(\mathbf{X}|\mathbf{S})\delta(\gamma|\mathbf{X}). \tag{1.4.13}$$

Since we shall be concerned in what follows almost exclusively with nonrandomized decision rules, particularly in applications, we see that $\delta(\gamma|\mathbf{X})$ may be expressed as

$$\delta(\boldsymbol{\gamma}|\mathbf{X}) = \delta[\boldsymbol{\gamma} - \boldsymbol{\gamma}_\sigma(\mathbf{X})], \tag{1.4.14}$$

where the δ of the right-hand member is now the Dirac δ-function. Here it is essential to distinguish between the decisions $\boldsymbol{\gamma}$ and the functional operation $\gamma_\sigma(\mathbf{X})$ performed on the data by the system. The subscript σ reminds us that this operation depends in general on signal statistics. With Eq. (1.4.14), the probability (density) of decisions $\boldsymbol{\gamma}$ on condition **S**, which may represent correct or incorrect decisions, can be written

$$p(\boldsymbol{\gamma}|\mathbf{S}) = \int_\Gamma F_J(\mathbf{X}|\mathbf{S})\delta[\boldsymbol{\gamma} - \boldsymbol{\gamma}_\sigma(\mathbf{X})]\mathbf{dX} = \int_{-\infty}^{\infty}\cdots\int e^{i\tilde{\boldsymbol{\xi}}\boldsymbol{\gamma}}\frac{d\xi}{(2\pi)^L}\int_\Gamma F_J(\mathbf{X}|\mathbf{S})e^{-i\tilde{\xi}\gamma\sigma(\mathbf{X})}\mathbf{dX}, \tag{1.4.15}$$

(cf. [1], Section 17.2.1). This reveals the explicit system operation. Equation (1.4.15) in particular provides a direct way of calculating $p(\boldsymbol{\gamma}|\mathbf{S})$ for any system once its system structure $\boldsymbol{\gamma}_\sigma(\mathbf{X})$ is known. As for Eq. (1.4.13), we can also obtain the probability density of $\boldsymbol{\gamma}$ itself by averaging $p(\boldsymbol{\gamma}|\mathbf{S})$ [Eq. (1.4.15)] over **S**.

Comparison of explicit decision systems now follows directly. For example, this may be done by determining which has the smallest average loss rating $L(\sigma,\delta)$, which, in terms of *average* risk Eq. (1.4.6a), involves the comparison of $R(\sigma,\delta_1)$ and $R(\sigma,\delta_2)$ for two systems with system functions $\boldsymbol{\gamma}_\sigma(\mathbf{X})_1$ and $\boldsymbol{\gamma}_\sigma(\mathbf{X})_2$ [Eq. (1.4.14)]. In a similar fashion, one can compare also $H(\sigma,\delta_1)$ and $H(\sigma,\delta_2)$ [Eq. (1.4.8)]. Note that not only optimum but suboptimum systems may be so handled once $\boldsymbol{\gamma}_\sigma(\mathbf{X})$ is specified, so that now one has a possible quantitative method of deciding in practical situations between "good," "bad," "fair," "best," and so on, where the comparisons are consistently made within a common criterion and where the available information can be incorporated in ways appropriate to each system under study. We emphasize that this consistent framework for system comparison is one of the most important practical features of the theory, along with its ability to indicate the explicit structure of optimum and suboptimum systems, embodied in the decision rule $\delta(\boldsymbol{\gamma}|\mathbf{X})$.

1.4.3 Optimization: Bayes Systems

In Section 1.4.1, we have seen how average and conditional loss ratings may be assigned to any system, once the evaluation and cost functions have been selected. We now define what we mean by an optimum decision system. We state a definition first for the case where

complete knowledge of the *a priori* signal probabilities $\sigma(\mathbf{S})$ is assumed and in which, from what has been said above, evaluation from the point of view of the *average* loss $R(\sigma, \delta)$ is most appropriate. Consider, then, that *one system is* "*better*" *than another if its average loss rating is smaller for the same application (and criterion), and that the* "*best,*" *or optimum, system is the one with the smallest average loss rating*. (The preassigned costs, of course, are the same.) We call this optimum system a *Bayes system*:

> *A Bayes system obeys a Bayes decision rule* δ^*, *where* δ^* *is a decision rule whose average loss rating* L *is smallest for a given a priori distribution* σ. (1.4.16)

For the risk and information criteria of Eqs. (1.4.6a) and (1.4.8a), this becomes

$$R^* = \min_{\delta} R(\sigma, \delta) = R\big(\sigma, \delta^*\big) : \textit{Bayes risk}, \tag{1.4.16a}$$

and

$$H^* = \min_{\delta} H(\sigma, \delta) = H\big(\sigma, \delta^*\big) : \textit{Bayes equivocation}. \tag{1.4.16b}$$

The former minimizes the average risk (or cost), while the latter minimizes the equivocation. Bayes decision rules (for the given F) form a *Bayes class*, each member of which corresponds to a different *a priori* distribution[19] $\sigma(\mathbf{S})$.

1.4.4 Optimization: Minimax Systems

When the *a priori* signal probabilities are not known or are only incompletely given, definition of the optimum system is still open. A possible criterion for optimization in such cases is provided by the *Minimax decision rule* δ_M^*, or Bayes rule associated with the conditional risk $r(\mathbf{S}, \delta)$. As indicated by our notation, there is one conditional risk figure attached to each possible signal $\mathbf{S}$. In general, these risks will be different for different signals, and there will be a minimum among them, say $r(\mathbf{S}, \delta)_{\max}$. The Minimax rule is, roughly speaking, the decision rule that reduces this maximum as far as possible. More precisely:

> *The Minimax decision rule* δ_M^* *is the rule for which the maximum conditional loss rating* $\mathsf{L}(\mathbf{S},\delta)_{\max}$, *as the signal* $\mathbf{S}$ *ranges over all possible values*, *is not greater than the maximum conditional loss rating of any other decision rule* δ. (1.4.17)

Thus, in terms of conditional risk r, or conditional information loss h, we may write

$$\begin{aligned} \max_{\mathbf{s}} r(\mathbf{S}, \delta_{\mathrm{M}}^*) &= \max_{\mathbf{s}} \min_{\delta} r(\mathbf{S}, \delta) \le \max_{\mathbf{s}} r(\mathbf{S}, \delta), \\ \max_{\mathbf{s}} h(\mathbf{S}, \delta_{\mathrm{M}}^*) &= \max_{\mathbf{s}} \min_{\delta} h(\mathbf{S}, \delta) \le \max_{\mathbf{s}} h(\mathbf{S}, \delta). \end{aligned} \tag{1.4.17b}$$

[19] Of course, it is possible that different $\sigma(\mathbf{s})$ may lead to identical decision rules, but aside from this possible ambiguity we observe that a Bayes criterion is entirely appropriate when $\sigma(\mathbf{s})$ is known, since it makes full use of all available information.

Wald has shown[20] under certain rather broad conditions (see Sections 1.5.2 and 1.5.3 ff.) that $\max_{\mathbf{S}} \min_{\delta} r(\mathbf{S}, \delta) = \min_{\delta} \max_{\mathbf{S}} r(\mathbf{S}, \delta)$ for the risk (i.e., simple cost) formulations, from which the significance of the term "Minimax" becomes apparent. Whether or not a corresponding result holds for the information-loss formulation remains to be established.

We may also express the Minimax decision process in terms of the resulting average risk. From Section 1.5.3 ff., Theorems 1, 4, 5, 9, we have the equivalent Minimax formulation

$$\left.\begin{aligned} R^*_{\mathrm{M}}(\sigma_0, \delta^*_{\mathrm{M}}) = \max_{\sigma} R^*(\sigma, \delta^*) &= \max_{\sigma} \min_{\delta} R(\sigma, \delta) \\ &= \min_{\delta} \max_{\sigma} R(\sigma, \delta) \end{aligned}\right\} = \text{Minimax average risk}^{21} \tag{1.4.18}$$

this last from Eq. (1.4.16a) and Section 1.5.2 ff., definition 7a. Thus, the Minimax average risk is the largest of all the Bayes risks, considered over the class of *a priori* signal distribution $\{\sigma(\mathbf{S})\}$. The distribution $\sigma_0 (= \sigma^*_{\mathrm{M}})$ for which this occurs is called the *least favorable distribution*. Accordingly, the Minimax decision rule δ^*_{M} [obtained by adjusting the Bayes rule δ^* as σ is varied, cf. Eq. (1.4.18)] is one which gives us the least favorable, or "worst," of all Bayes — that is, "best" — systems. Geometrically, the Minimax situation of $\sigma \to \sigma_0,\ \delta \to \delta^*_{\mathrm{M}},\ R(\sigma, \delta) \to R^*_{\mathrm{M}}(\sigma_0, \delta^*_{\mathrm{M}})$ is represented by a saddle point of the average-risk surface over the (σ, δ) plane, as Fig. 1.6 indicates. The existence of $\sigma_0, \delta^*_{\mathrm{M}}$ and this saddle point follows from the appropriate theorems (cf. Section 1.5.3).

The Minimax decision rule has been the subject of much study and also of some adverse criticism. It has been argued that it is often too conservative to be very useful. However, it is also true that there are situations in which the Minimax rule is unquestionably an excellent choice. Figure 1.7a illustrates these remarks.

Here we have presented the case where the maximum conditional loss rating of all other decision rules $\delta_1, \delta_2, \ldots$ exceeds that for δ^*_{M} and where even most of the minimum loss ratings are also noticeably larger than the corresponding minimum for δ^*_{M}. Sometimes, however, we may have the situation shown in Fig. 1.7b, where δ^*_{M} leads to excessive loss ratings, except for a comparatively narrow range of values of $\mathbf{S}$. In the latter case, δ^*_{M} is perhaps too conservative, and a more acceptable decision rule might be sought.[22] The Minimax procedure does, at any rate, have the advantage of guarding against the worst case, but also may be too cautious for the more probable states of the input to the system. When the costs are preassigned and immutable, the possible conservatism of Minimax

[20] The minimax theorem was first introduced and proved by Von Neumann, in an early paper on the theory of games [27]. For a further account, see Von Neumann and Morgenstern [28], Section 17.6, p. 154; also Ref. [5].

[21] This is the average risk associated with the Minimax decision rule.

[22] These Minimax risk curves have a single distinct maximum. The least favorable *a priori* distribution σ_0 is this case consequently concentrates all of its probability mass at the signal value corresponding to the maximum conditional risk (a δ-function distribution for continuous signal space), since by definition σ_0 must maximize Bayes (average) risk. Existence of a least favorable distribution is here ensured by our assumptions *A* to *D* of Section 1.5.3, which correspond to Wald's assumptions 3.1–3.7 [25]. Roughly speaking, the Minimax conditional risk must equal its maximum value for all signals to which the least favorable distribution assigns a nonzero *a priori* probability (see Wald's theorems 3.10 and 3.11). Thus, a Minimax conditional risk curve with two distinct and equal maxima could have a corresponding σ_0 with probability concentrated at either of the two maxima or distributed between them, while if one maximum were larger than the other, the mass would have to be concentrated only at the larger, and so on. Or again, if σ_0 were nonzero over a finite interval, the corresponding Minimax conditional risk would be constant over this range (but might take on other, smaller values outside).

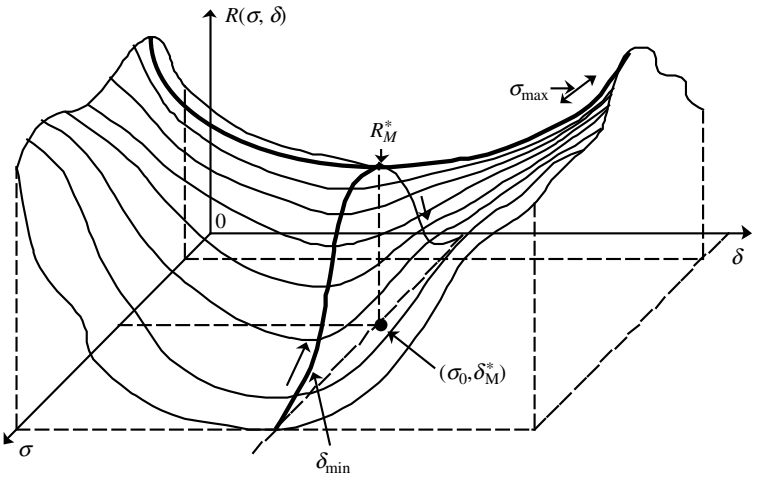


FIGURE 1.6 Average risk as a function of decision rule and *a priori* signal distribution, showing a Minimax saddle point.

cannot be avoided without choosing another criterion.[23] However, in many cases where the actual values of the preassigned costs are left open to an *a posteriori* adjustment, it may be possible by a more judicious cost assignment to modify δ_M^* more along the lines of Fig. 1.7a, where the "tails" of $\max_{\mathbf{S}} r(\mathbf{S}, \delta_M^*)$ are comparable to those of $r(\mathbf{S}, \delta)$ (all $\mathbf{S}$), and thus eliminate, at least in part, the conservative nature of the decision process.[24]

The Bayes decision rule makes the fullest use of *a priori* probabilities (when these are known) and in a sense assumes the most favorable system outcome. The Minimax decision rule, on the other hand, makes no use at all of these *a priori* probabilities (for the good reason that they are not available to the observer) and in the same sense assumes the worst case [cf. Eq. (1.4.18) and Fig. 1.7]. In practical cases, an important problem is to find δ_M^*. No general simple procedure is available, although δ_M^* always exists in the risk formulation. From the definitions of δ^* and δ_M^*, however, it can be shown that a *Bayes decision rule whose*

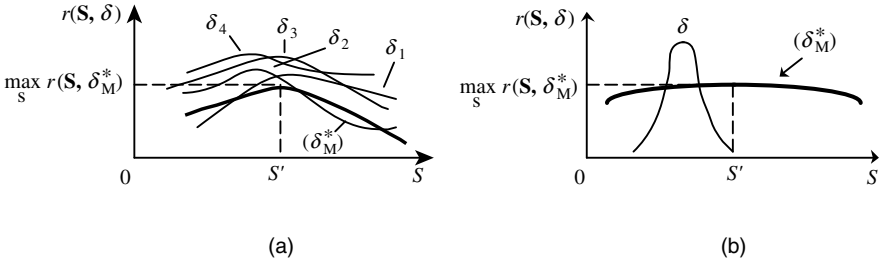


FIGURE 1.7 (a) An acceptable Minimax situation. (b) A Minimax situation that is possibly too conservative.

[23] For example, the "Minimax regret" criterion or the Hurwicz criterion, and so on [29].

[24] Hodges and Lehmann [30] have discussed some intermediate situations where $\sigma(\mathbf{s})$ is partially known on the basis of previous experience. Also, see Section 1.5 ff.

conditional loss rating L *is the same for all signals is a Minimax rule* (see Section 1.5.3 and Theorem 7). Thus, if we can find a δ^* for which this is the case, we have also determined a least favorable *a priori* distribution $\sigma_M^*(S)$ for this $\delta^*(=\delta_M^*)$, which follows from the above, that is,

$$r(\mathbf{S},\delta^*)\,[\text{or } h(\mathbf{S},\delta^*)] = \text{constant, all } \mathbf{S}. \tag{1.4.19}$$

Note that a non-Bayes rule whose conditional loss rating is constant for all $\mathbf{S}$ is not necessarily Minimax. When Eq. (1.4.19) holds, it furnishes a useful method for finding δ_M^* in each case.

1.5 A SUMMARY OF BASIC DEFINITIONS AND PRINCIPAL THEOREMS

We conclude this chapter now with a short summary of some of the principal results, theorems, and so on which give decision theory its general scope and power in our present application to communication problems. For proofs and further discussion, the reader is referred to the appropriate sections of Wald [25] and other pertinent references.

1.5.1 Some General Properties of Optimum Decision Rules

The practical utility of an optimum procedure lies to a considerable extent in its uniqueness: there is not a number of different reception systems with the same optimal properties. For the unique optimum, the problem of choosing the simplest (or least expensive) from the point of view of design is automatically resolved. For this reason and for the central one of optimization itself, it is important to know the properties of optimum decision rules and when they may be expected to apply in physical systems. We state now two of the main results on which subsequent applications are based:

1.5.1.1 Admissible Decision Rules We note, first, that the conditional loss rating of a decision rule depends, of course, on the particular signal present at the input. One decision rule may have a smaller rating than another for some signals and a larger one for others. If the conditional loss rating of δ_1 never exceeds that of δ_2 for *any* value of $\mathbf{S}$, and is actually less than that of δ_2 for at least one $\mathbf{S}$, then δ_1 is said to be *uniformly better* than δ_2. This leads, accordingly, to the notion of *admissibility*:

A decision rule is admissible if no uniformly better one exists. (1.5.1)

Observe that with this definition an admissible rule is not necessarily uniformly better than any other; other rules can have smaller ratings at particular $\mathbf{S}$ (Fig. 1.7a). However, they cannot be better for *all* $\mathbf{S}$.

It follows, then, that, *if a Bayes or Minimax rule is unique, it is admissible*. The converse is not true, since an admissible rule is not necessarily Bayes or Minimax. Accordingly, no system that does not minimize the average risk (loss rating) can be uniformly better than a Bayes system [for the same $\sigma(\mathbf{S})$], and no system that does not minimize the maximum conditional risk (loss rating) can be uniformly better than a Minimax system. Admissibility is an important additional optimum property of unique Bayes and Minimax decision systems.

1.5.1.2 The Complete Class Theorem This is Wald's fundamental theorem ([25], Theorem (3.20)) concerning complete classes of decision rules. We say first that a class D of decision rules is *complete* if, for any δ not in D, we can find a δ^* in D such that δ^* is uniformly better than δ. If D contains no subclass which is complete, D is a *minimal complete class*. Wald has shown that *for the simple loss functions* [Eqs. (1.4.3), (1.4.5), (1.4.6)] *the class of all admissible Bayes decision rules is a minimal complete class*, under a set of conditions that are certainly satisfied for most if not all physical situations (cf. Sections 1.5.2 and 1.5.3 ff.). For the same set of conditions, any Minimax decision rule can be shown to be a Bayes rule with respect to a certain *least favorable a priori* distribution $\sigma_{\mathrm{M}}^*(\mathbf{S})$, and the existence of $\sigma_{\mathrm{M}}^*(\mathbf{S})$, as well as of the Bayes and Minimax rules themselves, is assured. The complete class theorem thus establishes an optimum property of the Bayes class as a whole. To the author's knowledge no complete class theorem has as yet been demonstrated for the information loss ratings of Eqs. (1.4.7) and (1.4.8), nor have the general conditions for the existence of Bayes and Minimax rules for such measures been established. However, some results on the characterization of Bayes tests with this measure, for detection, are given in Ref. [1], Chapter 22.

1.5.2 Definitions[25]

It is assumed that decisions $\boldsymbol{\gamma}$ are to be made about a signal $\mathbf{S}$, based on observations $\mathbf{X}$ whose occurrence is governed by the conditional distribution–density function $F_j(\mathbf{X}|\mathbf{S})$. The decision rule $\delta(\boldsymbol{\gamma}|\mathbf{X})$ is the probability (density) that $\boldsymbol{\gamma}$ will be decided when the observation is $\mathbf{X}$, regardless of $\mathbf{S}$.

Risk theory is based on the following definitions:

(1) It is assumed that a *cost* $C(\mathbf{S}, \boldsymbol{\gamma})$ is preassigned to every possible combination of signal $\mathbf{S}$ and decision γ_l, $l = 1, \ldots, L$, in the problem.

(2) The *conditional risk* $r(\mathbf{S}, \delta)$ of using a decision rule δ is the expected value of the cost when the signal is S:

$$r(\mathbf{S}, \delta) = \int_{\Gamma} \int_{\Delta} C(\mathbf{S}, \boldsymbol{\gamma}) \delta(\boldsymbol{\gamma}|\mathbf{X}) F_J(\mathbf{X}|\mathbf{S}) \mathbf{dX}\, \mathbf{d}\boldsymbol{\gamma}. \tag{1.5.2}$$

(3) The *average risk* $R(\sigma, \delta)$ of using δ is the expected value of $r(\mathbf{S}, \delta)$ in view of the *a priori* probability (density) $\sigma(\mathbf{S})$:

$$R(\sigma, \delta) = \int_{\Omega} r(\mathbf{S}, \delta) \sigma(\mathbf{S}) \mathbf{dS}. \tag{1.5.3}$$

(4) A *Minimax* decision rule δ_{M}^* is one whose maximum conditional risk is not greater than the maximum conditional risk of any other δ:

$$\max_{\mathrm{s}} r\left(\mathbf{S}, \delta_{\mathrm{M}}^*\right) \leq \max_{\mathrm{s}} r(\mathbf{S}, \delta), \text{ for all } \delta. \tag{1.5.4}$$

[25] Wald's book [25] is referred to in the following as SDF. Wald's paper [25a] is recommended as an introduction to the subject.

(5) A Bayes decision rule δ^* is one whose average risk is smallest for a given *a priori* distribution $\sigma(S)$:

$$R(\sigma, \delta^*) = \min_{\delta} R(\sigma, \delta), \text{ for all } \delta. \tag{1.5.5}$$

(6) A decision rule δ_1 is *uniformly better* than a decision rule δ_2 if the conditional risk of δ_1 does not exceed that of δ_2 for any value of $\mathbf{S}$ and is actually less than that of δ_2 for some particular $\mathbf{S}$.

(7) A decision rule is *admissible* if no uniformly better one exists.

   a. An admissible rule is *not* necessarily uniformly better than any other; that is, other rules can have smaller risks at a particular value of $\mathbf{S}$. The point is that they cannot be better at *all* values of $\mathbf{S}$.

   b. An admissible rule need not be Minimax. Clearly, δ^*_{M} could have a larger risk than δ at some values of $\mathbf{S}$ and still have a smaller maximum risk.

(8) A class D of decision rules is complete if for any δ not in D we can find a δ'^* in D such that δ'^* is uniformly better than δ. If D contains no subclass which is complete, it is a *minimal complete* class.

1.5.3 Principal Theorems

We assume that the following conditions are fulfilled:

A. $F_J(\mathbf{X}|\mathbf{S})$ is continuous in $\mathbf{S}$.

B. $C(\mathbf{S}, \boldsymbol{\gamma})$ is bounded in $\mathbf{S}$ and $\boldsymbol{\gamma}$.

C. The class of decision rules considered is restricted to either (1) nonsequential rules or (2) sequential rules.

D. $\mathbf{S}$ and $\boldsymbol{\gamma}$ are restricted to finite closed domains.

These conditions are more restrictive in some cases than those imposed by Wald but are sufficient for our purposes. Specifically, Wald's assumptions [25]: (3.1), (3.2), (3.3) are covered by conditions A and B, (3.5) and (3.6) by condition C, and (3.4) and (3.7) by condition D.

Under these assumptions the following theorems exist:

(1) The decision problem viewed as a zero-sum two-person game is *strictly determined*:

$$\max_{\sigma} \min_{\delta} R(\sigma, \delta) = \min_{\delta} \max_{\sigma} R(\sigma, \delta), \qquad \text{(SDF Theorem 3.4)} \tag{1.5.6}$$

(2) For any *a priori* $\sigma(\mathbf{S})$, there exists a Bayes decision rule δ^* relative to $\sigma(\mathbf{S})$ (SDF Theorem 3.5).

(3) A Minimax decision rule exists (SDF Theorem 3.7).

(4) A *least favorable a priori* distribution $\sigma_0(\mathbf{S})$ exists:

$$\min_{\delta} R(\sigma_0, \delta) = \max_{\sigma} \min_{\delta} R(\sigma, \delta), \qquad \text{(SDF Theorem 3.14)} \tag{1.5.7}$$

(5) Any Minimax decision rule is Bayes relative to a least favorable *a priori* distribution (SDF Theorem 3.9).

(6) The class of all Bayes decision rules is complete relative to the class of all decision rules for which the conditional risk is a bounded function of **S** (SDF Theorem 3.20). (Kiefer [31] shows that the restriction of the set of decision rules to those for which the conditional risk is a bounded function of **S** is unnecessary. He also shows that the class of all admissible decision functions is minimal complete. See in addition Wald's remarks following Theorem 3.20 in SDF.)

The facts below follow from the definitions of Section 1.5.2:

(7) A Bayes decision rule δ_{M}^{*} whose conditional risk is constant is a Minimax decision rule. [This follows from definitions 4, 5, and 7. For suppose δ_{M}^{*} were not Minimax. Then there would exist a δ' with smaller maximum risk and smaller average risk with respect to $\sigma_0(\mathbf{S}) = \sigma_{\mathrm{M}}^{*}(\mathbf{S})$. This contradicts the definition of δ_{M}.]

(8) If a Bayes decision rule is unique, it is admissible. [For suppose δ^* were Bayes with respect to $\sigma(\mathbf{S})$ and not admissible. Then a uniformly better δ' would exist; that is, $r(\mathbf{S}, \delta') \leq r(\mathbf{S}, \overline{\delta^*})$ for all **S**, with equality for some **S**. But this implies that the average risk of δ' with respect to $\sigma(\mathbf{S})$ is less than that of δ^* with respect to $\sigma(\mathbf{S})$. This contradicts the definition of δ^*.]

(9) A Minimax decision rule has a smaller *maximum* average risk than any other. [This follows from the fact that the average risk cannot exceed the maximum conditional risk. Of course, for some particular $\sigma(\mathbf{S})$ another test might have smaller average risk than the Minimax with the same $\sigma(\mathbf{S})$.]

Finally, it is of considerable importance practically to be able to avoid randomized decision rules. We have quoted one theorem due to Hodges and Lehmann [15] on this point. Others may be found in some work of Dvoretzky et al. [32].

1.5.4 Remarks: Prior Probabilities, Cost Assignments, and System Invariants

From our discussion, it is clear that *a priori* probabilities play an essential part in the formulation and application of decision theory. In a general way we may say that Bayesian methods of statistical analysis offer two main approaches to providing prior probabilities. One approach, the "subjective" approach, treats probability as the measure of confidence, or plausibility, which we are willing to assign to a proposition or event. The other approach, the so-called "objective" approach, is based on the classic "frequency of occurrence" or prior history of the event in question. Both are, and have been, open to criticism: the subjective viewpoint, of course, introduces the observer's judgment, albeit quantitatively as a probability assignment. On the other hand, the "objective" alternative is limited by its dependence on a "history," or frequency of occurrence, which may not exist. If the event has no history up to the present but can be conceived as a physical possibility, one possible way out is to create an ensemble of virtual event outcomes, and hence generate a resulting "prior" probability measure of this physically possible event. An important pragmatic justification of the subjective viewpoint is that it couples the observer's probabilistic models and decision making to the real world. And this is

accomplished by providing the "plausible" priors, which the Bayesian formulation requires[26].

Both viewpoints offer the needed coupling of the observer's models to real-world applications, although the subjective approach appears to be the one more favored by scientists and engineers. In either case, statistical decision theory (SDT) can be used with various methods (Minimax among them) to provide the needed distributions. For these reasons, the quantification of *a priori* distributions is usually one of the chief problems to be faced in practical situations. In any case, the rôle of *a priori* information cannot be shrugged off or avoided [33–35].

The problem of cost assignment also must be carefully examined, since it provides another important link to the actual situation and its significance in the larger world of events. In this way, the connection between, say, the design of an optimum or near-optimum system and the operational aspects of the original problem is made with a theory of values, which seeks some over-all *raison d'être* for cost assignments in the particular case. Similar remarks apply for other loss functions. Thus, an ongoing task is to seek out other meaningful criteria (besides F_1 and F_2) and establish (if this be possible) similar optimal properties, such as admissibility and the complete class theorem for them also.

Another problem of general importance is to discover the "invariants" of various classes of detection and extraction systems. Perhaps the most important example here is *threshold* or *weak-signal reception, and in particular, reception in non-Gaussian noise*. This is because predicting the level of acceptable weak-signal reception provides limiting lower bounds on performance, expressed for instance in terms of "minimum detectable signal," or minimum acceptable estimation error. A canonical theory of threshold reception is not only generally possible, under benign constraints, but has been evolving for general noise and signals over the last four decades[27] in the Bayesian statistical decision theory (BSDT) formulation. This approach provides optimal processing algorithms, probability measures of performance, and permits evaluation and comparisons with suboptimum procedures.

Finally, it is evident that types of optimization other than Bayes and Minimax are possible: one can take as a criterion minimum average risk *with constraints*, say, on higher moments of the risk function for example, or other evaluation functions like F_2 [Eq. (1.4.4)]. However, an analytical theory for uniqueness, admissibility, and so on, comparable to Wald's for the simple cost function remains to be developed in such instances.

We turn now to the further development of Sections 1.1–1.5. Having presented the underlying theory above, let us reserve the detailed treatment of estimation to Chapter 6 and proceed to the detection problem and some of its more explicit consequences in Sections 1.6–1.10. Examples on estimation are presented in Chapter 5. Extensions are given

[26] Closely related to, and supported by, the Bayesian idea of probability measures of the plausibility assigned to an event or hypothesis, is the very old principle of *Ockham's razor*, which for our scientific purposes may be stated in contemporary terms as "choose, or favor, the simplest hypothesis over more complicated competing alternatives," which in Bayesian terms means selecting the simpler hypothesis as being more likely (i.e., having a larger probability) of being correct. These ideas are discussed more fully, with physical examples, in Refs. [33–35]; see also the Introduction here.

[27] Earlier attempts at a canonical treatment are given by the author in Ref. [19], Sections 19.4 and 21.2.5 (1960). Along similar lines we also note Section 2.7 of Ref. [36].

in Chapters 3 and 4. New material in Chapters 6 and 7 considers the related problem of joint detection and estimation.

1.6 PRELIMINARIES: BINARY BAYES DETECTION [19, 21, 36–38]

In this chapter so far we have described the main elements of space–time signal processing, namely, detection, estimation, and related applications, from a general Bayesian viewpoint. This includes employing statistical decision theoretic methods and parametric statistical models. Here we shall focus in more detail on general formulations of optimal and suboptimal detection. Specifically, in the context of the generic structures presented in Sections 1.1–1.5 above, we shall outline a general theory of single-alternative detection systems $T_{\mathrm{R}}^{(N)} = (T_{\mathrm{R}}^{(N)})_{\mathrm{det}}$, for the common and important cases where the data acquisition period (or, as we shall somewhat more loosely call it, the observation period) is fixed at the outset.[28] Since the decisions treated here have only two possible outcomes, we call them *binary decisions*, and the corresponding detection process, *binary detection*, in order to distinguish them from the multiple-alternative situations examined later in Chapter 4.

There are two types of binary detection processes, depending on whether the hypothesis classes refer to a decision between one or two possible signals. Thus, from Section 1.2 previously, we write symbolically

$$\text{I.} \quad H_1 : \mathbf{S} \otimes \mathbf{N} \text{ versus } H_0 : \mathbf{N} \tag{1.6.1a}$$

for the situation where we are asked to decide between H_1: received signal of class $\mathbf{S}$ with noise, versus H_0: noise alone. For the second situation we write

$$\text{II.} \quad H_2 : \mathbf{S}_2 \otimes \mathbf{N} \text{ versus } H_1 : \mathbf{S}_1 \otimes \mathbf{N} \tag{1.6.1b}$$

in which the decision is between the choice of a received signal of class $\mathbf{S}_2$ versus one from class $\mathbf{S}_1$, where both signals are accompanied [$\otimes$] by noise, not necessarily additive. In both instances the decision is to be made from the received data $\mathbf{X}$. We can further anatomize the structure represented by I and II above, according to their general application, as summarized in Table 1.1.

In the radar and sonar cases the received signal (S) represents a target. In a telecommunications environment S is the desired, received communication waveform and the ambient noise N_{A} embodies the (usually) similar signals or "interference," while N_{REC} is receiver noise. Here $N_{\mathrm{A+REC}} = N_{\mathrm{A}} + N_{\mathrm{REC}}$. In nonadditive situations such as envelope detection, signal and noise are combined ($S \otimes N$) nonlinearly, including the receiver noise as well as any ambient noise and interference (i.e. unwanted signals). Here and subsequently unless otherwise indicated, the term "signal" shall refer to the received (desired) signal at the output of the receiving aperture or array, $\hat{R}_0$. (see, the Introduction.)

We begin our discussion with a Bayesian formulation of the one-signal or "on–off" cases (I), Sections 1.6–1.8, including performance measures and a structure of system

[28] Variable observation periods are referred to in Section 1.8.5.

TABLE 1.1 Binary Hypothesis Classes

I.	Radar, Sonar	$H_1 : S(S_{\text{in}}) + N(S_{\text{in}}) + N_{\text{A+REC}}$	versus $H_0 : N(S_{\text{in}}) + N_{\text{A+REC}}$	: "Addition" : $S \otimes N = S + N$
"on–off"	Telcom	$H_1 : S(S_{\text{in}}) + N_{\text{A+REC}}$	versus $H_0 : N_{\text{A+REC}}$	: " "
II.	Telcom	$H_2 : S_2(S_{\text{in}}) + N_{\text{A+REC}}$	versus $H_1 : S_1(S_{\text{in}}) + N_{\text{A+REC}}$	: " "
	Telcom + Scatter	$H_2 : S_2(S_{\text{in}}) + N(S_{\text{min}}) + N_{\text{A+REC}}$	versus $H_1 : S_1(S_{\text{in}}) + N(S_{\text{min}}) + N_{\text{A+REC}}$	: " "

comparisons (Section 1.9). This is followed by the formal extension of the theory to the two-signal cases (II), Section 1.10. (A summary is given in Chapter 2 of some illustrative exact results.) We observe, moreover, that exact results are the exception rather than the rule in practical applications, so that approximate methods must be employed if we are to achieve useful analytical and numerical results.

1.6.1 Formulation I: Binary On–Off Signal Detection

Binary detection problems in communication systems have been studied probabilistically in terms of tests of hypotheses since the 1940s. The original formulation in terms of statistical decision theory stems from the 1950s ([1], Chapter 18 [19]).[29] The principal objectives of the following sections here are to obtain (1) a formulation of the binary detection problem itself and (2) by so doing, to indicate how these different viewpoints can not only be reëstablished by the decision theoretical approach but extended to include situations of general practical significance. Specifically, we first derive a general class of Bayes systems and the rather well-known result that several other detection systems considered previously are special cases of this.

Before we begin to develop the elements of statistical communication theory (SCT) outlined in Sections 1.1–1.5, let us establish the effects of sampling of the input fields at the receiver. We have considered principal modes of sampling the continuous input field: (1) a continuous procedure, which essentially reproduces the original field and (2) a discrete sampling procedure that produces a series of sampled values, at the space–time points $(\mathbf{r}_m, t_n)$, where both are obtained during a finite (or infinite) interval Δ (or $\Delta = \infty$). These operations are respectively represented by

$$T_{\mathrm{S}}(\alpha(\mathbf{r},t))_{\mathrm{C}} = X(\mathbf{r},t), \quad \text{and} \quad T_{\mathrm{S}}(\alpha(\mathbf{r},t))_{\mathrm{D}} = X(\mathbf{r}_m,t_n) = X_j, \quad \text{on} \quad \Delta = |\mathbf{R}_0| T \leq \infty, \tag{1.6.2a}$$

where explicitly

$$T_S(\,)_{\mathrm{C}} \equiv \int_{-\Delta/2}^{\Delta/2} \delta(\mathbf{r}'-\mathbf{r})\delta(t'-t)(\,)\mathbf{dr}'dt'; \quad T_{\mathrm{S}}(\,)_{\mathrm{D}} = \int_{-\Delta/2}^{\Delta/2} \delta(\mathbf{r}'-\mathbf{r}_m)\delta(t'-t_n)(\,)\mathbf{dr}'dt' \tag{1.6.2b}$$

and $\mathbf{dr}' = dx'dy'dz' = dr_{x'}dr_{y'}dr_{z'}$ or a lesser dimensionality, depending on the sampling process employed. The effects of these two sampling procedures on the input field, as we shall see later in Sections 2.3.1 and 2.3.2, are quite different when applied to ordered data streams in the discrete and continuous cases, for example, in the formation of apertures, arrays and beam patterns (Section 2.5).

[29] For the earlier studies, based for the most part on a second-moment theory (e.g., signal-to-noise ratios, etc.), see the references at the end of Chapters 19 and Reference Supplements, pp. 1103–1109 (1960); pp. 1111–1120 (1996), of Ref. [1]. Somewhat later studies, also included therein, employing a more complete statistical approach and, leading up to and in some instances coinciding with certain aspects of the present theory, are described more fully here and in Ref. [1], Part 4 and Ref. [19]. For more recent work see the references at the end of this chapter [21, 37].

1.6.2 The Average Risk

First we use F_1 [Eq. (1.4.3)] as our loss function and determine optimum systems of the Bayes class, which, as we have seen [Eq. (1.4.6b)], are defined by minimizing *the average risk*

$$R(\sigma,\delta)=\int_{\Omega}\mathbf{ds}\sigma(\mathbf{S})\int_{\Gamma}F_J(\mathbf{X}|\mathbf{S})\mathbf{dX}\int_{\Delta}\mathbf{d}\boldsymbol{\gamma}C(\mathbf{S},\gamma)\delta(\gamma|\mathbf{X}) \tag{1.6.3}$$

We recall that in binary detection we test the hypothesis H_0 that noise alone is present against the alternative H_1 of a signal and noise, so that there are but two points $\gamma=(\gamma_0,\gamma_1)$, respectively, in decision space Δ. For the moment, allowing the possibility that the decision rule δ may be randomized, we let $\delta(\gamma_0|\mathbf{X})$ and $\delta(\gamma_1|\mathbf{X})$ be the probabilities that γ_1 and γ_0 are decided,[30] given $\mathbf{X}$. Since definite, terminal decisions are postulated here, some decision is always made and therefore

$$\delta(\gamma_0|\mathbf{X})+\delta(\gamma_1|\mathbf{X})=1. \tag{1.6.3a}$$

Denoting by $\mathbf{S}$ the input signal that may occur during the observation interval, we may express the two hypotheses concisely as $H_0: S\varepsilon\Omega_0$ and $H_1: S\varepsilon\Omega_1$, where Ω_0 and Ω_1 are the appropriate nonoverlapping hypothesis classes, as discussed in Section 1.2.1.2. In binary detection, the null class Ω_0 usually contains only one member, corresponding to no signal. The signal class Ω_1 may consist of one or more nonzero signals. It is now convenient to describe the occurrence of signals within the nonoverlapping classes Ω_0, Ω_1 by density functions $w_0(\mathbf{S})$, $w_1(\mathbf{S})$, normalized over the corresponding spaces, for example,

$$\int_{\Omega_0}w_0(\mathbf{S})\mathbf{ds}=1\quad\int_{\Omega_1}w_1(\mathbf{S})\mathbf{ds}=1. \tag{1.6.4}$$

If q and $p\,(=1-q)$ are respectively the *a priori* probabilities that some one signal from Ω_0 and Ω_1 will occur, the *a priori* probability distribution $\sigma(\mathbf{S})$ over the total signal space $\Omega=\Omega_0+\Omega_1$ becomes

$$\sigma(\mathbf{S})=qw_0(\mathbf{S})+pw_1(\mathbf{S})=q\delta(\mathbf{S}-0)+pw_1(\mathbf{S}), \tag{1.6.5}$$

this last when there is but one (zero) signal in class Ω_0. Equation (1.6.5) represents the *one-sided alternative* mentioned in Section 1.2.1.2, while if there is only a single signal in class Ω_1 as well, Eq. (1.6.5) becomes $\sigma(\mathbf{S})=q\delta(\mathbf{S}-0)+p\delta(\mathbf{S}-\mathbf{S}_1)$, $(\mathbf{S}_1\neq 0)$, and we have an example of the *simple alternative* situation. In both cases, $\int\sigma(\mathbf{S})\mathbf{ds}=1$, by definition of p, q, and w.

1.6.3 Cost Assignments

The next step in our application of risk theory is to assign a set of costs to each possible combination of signal input and decision. For this we chose $\mathrm{F}_1=C(\mathbf{S},\boldsymbol{\gamma})$, Eq. (1.4.3), as our cost function. We illustrate the discussion with the assumption of one-sided alternatives,

[30] Since the number of alternatives is finite and discrete, the decision rule is represented by a probability (cf. the remarks following the statement of the general reception problem in Section 1.3.3).

noted above, and uniform costs, although the method is not restricted by such choices. Thus, for the binary on–off cases considered here there are four cost assignments: two for possible correct decisions and two for possible incorrect decisions. It is convenient to represent these by a (2×2) *cost matrix* $\mathbf{C}(\mathbf{S}, \boldsymbol{\gamma})$:

$$\mathbf{C}(\mathbf{S}, \boldsymbol{\gamma}) = \begin{bmatrix} C_{1-\alpha} & C_{\alpha} \\ C_{\beta} & C_{1-\beta} \end{bmatrix} \equiv \begin{bmatrix} C_0^{(0)} & C_1^{(0)} \\ C_0^{(1)} & C_1^{(1)} \end{bmatrix}, \tag{1.6.6}$$

where the rows represent costs associated with the hypothesis states H_0, H_1, and the columns costs assigned to the various decisions γ_0, γ_1. Thus, we write

"failure"
- $C_{\alpha} \equiv C_1^{(0)}$, cost of deciding (incorrectly) that a signal is present, when actually only noise occurs; the decision H_1 is false.
- $C_{\beta} \equiv C_0^{(1)}$, cost of deciding (incorrectly) that a signal is *not* present, when it actually is; the decision H_0 is false.

"success"
- $C_{1-\alpha} \equiv C_0^{(0)}$, cost of deciding (correctly) that there is no signal, only noise, that is, the decision H_0 is true.
- $C_{1-\beta} \equiv C_1^{(1)}$, cost of deciding (correctly) that a signal is present; the decision H_1 is true.

Consistent with the meaning of "correct" and "incorrect", that is, equivalently, "success" and "failure," with respect to the possible decisions, we require that

$$\begin{aligned} &C_{1-\alpha} < C_{\alpha}; C_{1-\beta} < C_{\beta} : \text{ "failure"costs more than "success";} \\ &\therefore \det\mathbf{C} = C_{1-\alpha}C_{1-\beta} - C_{\alpha}C_{\beta} < 0. \end{aligned} \tag{1.6.6a}$$

Here observe that the costs are assigned vis-à-vis the possible *signal classes* (hypothesis states) and not with respect to any one signal in a signal class, which in the case of composite hypotheses, contains more than one member, (Section 1.2.1.2). Similarly, H_0 here refers to noise only, representing a specified class of noise processes where, without loss of generality, we can also postulate that $C_{1-\alpha} = 0,\ C_{1-\beta} = 0$, that is, there is no net gain or "profit" from a correct decision.[31] The best we can expect in this situation, if we are forced to adjust the costs, in that success may cost us nothing: $C_{1-\alpha} = C_{1-\beta} = 0$.

We specify next that $F_J(\mathbf{X}|\mathbf{S})$ is continuous in $\mathbf{S}$, that fixed-sample tests only are considered, and that the assumptions needed for the validity of risk theory (Section 1.5.3) are applicable to the received data $\mathbf{X}$ and signals $\mathbf{S}$ in the following, and that these are random or deterministic quantities. Thus, the average cost or risk may now be found from (1.4.6b) by integrating over the two points (γ_1, γ_2) in the decision space Δ. The result is

$$\begin{aligned} R(\sigma, \delta) = \int_{\Gamma} \{ &\left[qC_{1-\alpha}F_J(\mathbf{X}|\mathbf{0}) + pC_{\beta}\langle F_J(\mathbf{X}|\mathbf{S})\rangle_S\right]\delta(\gamma_0|\mathbf{X}) \\ &+ \left[qC_{\alpha}F_J(\mathbf{X}|\mathbf{0}) + pC_{1-\beta}\langle F_J(\mathbf{X}|\mathbf{S})\rangle_S\right]\delta(\gamma_1|\mathbf{X})\}\mathbf{dX}, \end{aligned} \tag{1.6.7}$$

[31] This is achieved by setting $C_{\alpha} \to C_{\alpha'} = C_{\alpha} - C_{1-\alpha}, C_{\beta} \to C_{\beta'} = C_{\beta} - C_{1-\beta}$.

and

$$p\big\langle F_j(\mathbf{X}|\mathbf{S})\big\rangle_S = \int_{\Omega_1} \sigma(\mathbf{S}) F_J(\mathbf{X}|\mathbf{S})\mathbf{ds} = p\int_{\mathrm{S}} w_1(\mathbf{S}) F_J(\mathbf{X}|\mathbf{S})\mathbf{ds}, \tag{1.6.7a}$$

from (1.6.5). When the signal processes owe their statistical natures solely to a set of random parameters $\boldsymbol{\theta}$, that is, when the signals are deterministic (a usual case is practice), then (1.6.7a) has the equivalent form

$$p\big\langle F_J(\mathbf{X}|\mathbf{S}(\boldsymbol{\theta}))\big\rangle_{\boldsymbol{\theta}} = p\int_{\theta} w_1(\boldsymbol{\theta}) F_J(\mathbf{X}|\mathbf{S}(\boldsymbol{\theta}))\mathbf{d}\boldsymbol{\theta}. \tag{1.6.7b}$$

In detail, we have accordingly from [(1.6.7a) and (1.6.7b)] the defining relations

$$\langle F_J(\mathbf{X}|\mathbf{S})\rangle_{\mathbf{S}} = \int_{\mathbf{S}} w_1(\mathbf{S}) F_J(\mathbf{XS}))\mathbf{ds};\ \langle F_J(\mathbf{X}|\mathbf{S}(\boldsymbol{\theta}))\rangle_{\boldsymbol{\theta}} = \int_{\boldsymbol{\theta}} w_1(\boldsymbol{\theta}) F_J(\mathbf{X}|\mathbf{S}(\boldsymbol{\theta}))\mathbf{d}\boldsymbol{\theta} \tag{1.6.7c}$$

where the dimensionality of w_1 is w_J or w_L.

1.6.4 Error Probabilities

The average cost (1.6.7) can be more compactly expressed in terms of the conditional error probabilities and conditional probabilities of correct decisions. To see this, let us begin by introducing the conditional and total error probabilities:

$$\begin{cases} \alpha \equiv \alpha(\gamma_1|H_0) = \text{conditional probability of incorrectly deciding that a signal is present when only noise occurs. This is known in statistics as a Type I error probability. Here in SCT it is called the } \textit{false alarm probability}, \text{ for example, } \alpha = \alpha_F \equiv p_F \\ \beta \equiv \beta(\gamma_0|H_1) = \text{conditional probability of incorrectly deciding that only noise occurs, when a signal (in class } H_1) \text{ is actually present. Analogously to the above, this is often called a Type II error probability, or in SCT, the false rejection probability of the signal.} \end{cases} \tag{1.6.8}$$

The corresponding total error probabilities are $q\alpha$ and $p\beta$, where α and β are now specified in detail by

$$\alpha = \int_{\Gamma} F_J(\mathbf{X}|\mathbf{0})\delta(\gamma_1|\mathbf{X})\mathbf{dX} \text{ and } \beta = \int_{\Gamma} \langle F_J(\mathbf{X}|\mathbf{S})\rangle_S \delta(\gamma_0|\mathbf{X})\mathbf{dX}. \tag{1.6.8a}$$

Alternatively, the conditional and total probabilities of correct decisions are

$$1-\alpha = \int_{\Gamma} F_J(\mathbf{X}|\mathbf{0})\delta(\gamma_0|\mathbf{X})\mathbf{dX} \equiv 1 - p_{\mathrm{F}};\quad 1-\beta = \int_{\Gamma} \langle F_J(\mathbf{X}|\mathbf{S})\rangle_S \delta(\gamma_1|\mathbf{X})\mathbf{dX} \equiv p_{\mathrm{D}}, \tag{1.6.9}$$

where we have used (1.6.3). The quantity $1-\beta$, (1.6.9), is then the conditional probability of (correct) signal detection p_D or in statistical terminology, the *power of the test,* while $\alpha(= p_{\mathrm{F}})$, (1.6.8a), is called the *significance level*, or *test size*.

Applying (1.6.5) to (1.6.9) for the total probability of a decision $\boldsymbol{\gamma} = \gamma_0 = H_0$: no signal, or $\boldsymbol{\gamma} = \gamma_0 = H_1$: a signal in noise, we find respectively that

$$p(\gamma_0) = q(1-\alpha) + p\beta = \int_\Gamma \big[qF_J(\mathbf{X}|\mathbf{0}) + p\big\langle F_J(\mathbf{X}|\mathbf{S})\big\rangle_S\big]\delta(\gamma_0|\mathbf{X})\mathbf{dX}, \tag{1.6.10a}$$

$$p(\gamma_1) = q\alpha + p(1-\beta) = \int_\Gamma \big[q\big\langle F_J(\mathbf{X}|\mathbf{0})\big\rangle + p\big\langle F_J(\mathbf{X}|\mathbf{S})\big\rangle_S\big]\delta(\gamma_1|\mathbf{X})\mathbf{dX}, \tag{1.6.10b}$$

from which we note that

$$p(\gamma_0) + p(\gamma_1) = 1,\ (p \neq q = 1), \tag{1.6.10c}$$

as expected. Using (1.6.8)–(1.6.10), we readily obtain a more compact form for the average risk (1.6.7), namely,

$$R(\sigma,\delta) = \big\{qC_{1-\alpha} + pC_\beta\big\} - q\{C_\alpha - C_{1-\alpha}\}\alpha + p\big\{C_\beta - C_{1-\beta}\big\}\beta \tag{1.6.11}$$

in terms of the Types 1 and 2 error probabilities, cf. Eqs. (1.6.8 and 1.6.9). In terms of the probabilities of *correct decisions* this becomes

$$R(\sigma,\delta) = \big\{qC_\alpha + pC_\beta\big\} - q\{C_\alpha - C_{1-\alpha}\} - p\big\{C_\beta - C_{1-\beta}\big\}(1-\beta) \tag{1.6.11a}$$

$$= R_0 - q(C_\alpha - C_{1-\alpha})(1-p_{\mathrm{F}}) - p\big(C_\beta - C_{1-\beta}\big)p_{\mathrm{D}} \tag{1.6.11b}$$

cf. (1.6.8 and 1.6.9) above. The quantity

$$R_0 \equiv qC_{1-\alpha} + pC_{1-\beta}(\geq 0), \tag{1.6.12}$$

is called the *irreducible risk,* here a quantity that is prefixed once the costs and *a priori* probabilities ($p_1 = 1 - q$) are established. Thus, the corresponding average risk $R(\sigma,\delta)$ here deals with all signals in H_1, as well as the noise (H_0).

From Eq. (1.4.5), the *conditional risk* becomes similarly

$$r(\mathbf{S}) = (1-\alpha')C_{1-\alpha} + \alpha' C_\alpha; \mathbf{S} = \mathbf{0}; \quad [1-\beta'(\mathbf{S})]C_{1-\beta} + \beta'(\mathbf{S})C_\beta; \mathbf{S} \neq \mathbf{0}, \tag{1.6.13}$$

where now the *simple conditional error probabilities* α' *and* β' are distinguished from the *class conditional error probabilities* α *and* β, Eq. (1.6.8a) above, according to

$$\alpha' \equiv \int_\Gamma F_J(\mathbf{X}|\mathbf{0})\delta(\gamma_1|\mathbf{X})\mathbf{dX}(=\alpha); \quad \beta'(\mathbf{S}) = \int_\Gamma F_J(\mathbf{X}|\mathbf{S})\delta(\gamma_0|\mathbf{X})\mathbf{dX} \neq \beta. \tag{1.6.13a}$$

Finally, note that all of the above go over directly into analogous expressions in the case of random signal parameters.

1.7 OPTIMUM DETECTION: ON–OFF OPTIMUM PROCESSING ALGORITHMS

The criterion of optimization here (and subsequently) is chosen to be the minimization of average risk (1.6.11). Thus, by suitable choice of decision rule δ_0 [or δ_1, since δ_0 and δ_1 are related by Eq. (1.6.3) in these binary cases], the average risk $R(\sigma,\delta)$, Eq. (1.6.7) or (1.6.11),

is minimized by making the error probabilities as small as possible, consistent with Eq. (1.6.3) and the constraints (1.6.6) et seq. on the preassigned costs.[32] We assume here, for the moment, that the cost function $C(\mathbf{S}, \gamma)$ is chosen so that overlapping hypothesis (or signal) classes are not included.[33] Here and in Sections 1.8 and 1.9 we first derive the *optimal processing* or detection *algorithms* in their various generic binary forms. In Section 1.9, we shall then consider the evaluation of performance, as measured by the Bayes risk, or its equivalent probabilities of error and correct signal detection.

Eliminating $\delta(\gamma_1|\mathbf{X})$ with the help of Eq. (1.6.3a), we may express Eq. (1.6.9) as

$$R(\sigma, \delta) = \mathsf{R}_0 + p\big(C_\beta - C_{1-\beta}\big)\int_\Gamma \delta(\gamma_0|\mathbf{X})[\Lambda(\mathbf{X}) - K]F_J(\mathbf{X}|\mathbf{0})\mathbf{dX} \tag{1.7.1}$$

where

$$\Lambda(\mathbf{X}) \equiv \frac{p}{q}\frac{\langle F_J(\mathbf{X}|\mathbf{S})\rangle}{F_J(\mathbf{X}|\mathbf{0})} \tag{1.7.2}$$

is a *generalized likelihood ratio* GLR[34] and K is a *threshold*:

$$K \equiv \frac{C_\alpha - C_{1-\alpha}}{C_\beta - C_{1-\beta}} (> 0), \tag{1.7.3}$$

with $\mathrm{R}_0 = qC_0^{(0)} + pC_1^{(1)}$ the *irreducible risk*(1.6.12). Since δ_0, $F(\mathbf{X}|\mathbf{0})$, $C_\beta - C_{1-\beta}$, and so on, are all positive (or zero), we see directly that R can be minimized by choosing $\delta(\gamma_0|\mathbf{X}) \to \delta^*(\gamma_0|\mathbf{X})$ to be unity when $\Lambda < 0$ and zero when $\Lambda \geq K$. Thus, we decide

$\gamma_0 : H_0$ if $\Lambda(\mathbf{X}) < K$
namely, we set $\delta^*(\gamma_0|\mathbf{X}) = 1$ for any $\mathbf{X}$ that yields this inequality. From Eq. (1.6.3) this means also that

$$\delta^*(\gamma_1|\mathbf{X}) = 0. \tag{1.7.4a}$$

The acceptance region of X for which $\delta_0^* = 1$, $\delta_1^* = 0$ is Γ_0, that is, Γ_0 contains all $\mathbf{X}$ satisfying the inequality $\Lambda(\mathbf{X}) < K$.

$\gamma_1 : H_1$ if $\Lambda_J(\mathbf{X}) \geq K$,
that is, we choose $\delta^*(\gamma_1|\mathbf{X}) = 1$ for all $\mathbf{X}$ satisfying this inequality (and equality) and consequently require that

$$\delta^*(\gamma_0|\mathbf{X}) = 0. \tag{1.7.4b}$$

Here Γ_1 denotes the acceptance region of Γ for which

$$\delta_0^* = 0,\ \delta_1^* = 1.$$

[32] Equivalently, $R(\sigma, \delta)$ is minimized by *maximizing* the probabilities of correct decisions; cf. Eq. (1.6.11a).

[33] For the generalization to include overlapping classes, including stochastic as well as deterministic signals, see Section 1.10 ff.

[34] We note that Λ is more general than the classical likelihood ratio $F_J(\mathbf{X}|\mathbf{S})/F_J(\mathbf{X}|\mathbf{0})$, that is, the ratio of the conditional probability densities of $\mathbf{X}$ with and without $\mathbf{S}$ fixed. The generalized likelihood ratio (1.7.2) reduces to this form when the *a priori* probabilities p and q are equal and the signal space contains but one point, corresponding to the very special case of a completely deterministic signal.

We remark that $\delta^*_{1,0}$ are nonrandomized decision rules directly deduced from the minimization process itself. From Eq. (1.6.8) for these optimum rules we may write the Bayes or minimum average risk specifically as

$$R^*(\sigma,\delta^*) = \mathbf{R}_0 + p(C_\beta - C_{1-\beta})\left[\frac{K}{\mu}\alpha^* + \beta^*\right],\ \mu \equiv p/q. \tag{1.7.5}$$

The procedures described above in (1.7.4a and 1.7.4b) present a form of *generalized likelihood-ratio test* (*GLRT*). This general Bayesian definition must be distinguished from GLRT alternatives when the *a priori* probability distributions of the signal or signal parameters are replaced by their conditional maximum likelihood estimates. [As optimal likelihood-ratios, albeit constrained, these conditional GLRTs are a special subset of the general Bayes class of likelihood-ratio detectors minimizing average risk.]

1.7.1 The Logarithmic GLRT

In actual applications it is usually much more convenient to replace the likelihood-ratio Λ by its logarithm, as we shall see presently.[35] This in no way changes the optimum character of the test, since any monotonic function of Λ may serve as test function. Thus, the optimum decision process [Eqs. (1.7.4a and 1.7.4b)] is simply reëxpressed as

Decide

$$\gamma_0 : H_0 \ \textit{if}\ \log\Lambda(\mathbf{X}) < \log K\ \textit{with} \quad \text{or} \quad \gamma_1 : H_1\ \textit{if}\ \log\Lambda(\mathbf{X}) \geq \log K\ \textit{with}$$

$$\delta^*(\gamma_0|\mathbf{X}) = 1 \qquad \delta^*(\gamma_0|\mathbf{X}) = 0$$

$$\delta^*(\gamma_1|\mathbf{X}) = 0, \qquad \delta^*(\gamma_1|\mathbf{X}) = 1. \tag{1.7.6}$$

The likelihood ratio, and equivalently here its logarithm, embody the actual receiver structure $\mathbf{T}_{\mathrm{R}}^{(N)}$; namely, the operation the detector must perform on the received data $\mathbf{X}$ in order to reach an optimal decision as to the presence or absence of a signal (of class S) in noise. The optimum detection situation is schematically illustrated in Fig. 1.8.

Thus, choosing $\delta_0 \to \delta_0^*$, $\delta_1 \to \delta_1^*$(1.7.4a and 1.7.4b) or (1.7.6), may be stated alternatively: *Make the decision for which the a posteriori risk (or cost) is least*. It is important to observe that this is clearly a direct extension of the original *Theorem of Bayes, or Bayes' Rule*, namely, "chose that hypothesis with the greatest *a posteriori* probability, given the (data) $\mathbf{X}$," to include now the various costs associated with the decision process.

1.7.2 Remarks on the Bayes Optimality of the GLR

The complete class theorem (see Section 1.2.1) for the risk formulation assures us that we have an optimum test and *that all such tests based on the likelihood ratio* [Eq. (1.7.2)] *are*

[35] In fact, any monotonic function of the likelihood ratio (1.7.2) is potentially suitable as an optimal (Bayes) test statistic, since (1.7.2), and $F_{\mathrm{mono}}(\Lambda)$, are *sufficient statistics* because Λ, and $F_{\mathrm{mono}}(\Lambda)$, contain all the relevant information for deciding H_1 versus H_0. [See Section 1.9.1.1. for a more detailed discussion of *sufficiency*.]

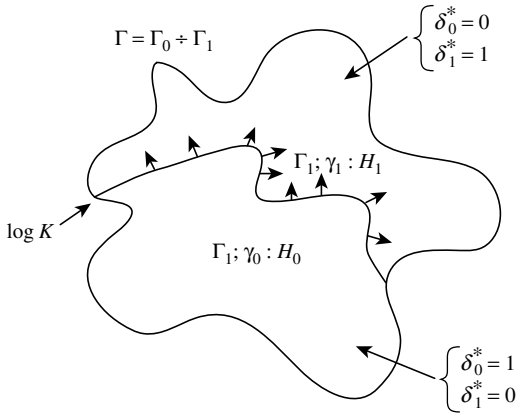


FIGURE 1.8 Optimum binary on–off detection.

Bayes tests (see Section 1.5.1). The Bayes risk R^*(1.7.5) and the average risk R [Eqs. (1.6.11)] for general (not necessarily optimum) systems become, respectively,

$$R^* = \mathsf{R}_0 + p\big(C_\beta - C_{1-\beta}\big)\left(\frac{K}{\mu}\alpha^* + \beta^*\right), \mu \equiv \frac{p}{q}, \tag{1.7.7a}$$

$$R = \mathsf{R}_0 + p\big(C_\beta - C_{1-\beta}\big)\left(\frac{K'}{\mu}\alpha + \beta\right) \tag{1.7.7b}$$

with K given by Eq. (1.7.3), while $R_0 = qC_{1-\alpha} + pC_{1-\beta}$ in either instance [Eq. (1.6.12)]. The threshold K' for the nonideal cases[36] may or may not be equal to K.

Note, incidentally, from Eq. (1.7.7a) that if we differentiate $\big(R^{(*)} - \mathsf{R}_0\big)/p\big(C_\beta - C_{1-\beta}\big)$ with respect to α^* or α we have at once

$$d\beta^*/d\alpha^* = -K/\mu;\ d\beta/d\alpha = -K'/\mu, \tag{1.7.8a}$$

relations that are of use in describing a receiver's performance characteristics. In fact, since $1 - \beta^* = p_{\mathrm{D}}^*$, (1.6.9), (1.6.11a), and since $\alpha^* = p_{\mathrm{F}}^*$, the (conditional) probability of "false alarm," use have

$$\frac{dp_{\mathrm{D}}^*}{dp_{\mathrm{F}}^*} = K/\mu, \quad \text{and} \quad \frac{d\big(pp_{\mathrm{D}}^*\big)}{d\big(qp_{\mathrm{F}}^*\big)} = \frac{dP_{\mathrm{D}}^*}{dP_{\mathrm{F}}^*} = K, \tag{1.7.8b}$$

where P_{D}^* and P_{F}^* are respectively the *un*conditional probabilities of correct detection and false alarm. In terms of the conditional probability, K/μ is the slope of the curve $p_{\mathrm{D}}^* = F^*\big(p_{\mathrm{F}}^*|\ldots\big)$, for example, $dp_{\mathrm{D}}^*/dp_{\mathrm{F}}^* = K/\mu$ when one plots $\beta_{\mathrm{D}}^* = 1 - \beta^*$ versus $p_{\mathrm{F}}^* = \alpha_{\mathrm{F}}^*$. This latter case is generally called the *receiver operating characteristic* (*ROC*) of the Bayes system here. Similarly, with $1 - \beta = p_{\mathrm{D}}, \alpha = p_{\mathrm{F}}$ for suboptimum systems, with slope $dp_{\mathrm{D}}/dp_{\mathrm{F}} = K/\mu, p_{\mathrm{D}}$ versus p_{F} is the corresponding (suboptimum) ROC curve.

[36] One always has a definite (nonzero) threshold when a definite decision is made.

Variants of the relations between $p_{\mathrm{D}}^{(*)}$ and $p_{\mathrm{F}}^{(*)}$ are noted in practice, with analogous relations involving p_{D}^{*}, P_{F}^{*} or P_{D} and P_{F}. [See Section 1.9 and Eq. (1.9.10a) et seq.]

1.8 SPECIAL ON–OFF OPTIMUM BINARY SYSTEMS

A variety of important special cases of the optimum general (fixed-sample) binary detection procedures discussed in Section 2.2 now requires our attention. These are all characterized by one or more constraints on the error probabilities in minimizing the average risk. We begin with the well known *Neyman–Pearson detector*.

1.8.1 Neyman–Pearson Detection[37] Theory

Here the constraint is on the false alarm probability $\alpha_{\mathrm{F}}(= p_{\mathrm{F}})$. We require it to remain fixed, hence the alternative designation of *constant false alarm* (*CFA*) *detector* or constant false alarm *rate* detector (CFAR), for sequences of decisions. Moreover, from the viewpoint of decision theory, we require the total Type I error probability $q\alpha_{\mathrm{F}}$ to remain fixed, while minimizing the total Type II error probability $p\beta$. This is expressed as

$$\mathsf{R}_{\mathrm{NP}}^{*} \equiv \min_{\delta} (p\beta + \lambda q\alpha) = p\beta_{\mathrm{NR}}^{*} + \lambda q\alpha, \tag{1.8.1}$$

where λ is an as yet undetermined multiplier. Minimization is with respect to the decision rule, in the usual way cf. (1.7.4a and 1.7.4b), (1.7.6) and subject to the fact of a definite decision. From Eq. (1.6.8a), we write explicitly

$$\begin{aligned}\mathsf{R}_{\mathrm{NP}}^{*} &= \min_{\delta}\left[p\int_{\Gamma}\langle F_J(\mathbf{X}|\mathbf{S})\rangle_G\,\delta(\gamma_0|\mathbf{X})\mathbf{d}\mathbf{X} + \lambda q\int_{\Gamma} F_J(\mathbf{X}|\mathbf{0})\delta(\gamma_1|\mathbf{X})\mathbf{d}\mathbf{X}\right]\\ &= \min_{\delta}\left\{\int_{\Gamma}\mathbf{d}\mathbf{X}\delta(\gamma_0|\mathbf{X})\left[p\langle F_J(\mathbf{X}|\mathbf{S})\rangle_S - \lambda q F_J(X|0)\right]\right\} + \lambda q.\end{aligned} \tag{1.8.2}$$

For this to be a minimum it is clear that we must choose the δ's such that when

$$\begin{aligned}&p\langle F_n(\mathbf{X}|\mathbf{S})\rangle_S \geq \lambda q F_n(\mathbf{X}|\mathbf{0}),\ \textit{we set}\ \delta^*(\gamma_1|\mathbf{X}) = 1,\ \delta^*(\gamma_0|\mathbf{X}) = 0\\ &\textit{and decide signal and noise},\end{aligned}$$

or when (1.8.3)

$$p\langle F_n(\mathbf{X}|\mathbf{S})\rangle_S < \lambda q F_n(\mathbf{X}|\mathbf{0}),\ \textit{we set}\ \delta^*(\gamma_0|\mathbf{X}) = 1,\ \delta^*(\gamma_1|\mathbf{X}) = 0$$

and decide noise alone,

cf. (1.7.4a and 1.7.4b). Thus, we have the GLRT

$$\mathbf{T}^{(N)}[\mathbf{X}]_{\mathrm{NP}} = \Delta_{\mathrm{NP}} = \frac{p\langle F_n(\mathbf{X}|\mathbf{S})\rangle_S}{qF_n(\mathbf{X}|\mathbf{0})}\begin{cases}\geq \lambda \text{ decide } \gamma_1, \text{ or}\\ < \lambda \text{ decide } \gamma_0,\end{cases} \tag{1.8.4}$$

which establishes the likelihood nature of the detection system [cf. Eq. (1.7.2)].

[37] See Section 19.2.1 and Ref. [26] therein of Ref. [1] for additional remarks regarding the classical Neyman–Pearson hypothesis test.

Comparison with Eq. (1.7.2) et seq. shows that the undetermined multiplier λ here plays the role of threshold K, and $\mathsf{R}^*_{\mathrm{NP}}$ is (except for a scale factor) the corresponding Bayes risk for this threshold $\lambda \equiv K_{\mathrm{NP}}$. (The decision regions Γ_1, Γ_0 for $\mathbf{X}$ are pictured in Fig. 1.8, for $\log \Lambda$. However, λ is not arbitrary but is determined by the constraint of a preassigned value of the conditional Type I error probability α, for example,

$$\alpha_{\mathrm{NP}} = \int_{\Gamma} F_J(\mathbf{X}|\mathbf{0})\delta^*(\gamma_1|\mathbf{X})\mathbf{dX} = \alpha_{\mathrm{NP}}(\lambda = K_{\mathrm{NP}}) \tag{1.8.5}$$

from Eq. (1.6.8a) and the nature of the optimum decision rule [Eq. (1.8.4)].

We can also write Eq. (1.8.4) in the more classical form, either by an obvious modification, or from a computation of $\min_{\delta}(\beta + \lambda\alpha)$ by precisely the same sort of argument given above, namely,

$$\Lambda_{\mathrm{NP}} \equiv \frac{\langle F_n(\mathbf{X}|\mathbf{S})\rangle_S}{F_n(\mathbf{X}|\mathbf{0})} = \frac{\lambda}{\mu} = K'_{\mathrm{NP}}, \tag{1.8.6}$$

with $\mu \equiv p/q$ (1.7.7a), and where now K'_{NP} is a new threshold or significance level, into which have been absorbed the *a priori* probabilities (p, q) and the cost ratio λ. Thus, in the subclass of Neyman–Pearson tests the significance level K'_{NP} is set by choosing α_{NP}, or, equivalently α_{NP} is specified for a predetermined level K'_{NP}. In either instance, it is clear from the preceding remarks that such a formulation *implies* a specific set of *a priori* probabilities (p, $q = 1-p$) and a cost ratio K_{NP} if we are to apply this optimum detection procedure to physical situations. In practice, as has been noted in Section 1.7.1 above, the logarithmic form of the GLRT (1.8.4), with $\lambda \to \log \lambda$ now, is the usually preferred form. CFAR (i.e., Neyman–Pearson) detectors are commonly used in radar and sonar applications, where the practical constraint is keeping the false alarm rate (for sequences of decisions) suitably low, for operational reasons.

1.8.2 The Ideal Observer Detection System

Another way of designing a fixed-sample one-sided alternative test is to require that the *total* probability of error $q\alpha + p\beta$ be minimized, instead of just $p\beta$ as above. An observer who makes a decision in this way is called an *Ideal Observer* [40]. As in the Neyman–Pearson case, this may be set up as a variational problem and shown to yield a likelihood-ratio test with $K = K_I = 1$. Specifically, we want

$$\mathsf{R}^*_I = \min_{\delta}(q\alpha + p\beta) = q\alpha^*_I + p\beta^*_I \tag{1.8.7a}$$

where now α and β are jointly minimized in the sum by proper choice of the decision rule. Using Eqs. (1.7.7a and 1.7.7b) again, we can write Eq. (1.8.7a) as

$$\mathsf{R}^*_I = \min_{\delta}\left\{\int_{\Gamma} \delta(\gamma_0|X)\left[p\langle F_J(\mathbf{X}|\mathbf{S})\rangle_S - qF_J(\mathbf{X}|\mathbf{0})\right]\mathbf{dX}\right\} + q. \tag{1.8.7b}$$

From this it follows at once that the decision procedure for the Ideal Observer is

Decide signal and noise when $\Lambda \geq 1$, that is, set

$$\delta^*(\gamma_1|\mathbf{X}) = 1 \qquad \delta^*(\gamma_0|\mathbf{X}) = 0,$$

or

Decide noise alone when $\Lambda < 1$, that is, set

$$\delta^*(\gamma_0|\mathbf{X}) = 1 \qquad \delta^*(\gamma_1|\mathbf{X}) = 0. \tag{1.8.8}$$

Accordingly, the Ideal Observer system $\mathbf{T}^{(N)}[\mathbf{X}]_{\mathrm{I}}$ is a Bayes detector with threshold K_I of unity. The fact that both the Neyman–Pearson and Ideal Observer systems yield likelihood-ratio tests of the type (1.7.2) follows from the optimum performance they require. Since they *are* likelihood-ratio tests, they belong to the Bayes risk class and accordingly share the general optimum properties possessed by that class, including uniqueness and admissibility (cf. Section 1.5.1). Unlike the Neyman–Pearson detectors (1.8.1), which are particularly appropriate to radar, sonar, and similar hypothesis situations where the decision costs are unsymmetrical, that is, $K \neq 1, > 0$, (1.8.2), the Ideal Observer is the usual choice in many telecommunication applications (e.g., telephone, wireless telephony, etc.) where the costs associated with each class of decision are equal, so that now $K = 1$, cf. (1.8.8). Note, finally, that both these special classes of optimum detection system employ nonrandomized decision rules.

1.8.3 Minimax Detectors

There is yet another possible solution to the detection problem that, like the two just discussed, is optimum in a certain sense and which leads to a likelihood-ratio test. This is the *Minimax detection rule*.

When the *a priori* signal probabilities $\sigma(\mathbf{S})$ are unknown, the Minimax criterion discussed in Section 1.4.4 provides one possible definition of optimum system performance. As we have seen, a Minimax system for binary detection (and the disjoint hypothesis classes of the present chapter) can be regarded as a likelihood-ratio system for some least favorable distribution $\sigma = \sigma_0 = \sigma_{\mathrm{M}}^*$. Once this distribution is found, the Bayes system is completely determined. Now, in order to find it, we may take advantage of the fact that a likelihood-ratio system with the same conditional risks for all signals is Minimax in consequence of the definitions of Bayes and Minimax systems (Section 1.4.4). The procedure is briefly described below.

First, we require the conditional probabilities $\alpha', \beta'|\alpha' = \beta'(S)$ [cf. Eq. (1.6.13a)] to be the result of a Bayes decision rule, which here means a likelihood-ratio test, with σ as yet unspecified. Then, as different σ are tried, different α', β' result (for the same threshold K, which depends only on the preassigned costs). The conditional risks [Eq. (1.6.13)] for $\mathbf{S} = 0$ and for $\mathbf{S} \neq 0$ (all $\mathbf{S}$) will vary. As one increases, the other must decrease, in consequence of the admissibility and uniqueness of the particular Bayes rule corresponding to our choice of σ. If, then, there exists a σ for which the conditional risks [Eq. (1.6.13)] are equal[38] for all $\mathbf{S}$, we have the required Minimax rule δ_{M}^* and associated least favorable prior distribution σ_{M}^*. We remember, moreover, that, if the equation between conditional risks has no solution, this does *not* mean that a Minimax rule or least favorable distribution does not exist, only that other methods must be discovered for determining it.

[38] There can be no system for which both conditional risks together can be less than this, since this is a Bayes test.

As an example, suppose we have the *simple* alternative detection problem, when p and q are unknown, and $w_1(\mathbf{S}) = \delta(\mathbf{S} - \mathbf{S}_I)$ here. From Eq. (1.7.2), we have for the Bayes test

$$\Lambda = \frac{pF_J(\mathbf{X}|\mathbf{S}_I)}{qF_J(\mathbf{X}|\mathbf{0})} \underset{<}{\geq} K \tag{1.8.9}$$

where $\alpha'(=\alpha)$, $\beta'(=\beta)$ are the corresponding error probabilities [cf. Eq. (1.6.8a)]. The α, β are functions of p and q since the decision rule $\delta = \delta_{\mathrm{M}}^*$ depends on p and q through Λ. Thus, as p and q are varied, α, β also are changed, as the boundary between the critical and acceptance region varies. Equating the conditional risks [Eq. (1.6.7)] in order to determine the least favorable $p = p_{\mathrm{M}}^*$, $q = q_{\mathrm{M}}^* (= 1 - p_{\mathrm{M}}^*)$, we have

$$\left[1 - \alpha\left(p_{\mathrm{M}}^*, q_{\mathrm{M}}^*\right)\right] C_{1-\alpha} + \alpha\left(p_{\mathrm{M}}^*, q_{\mathrm{M}}^*\right) C_\alpha = \left[1 - \beta\left(p_{\mathrm{M}}^*, q_{\mathrm{M}}^*\right)\right] C_{1-\beta} + \beta\left(p_{\mathrm{M}}^*, q_{\mathrm{M}}^*\right) C_\beta, \tag{1.8.10}$$

provided that a solution exists.

Alternatively, the least favorable a priori distribution σ_{M}^*, and therefore the Minimax decision rule, may be found in principle from the basic definitions (see Section 1.5.3, Theorems 1–5). That is, $\sigma_0 = \sigma_{\mathrm{M}}^*$ is the *a priori* distribution that maximizes the Bayes risk (cf. Fig. 1.6). Since every Bayes decision rule is associated with a specific *a priori* distribution, however, the Bayes rule changes as this distribution is varied for maximum risk. As a result, this method of finding the extremum may be technically difficult to implement. It is applicable, however, when the previous method (based on uniform conditional risk) fails.

In the case of the one-sided alternative where $w_1(\mathbf{S})$ is known but again p and q are not, the same procedure may be tried when now Eq. (1.7.2) is used in place of Eq. (1.8.9). Finally, if $w_1(\mathbf{S})$ is unspecified, or if neither p, q, nor $w_1(\mathbf{S})$ is given [i.e., if $\sigma(\mathbf{S})$ is completely unavailable to the observer], Eq. (1.8.10) with Eq. (1.7.2) still applies when a solution exists, although the task of finding σ_{M}^* may be excessively formidable. In any case, an explicit evaluation of $(\alpha')_{\mathrm{M}}^*$ and $(\beta')_{\mathrm{M}}^*$ from Eq. (1.6.13a) when $\delta = \delta_{\mathrm{M}}^*$ therein, may be carried out by methods outlined in Section 1.8.1 and illustrated in succeeding sections.

We observe that the Minimax error probabilities $(\alpha)_{\mathrm{M}}^*$, $(\beta)_{\mathrm{M}}^*$ are fixed quantities, independent of the actual *a priori* probabilities p, q, $w_1(\mathbf{S})$ chosen by nature. The average Minimax risk R_{M}^* is given formally by writing R_{M}^* for R^* in Eq. (1.7.7a) and replacing p, q, and so on, and α^*, β^* by p_{M}^*, $q_{\mathrm{M}}^*, \ldots$ and α_{M}^*, β_{M}^* therein. The difference $\left(R_{\mathrm{M}}^* - R^* \geq 0\right)$ between the Bayes (σ known) and Minimax average risk (σ unknown) is thus one useful measure of the price we must pay for our ignorance of nature's strategy (i.e., here nature's choice of p, q, etc.). For further discussion, see Section 20.4.8 of Ref. [1].

1.8.4 Maximum Aposteriori (MAP) Detectors from a Bayesian Viewpoint

Another approach to treating unknown, or unavailable *a priori* pdfs that are exceedingly difficult to evaluate in the likelihood-ratio Λ (1.7.2) (usually of random signal parameters $\boldsymbol{\theta}$ or waveform $\mathbf{S}$), is to employ a suitably optimized estimate of $\boldsymbol{\theta}$, or $\mathbf{S}$. "Suitably optimized" means here that an appropriate likelihood ratio results and hence belongs in the family of Bayes tests, that is, one which yields a minimum average risk, R_{MAP}^* consistent with the available prior information and the constraints imposed by the receiver's ignorance and/or

simplifications. For the latter reason, of course, $R^*_{\text{MAP}} - R^* \geq 0 : R^*_{\text{MAP}}$ is larger than (or at best equal to) the Bayes or minimum average risk with all prior information used, for reasons similar to the Minimax cases discussed above.

To see how such Bayes tests are obtained, we begin by considering the situation where the *a priori* pdf $w_L(\boldsymbol{\theta})$, of the L signal parameters $\boldsymbol{\theta}(= \boldsymbol{\theta}_1, \ \ldots, \ \boldsymbol{\theta}_m)$, is available but is difficult to treat in Λ, (1.7.2). Using Bayes's theorem we may write for the integrand of the numerator of Λ,

$$w_L(\boldsymbol{\theta})F_J(\mathbf{X}|\mathbf{S}(\boldsymbol{\theta})) = w_L(\boldsymbol{\theta}|\mathbf{X})W_J(\mathbf{X})(\equiv W_{J\times L}(\mathbf{X}, \boldsymbol{\theta})), \tag{1.8.11}$$

$$\text{with } W_J(\mathbf{X}) = \int_{\boldsymbol{\theta}} w_L(\boldsymbol{\theta})W_J(\mathbf{X}|\boldsymbol{\theta})\mathbf{d}\boldsymbol{\theta}. \tag{1.8.11a}$$

Consequently, we have the conditional pdf of $\boldsymbol{\theta}$ given $\mathbf{X}$, namely the *a posteriori* pdf of $\boldsymbol{\theta}$ represented by

$$w_L(\boldsymbol{\theta}|\mathbf{X}) = w_L(\boldsymbol{\theta})F_J(\mathbf{X}|\mathbf{S}(\boldsymbol{\theta}))/W_J(\mathbf{X}). \tag{1.8.11b}$$

Next, we use (1.8.11b) in (1.6.9) for the (conditional) Bayes average probability of correct signal detection, namely,

$$1 - \beta^* = \int_{\Gamma} \mathbf{dX}\, \delta(\gamma_1|\mathbf{X}) \int_{\boldsymbol{\theta}} w_L(\boldsymbol{\theta}|\mathbf{X})W_J(\mathbf{X})\mathbf{d}\boldsymbol{\theta}, \tag{1.8.12}$$

which we now maximize by choosing that estimate $\hat{\boldsymbol{\theta}}^*$ which in turn maximizes the integrand (in $\boldsymbol{\theta}$). Thus, we seek that estimate $\hat{\boldsymbol{\theta}}^*$ which maximizes the average value of correct signal detection, for example, $1 - \hat{\beta}^*$. The estimate $\hat{\boldsymbol{\theta}}$ is found from the $\boldsymbol{\theta} \in \Omega_0$ for which the *a posteriori probability* $w_L(\boldsymbol{\theta}|\mathbf{X})$ *is maximum*, namely, from

$$w_L\left(\hat{\boldsymbol{\theta}}^*|\mathbf{X}\right) \geq w_L(\boldsymbol{\theta}|\mathbf{X}), \tag{1.8.13a}$$

$$\text{or equivalently from (1.8.11b)} : \ w_L\left(\hat{\boldsymbol{\theta}}^*\right)F_J\left(\mathbf{X}\middle|\mathbf{S}\left(\hat{\boldsymbol{\theta}}^*\right)\right) \geq w_L(\boldsymbol{\theta})F_J(\mathbf{X}|\mathbf{S}(\boldsymbol{\theta})), \text{ all } \boldsymbol{\theta} \in \Omega_0, \tag{1.8.13b}$$

Here $W_J(\mathbf{X})$ is dropped as irrelevant to the estimation process because it does not contain $\boldsymbol{\theta}$. Accordingly, it is customary to call $\hat{\boldsymbol{\theta}}^* = \hat{\boldsymbol{\theta}}(\mathbf{X})^*$ here a *MAP*, or *maximum a posteriori probability* estimate, which depends, of course, on the received data $\mathbf{X}$. [Note that the *a posteriori* probability $w_L(\boldsymbol{\theta}|\mathbf{X})$ *depends explicitly on the prior probability* $w_L(\boldsymbol{\theta})$, as a consequence of (1.8.11b) in (1.8.12a and 1.8.12b).]

Our next step in obtaining the desired *MAP detector* Λ_{MAP} is to replace the pdf $w_L(\boldsymbol{\theta})$ in the GLR (1.7.2) by the new pdf[39]

$$\hat{w}_L(\boldsymbol{\theta})^* \equiv \delta\left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}^*(\mathbf{X})\right), \tag{1.8.14}$$

[39] For a Bayes formulation of estimation and associated Bayes risk, see chapter 5.

which corresponds to the maximizing operation (1.8.13a and 1.8.13b) for the integrand of the GLR (1.7.2). The result is directly the classical MAP test

$$\textit{Classical MAP Test}: \quad \Lambda_{\mathrm{MAP}}|_{\mathrm{classical}} = \mu F_J\left(\mathbf{X}\middle|\mathbf{S}\left(\hat{\boldsymbol{\theta}}^*\right)\right)/F_J(\mathbf{X}|\mathbf{0})\left\{\begin{array}{l}\geq K : \text{decide } S\otimes N\\ < \quad : \text{decide } N\end{array}\right\}. \tag{1.8.15}$$

Several important points need to be emphasized regarding the maximizing condition (1.8.13a and 1.8.13b) for $\hat{\boldsymbol{\theta}}^*$. First, as we shall see in Chapter 5 ff., (1.8.13a) defines an *unconditional maximum likelihood estimate* (*UMLE*), because of the presence of the *a priori* pdf $w_L(\boldsymbol{\theta})$. The conditional pdf $w_L(\boldsymbol{\theta}|X)$ must be determined from (1.8.11b), which in turn depends explicitly on $w_L(\boldsymbol{\theta})$. *Accordingly, MAP detectors cannot avoid the requirements of an explicit knowledge of* $w_L(\boldsymbol{\theta})$, which limits their use when $w_L(\boldsymbol{\theta})$, or $\sigma(\boldsymbol{\theta})$(1.6.5), is not available. A second and more serious difficulty with the classical result (1.8.15) above, and one which appears to be universally unacknowledged, is with the estimation process itself, as embodied in (1.8.13a and 1.8.13b). The point here is that $p(=p(H_1))$ is less than unity: there is a detection procedure indicated. *The data* $\mathbf{X}$ *do not always contain the desired signal* $\mathbf{S}(\boldsymbol{\theta})$: 100q% of the time the data sample $\mathbf{X}$ contains no signal, only noise. As such the result of employing (1.8.13a and 1.8.13b) which assumes $p = 1$, *yields a biased-estimate* $\boldsymbol{\theta}^* = \hat{\boldsymbol{\theta}}^*_{p=1}$, with an average positive bias *in the magnitude of the estimates,* cf. Chapter 6. However, as the analysis there shows, this situation can be remedied by using the *unbiased estimate* $\hat{\boldsymbol{\theta}}^*_{p<1} = p\hat{\boldsymbol{\theta}}^*_{p=1}$, appropriate to the UMLE process when $p < 1$ (and for $p = 1$). (It is shown in Chapters 5 and 6 that in the context of the Bayes risk formulation for detection and estimation, the UMLEs are derived by minimization of so called "simple cost" functions [cf. Sections 21.2.2 and 21.2.3 [1] and Chapter 5. ff.][40]). Accordingly, the classical result (1.8.15) needs to be replaced by the correct result:

$$\text{MAP Test } (p<1): \quad \Lambda_{\mathrm{MAP}}|_{p<1} = \mu F_J\left(\mathbf{X}\middle|\mathbf{S}\left(p\hat{\boldsymbol{\theta}}^*(\mathbf{X})_{p=1}\right)\right)/F_J(\mathbf{X}|\mathbf{0})\left\{\begin{array}{l}\geq K : \text{decide } S\otimes N\\ < \quad : \text{decide } N\end{array}\right\}. \tag{1.8.16}$$

As we shall see in Chapter 5 even $\hat{\boldsymbol{\theta}}^*(X)_{p=1}$, is itself not always easily obtained.

Finally, there is a variant of the $\mathrm{MAP}|_{p<1}$ detector that can be used when the *a priori* pdf $w_L(\boldsymbol{\theta})$ *is essentially uniform or is at least slowly varying over the range of values of* $\boldsymbol{\theta}$ *in* $F_J(X|S(\boldsymbol{\theta}))$ *where* F_J *is significant*. Thus our maximization of the (conditional) average probability of correct signal detection is accomplished by maximizing $\boldsymbol{\theta}$ according to the condition

$$F\left(\mathbf{X}\middle|\mathbf{S}\left(\hat{\boldsymbol{\theta}}\right)\right) \geq F(\mathbf{X}|\mathbf{S}(\boldsymbol{\theta})\,,) \quad \text{all } \boldsymbol{\theta}\in\Omega_{\boldsymbol{\theta}}. \tag{1.8.17}$$

Although $\hat{\boldsymbol{\theta}}$ is now *apparently* a conditional estimate (no *a priori* pdf $w_L(\boldsymbol{\theta})$), *from the unconditional Bayes viewpoint* (1.8.17) *implies that* $w_L(\boldsymbol{\theta})$ *is uniform, consistent with our requirement (1.8.17) represents the only significant variation of the integrand with* $\boldsymbol{\theta}$.

[40] For these maximum likelihood cases, including (1.8.18) below, the effective part of the maximizing estimation procedure is simply p. The relation (1.8.13a and 1.8.13b) is used when $p \leq 1$. See the analysis for Eqs. (6.3.24)–(6.3.30).

Thus, $w_L(\boldsymbol{\theta}) \doteq w_{oL}$, a constant, and now the integrand of Λ, (1.7.2), from (1.8.17), can be written

$$\int_{\boldsymbol{\theta}} w_{oL}(\boldsymbol{\theta}) F\left(\mathbf{X}\middle|\mathbf{S}\left(\hat{\boldsymbol{\theta}}(\mathbf{X})\right)\right)\mathbf{d}\boldsymbol{\theta} = F\left(\mathbf{X}\middle|\mathbf{S}\left(\hat{\boldsymbol{\theta}}(\mathbf{X})\right)\right), \tag{1.8.18}$$

since $\int_{\boldsymbol{\theta}} w_L(\boldsymbol{\theta})(= w_{0L})\mathbf{d}\boldsymbol{\theta} = 1$. Of course, the conditions on $w_L(\boldsymbol{\theta})$ leading to (1.8.17) must be obeyed for the results to be acceptably accurate. The resulting MAP Test, $p < 1$, here must also take into account the proper estimator for $p < 1$. Again, this is $p\hat{\boldsymbol{\theta}}_{p=1}(\mathbf{X})$, where $\hat{\boldsymbol{\theta}}_{p=1} = \hat{\boldsymbol{\theta}}^*_{p=1}|_{\text{uniform}}$, which is the equivalent UMLE now with a uniform *a priori* pdf $\boldsymbol{\theta} \in \boldsymbol{\Omega}_{\boldsymbol{\theta}}$. The MAP test here becomes from (1.8.15) and the above

$$\text{MAP Test}|_{\text{uniform}} : \Lambda_{\text{MAP}}|_{p<1,\text{ uniform}} = \mu F_J\left(\mathbf{X}\middle|\mathbf{S}\left(p\hat{\boldsymbol{\theta}}^*_{p=1,\text{ uniform}}\right)\right)/F_J(\mathbf{X}|\mathbf{0})\left\{\begin{matrix}\geq\\<\end{matrix}\right\} K \begin{matrix}: \text{ decide } S\otimes N\\: \text{ decide } N.\end{matrix} \tag{1.8.19}$$

Once more, this procedure maximizes the average probability of correct signal detection, now without detailed knowledge of a parametric *a priori* pdf $w_L(\boldsymbol{\theta})$, except again that it be effectively uniform over values of $\boldsymbol{\theta}$ where F is significant *and* with recognition of the fact that $p < 1$.

The associated Bayes risk for these MAP detectors has the form of (1.7.7a), namely,

$$R^*_{\text{MAP}} = R_0 + p(C_\beta - C_{1-\beta})\left(\frac{K}{\mu}\alpha^*_{\text{MAP}} + \beta^*_{\text{MAP}}\right). \tag{1.8.20}$$

We remark once more that $R^*_{\text{MAP}} \geq R^*$, the minimum average risk for the fully known prior pdf's, including $p\,(= 1 - q)$. For quantitative results we must of course evaluate the conditional error probabilities $\alpha^*_{\text{MAP}}, \beta^*_{\text{MAP}}$, and for comparison, α^*, β^*, as well. General expansions for α^*, β^*, α^*_{MAP}, and so on are derived in Section 1.9 ff., from which, in turn, explicit results can be obtained either exactly or approximately by a variety of analytical methods. In any case, the relation $R^*_{\text{MAP}} \geq R^*$ is basically attributable to the fact that R^*_{MAP} employs only partial information regarding the prior pdfs of (here) the parameters $\boldsymbol{\theta}$, in the form of estimates, whereas the Bayes risk R^* uses the true and entire pdfs for $\boldsymbol{\theta}$.

Similar arguments for MAP detection of received signal *waveforms* $\mathbf{S}$ give at once the desired counterparts to (1.8.16) and (1.8.19):

$$\begin{aligned}
&\text{MAP Test } (p<1)\text{: } \Lambda_{\text{MAP}}|_{p<1} = \mu F_J\left(X\middle|p\hat{\mathbf{S}}^*(\mathbf{X})_{p=1}\right)/F_J(\mathbf{X}|\mathbf{0})\\
&\text{MAP Test}|_{\text{uniform}} : \Lambda_{\text{MAP}}|_{p<1;\text{ uniform}} = \frac{\mu F_J\left(X\middle|p\hat{\mathbf{S}}^*(\mathbf{X})_{p=1|\text{uniform}}\right)}{F_J(\mathbf{X}|\mathbf{0})}\left\{\begin{matrix}\geq K:\text{ decide } S\otimes N\\<\ :\text{ decide } N\end{matrix}\right\},
\end{aligned} \tag{1.8.21}$$

where (1.8.13a and 1.8.13b) and (1.8.18), with $\boldsymbol{\theta}$ replaced by $\mathbf{S}$, now provide the maximizing condition for $\hat{\mathbf{S}}^*(\mathbf{X})_{p=1}$ in (1.8.21).

Finally, an alternative way of handling the MAP estimators in detection is discussed in Chapter 6. Here the biased nature of the estimator is handled by a strongly coupled

estimator — a detector system that does not require *explicit a priori* knowledge of $p\ (=p(H_1))$ when $p<1$. This is accomplished by the application of appropriate thresholds on the estimators and detector, along with feedback of the detector's decision (H_1 or H_0) regarding acceptance or rejection of these MAP estimators as well as presence or absence of the signal. Nevertheless, it should be remembered that in the Bayesian formulation prior probabilities (among them p or $q=1-p$), are always at least implied. An extensive discussion of the pros and cons of these methods is given in Section 23.4 of Ref. [1].

1.8.5 Bayesian Sequential Detectors

Other variations in the form of the likelihood detector are also possible. In all of the above, sample size (J) is fixed and minimization of the average risk, generally, involves minimizing the appropriate error probabilities. In sequential detection, however, the false alarm and Type II error probabilities are preset and the aim is to reach a decision ($S\otimes N$ or N) in the shortest time, that is, for the smallest sample size, *on the average*. Thus, sample size J is now the random variable (as well as the data $\mathbf{X}$). Minimization of the average risk is now minimization of the "average cost of experimentation," defined as being proportional to sample size. This Bayes risk can be expressed as

$$R^*_{\text{seq}} = q\alpha C_\alpha + p\beta C_\beta + pC_o \min_{\delta\to\delta^*} \left\langle J\left(\mathbf{X}|\mathbf{S},\delta\right)^* \right\rangle_{\mathbf{X}},\ j\to J^*(\text{termination}) \tag{1.8.22}$$

where C_o is the cost per unit trial (per unit of $j=11, 12, \ldots, J$). In many (but not all) cases, $\delta\to\delta^*_{\text{seq}}$ yields a likelihood detector for the optimum structure. If $y_s\left(\equiv \log\Lambda_{J-\text{seq}}\right)$ is this likelihood detector, then the best procedure involves a *double* threshold, instead of the single threshold ($\log K$) characteristic of the fixed-sample tests described above. The detection process is described by

$$\begin{aligned}
\text{Sequential (Binary) } \textit{Test}:\ &\text{If } B\,(=\beta/(1-\alpha)) < y_s < A\,(=(1-\beta)/\alpha):\ \text{continue test } j\to j+1\\
&\text{If } y_s \geq A:\ \text{test terminates } j\to J(\mathbf{X}|\mathbf{S},\delta_{\text{seq}})^*\ \text{decide } H_1:\ S\otimes N\\
&\text{If } y_s < B:\ \text{test terminates } j\to J(X|S,\delta_{\text{seq}})^*\ \text{decide } H_0:\ N.
\end{aligned} \tag{1.8.23}$$

The theory of sequential tests is due primarily to Wald [8], with its application to signal detection subsequently initiated by Bussgang and Middleton [9], with further development by Blasbalg [22] and others; see also Basseville and Nikiforov [23]. Chapter 9 of Helstrom [14] provides a comprehensive account of the subject with additional references. Further discussion here is outside the scope of this book.

1.9 OPTIMUM DETECTION: ON–OFF PERFORMANCE MEASURES AND SYSTEM COMPARISONS

A second and equally significant task of Bayes SCT, along with the determination and practical interpretation of the optimal data processing algorithms ($\Lambda(\mathbf{X})$, $\log\Lambda(\mathbf{X})$, $\Lambda_{\text{MAP}}(\mathbf{X})$, etc.), is the evaluation of optimum system performances and performance comparisons with suboptimum receivers $G(\mathbf{X})$. The latter is particularly important because practical systems are themselves never strictly optimum: optimality is an ideal,

to be approached under the inevitable constraints of usually limited knowledge of the environment and bounded economic resources. Nevertheless, optimality and its explicit formulations provide a guide to the key elements of (1) effective practical system design, (2) limiting measures of performance against which the practical system can be compared and often improved, and (3) insights regarding the critical channel structures that inhibit performance. Accordingly, modeling the communication environment, that is, translating the physics of the channel $\left(\hat{\mathbf{T}}^{(N)}\right)$ into relevant mathematical relationships becomes a third major task. This will be treated in detail in Chapters 8 and 9, but it needs to be borne in mind here, because of its ultimate influence on the actual probability measures that constitute the elements of the (Bayes) risks by which performance is evaluated.

Useful measures of performance all depend in some way on the conditional error probabilities $\alpha^*, \beta^*, \alpha, \beta, \ldots$, and so on, cf. Sections 1.6.2 and 1.7 above. Since the error probabilities are also functions of the received signal and of the parameters of the accompanying noise, comparisons of systems performance can also be made in terms of such quantities as well, under a variety of conditions, involving both optimality and suboptimality.

1.9.1 Error Probabilities: Optimum Systems

Our first problem is to provide some way of determining the error probabilities $\alpha^{(*)}$, $\beta^{(*)}$, α_{MAP}, β_{MAP}, and so on, which occur for Bayes and non-Bayes (suboptimum) systems.

We begin with the Bayes class, namely, those described in Sections 1.7 and 1.8 above, where the decision rules $\delta^*(\gamma_0|\mathbf{X}), \delta^*(\gamma_1|\mathbf{X})$ are determined according to (1.7.4a and 1.7.4b). With the help of the transformation $x = \log \Lambda(\mathbf{X})$, cf. (1.7.7a and 1.7.7b), we can write the following expressions for the conditional class probabilities of the Types I and II errors in the Bayesian cases

$$\alpha^* = \int_{\log K}^{\infty} dx \int_{\Gamma} F_J(\mathbf{X}|\mathbf{0})\delta[x - \log \Lambda(\mathbf{X})]\mathbf{dX} = \int_{\log K}^{\infty} Q_1(x)dx \tag{1.9.1a}$$

and

$$\beta^* = \int_{-\infty}^{\log K} dx \int_{\Gamma} \langle F_J(\mathbf{X}|\mathbf{S}(\boldsymbol{\theta}))\rangle_{\mathbf{S} \text{ or } \boldsymbol{\theta}} \delta[x - \log \Lambda(\mathbf{X})]\mathbf{dX} = \int_{-\infty}^{\log K} P_1(x)dx, \tag{1.9.1b}$$

Here Q_1 and P_1 are respectively given by

$$Q_1(x) = \int_{\Gamma} F_J(\mathbf{X}|\mathbf{0})\delta[x - \log \Lambda(\mathbf{X})]\mathbf{dX};\ P_1(x) = \int_{\Gamma} \langle F_J(\mathbf{X}|\mathbf{S}(\boldsymbol{\theta}))\rangle_{\mathbf{S} \text{ or } \boldsymbol{\theta}} \delta[x - \log \Lambda(\mathbf{X})]\mathbf{dX}. \tag{1.9.2}$$

From the fact that $F_J(\mathbf{X}|\mathbf{0})$ and $\langle F_J(\mathbf{X}|\mathbf{S}(\boldsymbol{\theta}))\rangle$ are themselves probability densities and that $x - \log \Lambda(\mathbf{X})$ is also a random variable when considered over the ensemble of possible values of $\mathbf{X}$, it follows that the Q_1, P_1 of (1.9.1a) and (1.9.1b) are the probability densities of x with respect to the distributions associated with the hypothesis states H_0 and H_1, respectively. We shall elaborate on this further below, cf. Eq. (1.9.4a) et. seq.

1.9.1.1 Sufficient Statistics and Monotonic Mapping The mapping for **X**- to x-space by means of the transformation $x - \log \Lambda(\mathbf{X})$ is mathematically quite arbitrary; (see Section 1.7.1). Any monotone function $x = F(\Lambda)$ can be used without altering the values of the error probabilities. This is to be expected, since $F(\Lambda)$, like Λ here, remains a *sufficient statistic*.[41] The analytic consequences of monotonicity in turn are readily demonstrated by the relations

$$\alpha^* \equiv \int_K^\infty q_1(y)^* dy = \int_K^\infty Q_1(F(y))F'(y)dy = \int_{K_{\text{new}}=F(K)}^\infty Q_1(x)dx, x = F(y). \quad (1.9.3)$$

Since $y = \Lambda$ and $x = F(y) = F(\Lambda) = \log \Lambda$ here, $F'(y)dy = dx = d\Lambda/\Lambda$ with $F(K) = \log K$, so that (1.9.1a) results. A similar procedure gives β^*(1.9.1b). In fact, for any monotonic relation $z = G(\mathbf{X})$, where $x = F(G) = F(z)$ it follows that the general (not necessarily optimum) conditional error probabilities α and β can be represented by

$$\begin{gathered}\alpha \equiv \int_K^\infty q_1(z)dz = \int_K^\infty \hat{Q}_1(F)\left(\frac{dF}{dz}\right)dz = \int_{K_{\text{new}}=F(K)}^\infty Q_1(x)dx; \\ \text{etc. for } \beta \equiv \int_{-\infty}^K \hat{p}_1(z)dz = \int_{-\infty}^{K_{\text{new}}=F(K)} \hat{P}_1(x)dx.\end{gathered} \quad (1.9.3a)$$

Although a new threshold $K_{\text{new}} = F(K)$ is established by the transformation F, *the key result is that the error probabilities $\alpha^{(*)}, \beta^{(*)}$ remain unchanged under any monotonic mapping.* Moreover, $\alpha_F^*, \beta^* \to 0$ and $P_D^* \to 1$ as the signal $S \to \infty$ vis-à-vis the accompanying noise, and likewise $(\alpha_F^*, \beta^*) \to 1$ when $S \to 0$. In the optimum cases this also means that $K_{\text{new}}[= K_{\text{new}}(S)]$ depends on the signal in such a way that these limiting results are achieved. The resulting decision process is *consistent* as $S \to \infty$. For suboptimum systems similar results will occur, depending on our choice of test statistic $z = G(\mathbf{X})$: however, not all choices lead to consistency.

Similarly, if several successive monotonic mappings are carried out, that is $x = G_1(y)$ and $z = G_2(x) = G_{21}(y)$, then (1.9.3a) becomes generally

$$\alpha = \int_K^\infty q_1(y)dy = \int_{K_1=G_1(K)}^\infty Q_1(x)dx = \int_{K_2=G_2(K_1)=G_{21}(K)} Q_2(z)dz, \text{ etc.}, \quad (1.9.3b)$$

and for the optimum cases $\alpha \to \alpha^*, y = \Lambda$, and so on. As noted above, monotonicity guarantees that $\alpha^{(*)}$ and $\beta^{(*)}$ remain unchanged in value, although their analytic forms are now different from the original expressions. The practical importance of this is that it very often allows us to evaluate performance analytically, without recourse to numerical methods, by suitable simplifying choices of monotonic transformations (e.g., $x = \log \Lambda$ instead of Λ itself). We shall see several examples employing these general results in Sections 3.2–2, 3, 5 subsequently.

[41] We recall that $\Lambda(\mathbf{X})$ is a sufficient statistic if specifying **X** in addition to $x = \Lambda(\mathbf{X})$ does not in any way increase our knowledge of the signal **S** (which is implicit in Λ, cf. $\langle F_J(\mathbf{X}|\mathbf{S})\rangle$(1.7.2)). Analytically, [Section 22.1.1 of Ref. [1] and, p. 1010] the n.+s. condition that Λ is a sufficient statistic is the requirement that the pdf $w_J(\mathbf{S}|\Lambda) = F_{\text{mono}}[f(\mathbf{S})g(\Lambda)]$, that is, that w_J is a monotonic function of the factors f, g here.

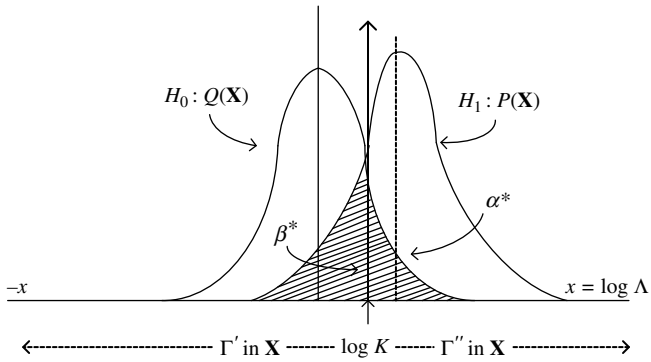


FIGURE 1.9 Probability densities and decision regions for $x = \log\Lambda$, Eq. (1.7.6).

In any case, we see that the δ-functions in (1.9.1a) and (1.9.2) pick out the region Γ^* in $\mathbf{X}$-space for which $\delta^*(\gamma_1|\mathbf{X}) = 1$ in the case of α^*, with a similar interpretation for β^* when $\delta^*(\gamma_0|\mathbf{X}) = 1$. Figure 1.9 shows the two regions in x-space when $x = \log\Lambda(\mathbf{X})$, with the dotted line at $x = \log K$ separating the decision regions for H_0 and H_1.

1.9.1.2 Error Probabilities and Contour Integration [24] Returning to Eqs. (1.9.1a and 1.9.1b) and (1.9.2) and using the integral exponential form for the δ-functions therein, we can write at once the characteristic functions (c.f.s) associated with the pdfs Q_1, P_1

$$F_1(i\xi)_Q = E_{H_0}\left\{e^{i\xi\log\Lambda(\mathbf{X})}\right\} = \int_\Gamma e^{i\xi\log\Lambda(\mathbf{X})}F_J(\mathbf{X}|\mathbf{0})\mathbf{dX} \tag{1.9.4a}$$

$$F_1(i\xi)_P = E_{H_1}\left\{e^{i\xi\log\Lambda(\mathbf{X})}\right\} = \int_\Gamma e^{i\xi\log\Lambda(\mathbf{X})}\langle F_J(\mathbf{X}|\mathbf{S}(\boldsymbol{\theta}))\rangle\mathbf{dX}, \tag{1.9.4b}$$

for which the corresponding pdfs are, from (1.9.2)

$$Q_1(x) = \mathsf{F}^{-1}\left\{F_1(i\xi)_Q\right\} = \int_{-\infty}^{\infty} e^{-i\xi x}\frac{d\xi}{2\pi}\int_\Gamma e^{i\xi\log\Lambda(\mathbf{X})}F_J(\mathbf{X}|\mathbf{0})\mathbf{dX} \tag{1.9.5a}$$

$$P_1(x) = \mathsf{F}^{-1}\left\{F_1(i\xi)_P\right\} = \int_{-\infty}^{\infty} e^{-i\xi x}\frac{d\xi}{2\pi}\int_\Gamma e^{i\xi\log\Lambda(\mathbf{X})}\langle F_J(\mathbf{X}|\mathbf{S}(\boldsymbol{\theta}))\rangle\mathbf{dX}. \tag{1.9.5b}$$

As noted earlier [[1], Eq. (19.32a) and Problem 17.8] there is a simple formal relation between P_1 and Q_1, which follows from the identity

$$\int_\Gamma e^{-i\xi x}F_J(\mathbf{X}|\mathbf{0})\mathbf{dX} \equiv \mu\int_\Gamma e^{(i\xi-1)x}\langle F_J(\mathbf{X}|\mathbf{S})\rangle\mathbf{dX};\ x = \log\Lambda = \frac{\mu\langle F_J(\mathbf{X}|\mathbf{S})\rangle_{\boldsymbol{\theta}}}{F_J(\mathbf{X}|\mathbf{0})} \tag{1.9.6}$$

(which is readily established on using $x = \log\Lambda$ explicitly in (1.9.6)). Thus, from (1.9.4a and 1.9.4b) it is seen at once that $F_1(i\zeta)_Q = \mu e^{-x}F_1(i\xi)_P$, so that using this relation in (1.9.5a and 1.9.5b) gives the desired relation

$$Q_1(X) = \mu e^{-x}P_1(x) \tag{1.9.7}$$

which is sometimes useful in the explicit evaluation of error probabilities, particularly when the (sometimes) easier to evaluate $Q_1(x)$ can be found.

Potentially useful alternative relations for the error probabilities, now in terms of the characteristic functions (1.9.4a and 1.9.4b), can be obtained as follows. Applying (1.9.4a and 1.9.4b) to (1.9.5a and 1.9.5b), and then in (1.9.1a and 1.9.1b), we first extend the domain of ξ by analytic continuation to appropriate regions of the complex ξ-plane, in order to ensure convergence of the integrals $\int_{\log K}^{\infty} e^{-i\xi x}dx$, $\int_{\infty}^{\log K} e^{-i\xi x}dx$ in the reversal of the orders of integration which we employed in the above. The results are then the inverse Fourier transforms

$$\alpha^* = \int_{-\infty-ic}^{\infty-ic} \frac{e^{-i\xi \log K}}{2\pi i\xi} F_1(i\xi)_Q d\xi = \int_{C^{(-)}} \frac{e^{-i\xi \log K}}{2\pi i\xi} F_1(i\xi)_Q d\xi, \tag{1.9.8a}$$

and

$$\beta^* = \int_{-\infty+ic}^{\infty+ic} \frac{e^{-i\xi \log K}}{-2\pi i\xi} F_1(i\xi)_P d\xi = \int_{C^{(+)}} \frac{e^{-i\xi \log K}}{-2\pi i\xi} F_1(i\xi)_P d\xi, \tag{1.9.8b}$$

where $\mathbf{C}^{(-)}, \mathbf{C}^{(+)}$ are respectively contours extending from $-\infty$ to $+\infty$ along the real axis, indented downward and upward about any singularities on this axis, usually at $\xi = 0$, as shown in Fig. 1.10. (We note the equivalence of the contours $[(-\infty \mp ic), (\infty \mp ic)]$ and $\mathbf{C}^{(-)}, \mathbf{C}^{(+)}$, since the contributions of the paths A_0A', A_1A''; $B'', B_1, B'B_0$ vanish at $\mp\infty$..) Simple poles on the ξ-axis or within the rectangular paths $\mathbf{C}^{(-)} + B_0B' + A'A_0$ and $\mathbf{C}^{(+)} + B_1B'' + A''A_1$ are handled in the usual way with the help of Cauchy's theorem,[42] extended to include any branch points by appropriate modification of the contours. For example, Fig. 1.11 shows some equivalent contours when the integrands (1.9.8a and 1.9.8b) contain a branch insert at $(= 0)$.[43] Equivalent contours are also obtained by

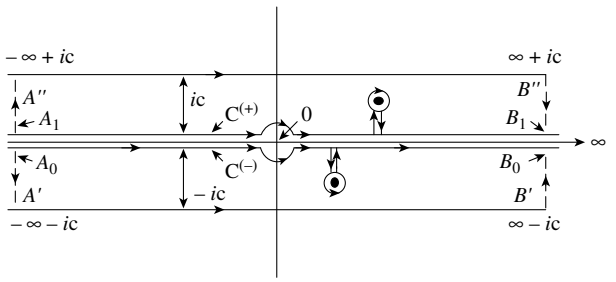


FIGURE 1.10 Equivalent contours of integration for the error probabilities (α^*, β^*), (1.9.8a and 1.9.8b).

[42] We cannot use circular arcs $\xi = pe^{i\phi_\xi}$ since $\rho \to \infty$ and $-\pi \le \phi_\xi \le 0$ or $0 \le \phi_\xi \le \pi$ generally, and have their contributions vanish, leaving $\mathbf{C}^{(\mp)}$. This depends on $F_1/\xi = F = \exp(-\xi^2/2)$, $A_1 > 0$, is a case in point. We can, however, eliminate A_1A'' by letting $0 \to +\infty$, etc., with $\mathbf{C}^{(+)}$ indented upward by ε, about any singularities on the Re ζ-axis, and so on.

[43] A rather extensive discussion of Fourier and Laplace transforms is available in Sections 2.2.4 and 2.2.3 of Ref. [1], including extensive references, along with applications to filters (Chapter 2), rectification, modulation (Chapters 5, 12, 13, 15), and Bayes detection results (Chapters 19, 20, 23), also in Ref. [1]. See Refs. [24] and [41] as well, along with [42, 43] for related analytical tools.

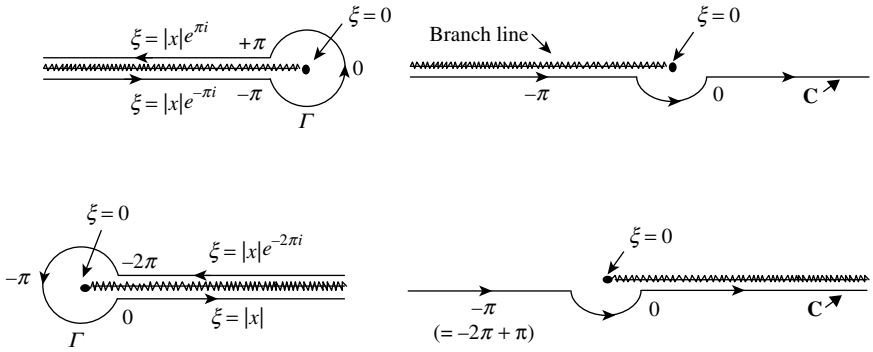


FIGURE 1.11 Some equivalent contours of integration when $F_1(i\xi)/\xi$ contains a branch point at $\xi = 0$; [24, 41].

setting $\xi = s/i$ in the above. The result is a rotation by $-i = e^{-\pi i/2}$ of the ξ-plane contours.[42]

Finally, it is useful for subsequent applications to provide explicit results for determining the (class) error probabilities (α^*, β^*) for the various types of binary Bayes, that is, optimum detectors discussed in Chapter 3, along with general expressions of their associated risks or costs. For this we shall formally employ the c.f.s (1.9.4a and 1.9.4b) to be used in (1.9.8a and 1.9.8b) directly:

$$\alpha^* = \int_{C^{(-)}} \frac{e^{-i\xi \log K}}{2\pi i\xi} F_1(i\xi)_Q d\xi \qquad \beta^* = \int_{C^{(+)}} \frac{e^{-i\xi \log K}}{-2\pi i\xi} F_1(i\xi)_P d\xi, \tag{1.9.9}$$

with the general Bayes case:

(1) *General On–Off Bayes.*

$$[\text{Section 1.7}]\ F_1(i\xi)_Q = \int_\Gamma e^{i\xi \log \Lambda} F_J(\mathbf{X}|\mathbf{0})\mathbf{dx}; \Lambda(\mathbf{X}) = \frac{\mu\langle F_J(\mathbf{X}|\mathbf{S}(\boldsymbol{\theta}))\rangle_{\boldsymbol{\theta}}}{F_J(\mathbf{X}|\mathbf{0})}, (1.7.2),\ (1.7.6), \tag{1.9.9a}$$

$$F_1(i\xi)_P = \int_\Gamma e^{i\xi \log \Lambda} \big\langle F_J(\mathbf{X}|\mathbf{S}(\boldsymbol{\theta}))\big\rangle_{\boldsymbol{\theta}} \mathbf{dx}; \tag{1.9.9b}$$

$$R^*(\sigma, \delta^*) = \mathsf{R}_0 + p(C_\beta - C_{1-\beta})\left(\frac{K}{\mu}\alpha^* + \beta^*\right), (1.7.1),\ (1.7.5), \tag{1.9.9c}$$

with α^*, β^* given by (1.9.1a), via (1.9.4a and 1.9.4b) in (1.9.5a and 1.9.5b).

(2) *Neyman–Pearson.*

$$[\text{Section 1.8.1}]\ \alpha^* = \alpha_{\text{F}} = \int_{C^{(-)}} \frac{e^{-i\xi \log K_{\text{NP}}}}{2\pi i\xi} F_1(i\xi)_Q d\xi, \qquad F_{1Q} = (\text{Eq.1.9.9a}); K = K_{\text{NP}}(\alpha_F) \tag{1.9.10a}$$

$$\beta^*_{\mathrm{NP}} = \int_{C^{(+)}} \frac{e^{-i\xi \log K_{\mathrm{NP}}(\alpha_{\mathrm{F}})}}{-2\pi i\xi} F_1(i\xi)_P d\xi,\ F_{1P} = \text{Eq.(1.9.9b)}; \tag{1.9.10b}$$

$$R^*_{\mathrm{NP}}\left(\sigma, \delta^*_{\mathrm{NP}}\right) = C_0\left[p\beta^*_{\mathrm{NP}} + K_{\mathrm{NP}}(\alpha_{\mathrm{F}})q\alpha_{\mathrm{F}}\right],\ \text{Eqs.(1.7.7a and 1.7.7b), (1.8.5).} \tag{1.9.10c}$$

(3) *Ideal-Observer.*

$$[\text{Section 1.8.2}]\ \alpha^*_I = \int_{C^{(-)}} F_1(i\xi)_Q \frac{d\xi}{2\pi i\xi};\ (1.9.9a)\ \text{for}\ F_1(i\xi)_Q;\ (K=1); \tag{1.9.11a}$$

$$\beta^*_I = \int_{C^{(+)}} F_1(i\xi)_P \frac{d\xi}{-2\pi i\xi},\ (1.9.9b)\ \text{for}\ F_1(i\xi)_P; \tag{1.9.11b}$$

$$R^*_I\left(\sigma, \delta^*_I\right) = C_0\left(q\alpha^*_I + p\beta^*_I\right),\ \text{Eqs. (1.8.7a and 1.8.7b).} \tag{1.9.11c}$$

(4) *Minimax.*

$$[\text{Section 1.8.3}]\ F_1(i\xi)_{Q_{\mathrm{M}}} = \int_\Gamma e^{i\xi \log \Lambda_{\mathrm{M}}} F_J(\mathbf{X}|\mathbf{0})\mathbf{dx};\ \Lambda = \Lambda_{\mathrm{M}} = \frac{\mu_{\mathrm{M}} \langle F_J(\mathbf{X}|\mathbf{S}(\boldsymbol{\theta}_{\mathrm{M}}))\rangle_{\boldsymbol{\theta}_{\mathrm{M}}}}{F_J(\mathbf{X}|\mathbf{0})} \tag{1.9.12a}$$

$$F_1(i\xi)_{P_{\mathrm{M}}} = \int_\Gamma e^{i\xi \log \Lambda_{\mathrm{M}}} \langle F_J(\mathbf{X}|\mathbf{S}(\boldsymbol{\theta}_{\mathrm{M}}))\rangle_{\boldsymbol{\theta}_{\mathrm{M}}} \mathbf{dx},\ \text{cf. (1.9.9a and 1.9.9b)} \tag{1.9.12b}$$

$$R^*\left(\sigma^*_{\mathrm{M}}, \delta^*\right) = \mathsf{R}_0 + p\left(C_\beta - C_{1-\beta}\right)\left(\frac{K}{\mu}\alpha^*_{\mathrm{M}} + \beta^*_{\mathrm{M}}\right); \tag{1.9.12c}$$

$\alpha^* \to \alpha^*_{\mathrm{M}},\ \beta^* \to \beta^*_{\mathrm{M}}$ fixed and determined from (1.9.12a and 1.9.12b) in (1.9.9a and 1.9.9b).

(5) *MAP Detectors, MAP1,2.*

$$\begin{aligned}[\text{Section 1.8.4}]\ F_1(i\xi)_{Q-\mathrm{MAP}_1} &= \int_\Gamma e^{i\xi \log \Lambda_{\mathrm{MAP}_1}} F_J(\mathbf{X}|\mathbf{0})\mathbf{dx};\\ \Lambda(\mathbf{X})_{\mathrm{MAP}_1} &= \frac{\mu F_J\left(\mathbf{X}|\mathbf{S}\left(p\hat{\boldsymbol{\theta}}^*_{p=1}\right)\right)}{F_J(\mathbf{X}|\mathbf{0})},\ \text{cf. Eq. (1.8.16)}\end{aligned} \tag{1.9.13a}$$

$$F_1(i\xi)_{P-\mathrm{MAP}_1} = \int_\Gamma e^{i\xi \log \Lambda_{\mathrm{MAP}_1}} \left\langle F_J(\mathbf{X}|\mathbf{S}(\boldsymbol{\theta}))\right\rangle_{\boldsymbol{\theta}} \mathbf{dx} \tag{1.9.13b}$$

[Here the actual or "true" pdf $w_L(\boldsymbol{\theta})$ is known, in order to obtain $\hat{\boldsymbol{\theta}}^*$ in the maximization process, cf. (1.8.13a) et. seq. Hence $\langle\ \rangle_{\boldsymbol{\theta}}$ involves $w_L(\boldsymbol{\theta})$ in $\langle F_J\rangle_{\boldsymbol{\theta}}$(1.9.13b): $p\hat{\boldsymbol{\theta}}^*(\mathbf{X})_{p=1}$ is the required *unbiased* estimate of $\mathbf{S}$.]

We observe that $\alpha^*_{\mathrm{MAP}_1}$, $\beta^*_{\mathrm{MAP}_1}$, follow from (1.9.13a and 1.9.13b) in (1.9.8), cf. (1.9.9c). Also, that

$$R^*_{\mathrm{MAP}_1} = R^*\left(\sigma, \delta^*_{\mathrm{MAP}_1}\right) = \mathsf{R}_0 + p(C_\beta - C_{1-\beta})\left(\frac{K}{\mu}\alpha^*_{\mathrm{MAP}_1} + \beta^*_{\mathrm{MAP}_1}\right), \text{ cf. (1.8.7) and (1.9.13c);}$$

MAP2.

$$F_1(i\xi)_{Q-\mathrm{MAP}_2} = \int_\Gamma e^{i\xi \log \Lambda_{\mathrm{MAP}_2}} F_J(\mathbf{X}|\mathbf{0})\mathbf{dx};\ \begin{cases} \Lambda(\mathbf{X})_{\mathrm{MAP}_2} = \dfrac{\mu\left\langle F_J\left(\mathbf{X}|\mathbf{S}\left(p\hat{\boldsymbol{\theta}}^*_{p=1|\mathrm{uniform}}\right)\right)\right\rangle}{F_J(\mathbf{X}|\mathbf{0})} \\ \text{cf. (1.8.19) : here } w_L(\boldsymbol{\theta}) \equiv \text{ uniform } \boldsymbol{\theta} \in \Omega_{\boldsymbol{\theta}} \end{cases}, \tag{1.9.14a}$$

$$F_1(i\xi)_{P-\mathrm{MAP}_2} = \int_\Gamma e^{i\xi \log \Lambda_{\mathrm{MAP}_2}} \left\langle F_J(\mathbf{X}|\mathbf{S}(\boldsymbol{\theta}))\right\rangle_{\boldsymbol{\theta}:\ \mathrm{uniform}} \boldsymbol{dX}; \tag{1.9.14b}$$

$(\alpha^*, \beta^*)_{p-\mathrm{MAP}_2}$ follow from (1.9.5a and 1.9.5b) in (1.9.8).

$$R^*_{(\mathrm{uniform}, \delta^*_{\mathrm{MAP}_2})} = R_0 + p(C_\beta - C_{1-\beta})(\frac{K}{\mu}\alpha^* + \beta^*)_{\mathrm{MAP}_2} \tag{1.9.14c}$$

(6) *Sequential Detection.* Here α, β are preset: the test statistic is

$$[\text{Section 1.8.5}]\ \Lambda = \Lambda(\mathbf{X}|J) = \mu \frac{\langle FJ(\mathbf{X}|\mathbf{S}(\boldsymbol{\theta}))\rangle_{\boldsymbol{\theta}}}{F_J(\mathbf{X}|\mathbf{0})} \tag{1.9.15a}$$

$$\left\{ \begin{array}{l} \dfrac{\beta}{1-\alpha} < \Lambda(\mathbf{X}|j) < \dfrac{1-\beta}{\alpha} : \text{ continue test, } j \to J+1; \\ \Lambda(\mathbf{X}|J^*) \geq \dfrac{1-\beta}{\alpha} : \text{ decide } H_1;\ \Lambda\left(\mathbf{X}|J^*\right) < \dfrac{\beta}{1-\alpha} : \text{ decide } H_0 \\ j \to J^* : \text{ terminating sample size.} \end{array} \right\} \tag{1.9.15b}$$

$$R^*_{\mathrm{seq}}\left(\sigma, \delta^*_{\mathrm{seq}}\right) = q\alpha C_\alpha + p\beta C_\beta + pC_0 \min_{\delta \to \delta^*_{\mathrm{seq}}} \left\langle J\left(\mathbf{X}|\mathbf{S}, J\right)^*\right\rangle^*_{\mathbf{X}}. \tag{1.9.15c}$$

In all of the above (except **(6)**) we observe that the various error probabilities $\left(\alpha^*, \beta^*\right)$ are also functions of the prior probabilities (*p, q*), as well as the parameters $\boldsymbol{\theta}$ of the signal and of the noise, through $\log\Lambda(\ )$ of the various optimal detectors above. When it is the signal waveform $\mathbf{S}$ with which we are directly concerned, rather than its parameters $\boldsymbol{\theta}$, we simply replace $\mathbf{S}(\boldsymbol{\theta})$, and so on, with $\mathbf{S}$, or $p\hat{\mathbf{S}}^*$ and so on, cf. **(5)** for the MAP detectors. All of the detectors are optimal within the general Bayesian framework here. However, since they all provide likelihood-ratio tests representative of the level of optimum performance which they demand, they differ in their average costs of decision (Bayes risk). This occurs primarily because of the various constraints imposed upon the signal parameters and their distributions: the more constrained and the more approximative of the actual distributions, the larger the Bayes risk. Thus, ignorance of the true distribution imposes an average risk penalty,

cf. remarks in Section 1.8.3, which can be determined by comparing R^*, (1.9.9c), with the other average risks, (1.9.10c), (1.9.11c), and so on.

1.9.2 Error Probabilities: Suboptimum Systems

The approach of Section 1.9.1 is in no way restricted to optimum systems. For example, in the case of an actual preselected detection system, with a threshold K' (implying at least a cost *ratio*), and a structure represented by $G(\mathbf{X})$ [$\neq \Lambda(\mathbf{X})$ usually] the conditional probabilities of the Type I and Type II errors are now described by analogues of (1.9.1a,b), namely,

$$\alpha = \int_{\log K'}^{\infty} dz \int_{\Gamma} F_J(\mathbf{X}|\mathbf{0})\delta[z - \log G(\mathbf{X})]\mathbf{dX} = \int_{\log K'}^{\infty} q_1(z)dz \tag{1.9.16a}$$

$$\beta = \int_{-\infty}^{\log K'} dz \int_{\Gamma} \langle F_J(\mathbf{X}|\mathbf{S}(\boldsymbol{\theta}))\rangle_{\boldsymbol{\theta} \text{ or } \mathbf{S}} \delta[z - \log G(\mathbf{X})]\mathbf{dX} = \int_{-\infty}^{\log K'} p_1(z)dz. \tag{1.9.16b}$$

The conditional error probabilities α', β', Eq. (1.6.13a), are again obtained on omitting the average $\langle\ \rangle_{\boldsymbol{\theta} \text{ or } \mathbf{S}}$ over parameters or waveform. The distributions (pdfs) of $y = \log G(\mathbf{X})$ are respectively given by (1.9.2), with $\Lambda(\mathbf{X})$ replaced by $G(\mathbf{X})$ under H_0, H_1. The c.f.s of q_1, p_1 are likewise described by

$$F_1(i\xi)_{q_1} = E_{H_0}\left\{e^{i\log G(\mathbf{X})}\right\} = \int_{\Gamma} e^{i\xi \log G(\mathbf{X})} F_J(\mathbf{X}|\mathbf{0})\mathbf{dX} \tag{1.9.17a}$$

$$F_1(i\xi)_{p_1} = E_{H_1}\left\{e^{i\log G(\mathbf{X})}\right\} = \int_{\Gamma} e^{i\xi \log G(\mathbf{X})} \langle F_J(\mathbf{X}|\mathbf{S}(\boldsymbol{\theta}))\rangle_{\boldsymbol{\theta} \text{ or } \mathbf{S}} \mathbf{dX}, \tag{1.9.17b}$$

cf. (1.9.4a and 1.9.4b). [Figure 1.9 applies here also, provided that we replace K by K', α^* by α', Q_1 by q_1, etc.]

The same procedure used to obtain α^* and β^*, (1.9.8a and 1.9.8b) et. seq., applies directly here for α and β. We have directly

$$\alpha = \int_{-\infty - ic'}^{\infty - ic'} \frac{e^{-i\xi \log K'}}{2\pi i\xi} F_1(i\xi)_{q1} d\xi = \int_{C^{(-)'}} \frac{e^{-i\xi \log K'}}{2\pi i\xi} F_1(i\xi)_{q1} d\xi \tag{1.9.18a}$$

$$\beta = \int_{-\infty + ic'}^{\infty + ic'} \frac{e^{-i\xi \log K'}}{-2\pi i\xi} F_1(i\xi)_{p1} d\xi = \int_{C^{(+)'}} \frac{e^{-i\xi \log K'}}{-2\pi i\xi} F_1(i\xi)_{p1} d\xi, \tag{1.9.18b}$$

where $c', \mathbf{C}^{(\pm)'}$ are similar to $c, \mathbf{C}^{(\pm)}$ in (1.9.8a and 1.9.8b) and in Fig. 1.10. Note, however, that the relation (1.9.7) connecting the pdfs Q_1 and P_1 of $x = \log \Lambda(\mathbf{X})$ under H_0 and H_1 does *not* hold for q_1 and p_1, cf. (1.9.16a and 1.9.16b). On the other hand, it is still true if $u = F(z)$, $z = G(\mathbf{X})$, that for any monotonic function $F(z)$, α and β remain unchanged: here $F = \log z = \log G(\mathbf{X})$ specifically, (1.9.3a) above.

Accordingly, with (1.9.16a and 1.9.16b) or (1.9.18a and 1.9.18b), we are able, at least in principle, to determine the average risk $R(\sigma, \delta)$

$$[(1.6.11)]: \qquad R(\sigma,\delta) = \mathsf{R}_0 + q(C_\alpha - C_{1-\alpha})\alpha + p\big(C_\beta - C_{1-\beta}\big)\beta \tag{1.9.19}$$

cf. (1.6.11a and 1.6.11b), and then compare the performance of the suboptimum system $G(\mathbf{X})$, or $\log G$, with that of the corresponding Bayes detectors (Sections 1.8 and 1.9.1), as outlined below.

1.9.3 Decision Curves and System Comparisons

The relations (1.9.1a and 1.9.1b) with the Bayes risk R^*, (1.7.7), and with the average risk R, enable us to compare the performance of actual and optimum systems for the same purpose and of course for the same input signals and noise statistics.

We note first that the error probabilities α^*, β^* and α, β, which appear in R^* and in R, are functions of a_0, the *input signal-to-noise* [(rms) amplitude] *ratio*[44], defined according to $a_0 = \left(\langle S^2\rangle / \langle N^2\rangle\right)^{1/2}$. Curves of average risk as a function of a_0 (or of any other pertinent signal parameters such as sample size J, and other structure parameters) are called *decision curves*. It is in terms of these that specific system comparisons may be made. Figure 1.12 illustrates a typical situation, involving an ideal and an actual detection system for the same purpose.[44] Thus, if we choose the same threshold ($K = K'$) and assign the same costs [Eq. (1.6.6)] to each possible decision, then the average risk R for all a_0 will exceed the corresponding Bayes risk R^*, as indicated.

One definition of *minimum detectable signal*[45] *is that input signal-to-noise ratio* $(a_0)_{\min}$ *that yields an average risk* R_0 *that is some specified fraction of the maximum average risk*, that is, *the* a_0 *for which* $R_0 = \eta R_{\max} (0 < \eta < 1)$. ($R_{\max}$, in physical situations at least, occurs for $a_0 = 0$.) System comparison can now be carried out in a variety of ways, of which the following are some examples (Fig. 1.12):

(1) $(a_0)^*_{\min\eta}$, versus $(a_0)_{\min\eta}$ (in general, $R_0 \neq R_0^*$ for the same η);
(2) $(a_0)^*_{\min\eta_1}$, versus $(a_0)_{\min\eta_1}$ (for $R_0^* = R_0''$, which determines η_1, η_2);
(3) R_0^* versus R' [for the same $(a_0)^*_{\min\eta}$].

Another definition of the *minimum detectable signal*, $\langle a_0^2\rangle^*_{\min}\left(\neq \langle a_0^2\rangle^2_{\min\eta}\right)$ which is explicitly related to detector structure and performance and easier to calculate, is obtained

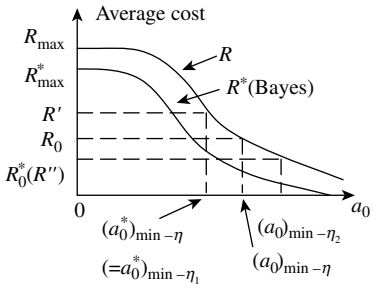


FIGURE 1.12 Typical situations of comparison, showing average and Bayes risks and minimum detectable signals.

[44] Frequently, a_0 is random over the signal class, so that the appropriate ratio is $\bar{a}_0$, or $\overline{a_0^2}$, and so on, depending on the system. See the examples in Section 3.2 ff.

[45] See Sections 19.3.3, 20.3.1, 20.4 of Ref. [1] and Refs. [1, 2] of Chapter 19, Ref. [1] for more details.

from the *detection parameter*. This quantity, in turn, explicitly determines the performance probabilities $P_{\mathrm{D}}^{(*)}$, or $P_{\mathrm{D}}^{(*)}$ and $P_{\mathrm{F}}^{(*)}$ in the optimum and suboptimum cases. It is defined and discussed in detail in Section 3.1.2 and is employed throughout this book.[46]

1.9.3.1 Betting Curves Another decision curve, also useful for comparison, is the *betting curve*, introduced originally by Siegert [40], which relates the probability $W_1(a_0; J)$ of a correct decision (Section 19.3.3 [1]) to the input signal-to-noise ratio a_0. This is defined by

$$W_1(a_0; J) = 1 - (\alpha q + \beta p) = 1 - (P_{\mathrm{F}} + P_{\mathrm{D}}). \tag{1.9.20}$$

For optimum systems Eq. (1.9.20) thus becomes

$$W_1(a_0; J) = 1 - \left(\alpha^* q + \beta^* p\right) = 1 - \left[P_{\mathrm{D}}^* + P_{\mathrm{F}}^*\right]. \tag{1.9.20a}$$

For the Neyman–Pearson and Ideal Observer we may replace α and β by the appropriate α^* and β^* [Eqs. (19.4.1a and 19.4.1b) in [1]], since these systems were shown in Sections 1.8.1 and 1.8.2 to be Bayes with suitable assumptions on the cost ratio (Fig. 1.13).

It is often convenient to use normalized betting curves, defined by ((20.135a and 20.135b) of Ref. [1]), which for the Neyman–Pearson and Ideal Observer become specifically here

$$W_1(a_0; J)_{\mathrm{NP}} = \left(W_1|_{\mathrm{NP}} - q\alpha_{\mathrm{F}}^*\right)/\left(1 - q\alpha_{\mathrm{F}}^*\right), \tag{1.9.21a}$$

$$W_1(a_0; J)_I = [W_{1-I} - (p \text{ or } q)]/[1 - (p \text{ or } q)], \tag{1.9.21b}$$

where (p or q) means that the larger of the two is to be used, and W_1 is given by (1.9.20). The Bayes risks (1.8.2) and (1.8.7a) can be expressed more compactly in terms of the betting curve (1.9.20a) by

$$R_{\mathrm{NP}}^* = C_0\{1 - W_{1-\mathrm{NP}} + \alpha_{\mathrm{NP}} q(K_{\mathrm{NP}} - 1)\}, \tag{1.9.22a}$$

$$R_I^* = C_0\{1 - (W_{1-I})\}, \tag{1.9.22b}$$

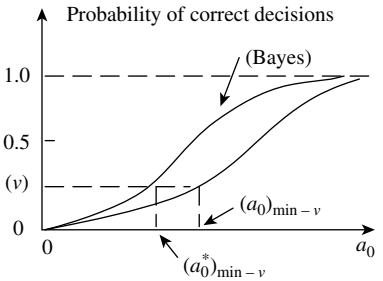


FIGURE 1.13 Betting curves and associated minimum detectable signals.

[46] Examples of $\langle a_0^2 \rangle_{\min}$ are also noted in Sections 20.3 and 20.4 of Ref. [1].

with more involved forms when the W_1s are replaced by their normalized representation. The conditional probabilities may be calculated as before [cf. Eqs. (1.9.1a and 1.9.1b) and (1.9.16a and 1.9.16b)] and may be used in a similar way for system comparison, with the minimum detectable signal defined now in terms of an input which leads to a given percentage ν of successful decisions at the output (cf. Fig. 1.9). Each point on the betting curve, considered as a function of α and β, corresponds to a point on the risk curve, which is also a function of these α, β. Thus, comparisons in terms of "success" are equivalent to those on a risk basis (except for the scale that sets the absolute cost).

1.9.3.2 Performance versus Sample Size An additional description and comparison of system performance, often of considerable interest, is given by the behavior of the minimum detectable signal as a function of the acquisition, or integration time T ($\sim J$, the sample size). This relationship is found from the set of average risk curves [Eq. (1.6.11)] (cf. Fig. 1.12) or the betting curves [Eqs. (1.9.20 and 1.9.20a)], as J assumes all allowed values. The examples considered in Section 20.4, [1] provide some further illustrations.

1.9.3.3 Other Performance Measures Still another variant of the general average risk curve is given by the probability of successfully deciding that a signal is present (i.e., the alternative hypothesis H_1), as a function of input signal-to-noise ratio or other significant system parameters, for example, the decision or "deflection" parameter. Thus, one may use the conditional probability $P_{\mathrm{D}}^{*} = 1 - \beta^{*}$, or the total probability $P_{\mathrm{D}}^{*} = pp_{\mathrm{D}}^{*} = p\left(1 - \beta^{*}\right)$, versus a_0, T (or J), and so on, for optimum systems. Here, usually, $\alpha\left(= \alpha_{\mathrm{F}}^{*}\right)$ is fixed, so that a Neyman–Pearson system is essentially employed. For suboptimum systems, one has similarly $P_{\mathrm{D}} = pp_{\mathrm{D}} = p(1 - \beta)$, versus a_0, and so on, for the desired performance curve and system comparisons.

It is also sometimes convenient to use a normalized decision curve for defining minimum detectable signals and making system comparisons. However, since normalization is an arbitrary procedure, there is no unique or compelling general reason for doing it. In any case, for system comparison care must be taken that common criteria be used under identical conditions. This usually means that comparisons should be made on the basis of the unnormalized or absolute risk curves instead, since normalization may sometimes disguise or diminish significant differences.

We remark, finally, that in the construction, operation, and evaluation of these binary detection processes we have assumed that the signal parameters, *or their average values*, if they are originally random, are known or preset beforehand from a decision curve and that they are then inserted into Λ [Eq. (1.7.2)] so that the scale of Λ can be fixed and an actual test [Eq. (1.7.4a)] carried out with the same parameter values. However, it may be that these "true" parameters, that is, the values actually occurring when a signal is present or average values appropriate to the signal class in question, are not given beforehand, in which case the test of Eq. (1.7.4) can still be carried out, but we are unable to specify the error probabilities α, β uniquely and so cannot determine the Bayes or average risk uniquely. We note some examples of this in the case of the Bayes sequential detectors referred to in Section 1.8.5. For the most part, however, it is not unrealistic to assume at least a knowledge of the required moments of the signal parameters or by some such process as Minimax, to define a class of Bayes receivers for the problem at hand which guards against least favorable situations in some operationally meaningful sense. Unless otherwise indicated, we shall assume henceforth that the appropriate statistics of the signals (and noise) parameters are specified and used.

1.10 BINARY TWO-SIGNAL DETECTION: DISJOINT AND OVERLAPPING HYPOTHESIS CLASSES

The previous on–off analysis here in this chapter is readily extended to the binary two-signal detection cases, where the hypothesis situation is now $H_1 : S_1 \otimes N$ versus $H_2 : S_2 \otimes N$. The two-signal cases are important in many telecommunications applications (Chapter 3 ff.) and when S_1 (or S_2) may represent an interfering or otherwise unwanted signal in radar and sonar environments. It is therefore both useful and instructive to generalize the on–off formalism to include the presence of a signal of class S_2 (in noise) vis-à-vis a signal of Class S_1 (also in noise). Unlike the on–off cases of the preceding sections, where the hypothesis classes are required always to be disjoint [cf. Fig. 1.1b and remarks after Eq. (1.6.3)], there is the additional possibility that the two-signal classes may overlap and thus be nondisjoint. For this latter situation, however, a more sophisticated viewpoint is required [44], as will be seen below.

1.10.1 Disjoint Signal Classes

We consider first the simpler case of disjoint signal classes, where now $\Omega = \Omega_1 + \Omega_2$ and $\Omega_1 \cap \Omega_2$ is empty. In place of (1.6.5), we have

$$\sigma(\mathbf{S}) = p_1 w_1(\mathbf{S}_1) + p_2 w_2(\mathbf{S}_2), \text{ and } \int_{\Omega} \sigma(\mathbf{S}) \mathbf{dS} = 1, \tag{1.10.1}$$

this last as before (cf. remarks after Eq. (1.6.5)), since $p_1 + p_2 = 1$, with p_1 and p_2, respectively, the *a priori* probabilities of a signal of Class 1 (or 2) occurring in the data sample $\mathbf{X}$. Equation (1.6.4) becomes

$$\int_{\Omega_1} w_1(\mathbf{S}_1) \mathbf{dS}_1 = \int_{\Omega_2} w_2(\mathbf{S}_2) \mathbf{dS}_2 = 1, \tag{1.10.2}$$

with $w_1,\ w_2$ the pdfs of $\mathbf{S}_1$ and $\mathbf{S}_2$, or their respective random parameters $\boldsymbol{\theta}_1,\ \boldsymbol{\theta}_2$ in $\mathbf{S}_{1,2}(\boldsymbol{\theta}_{1,2})$. Equation (1.6.3) and cost matrix (1.6.6) are modified in an obvious way to

$$\delta(\gamma_1|\mathbf{X}) + \delta(\gamma_2|\mathbf{X}) = 1;\ \mathbf{C}(\mathbf{S}, \boldsymbol{\gamma}) = \begin{bmatrix} C_1^{(1)} & C_2^{(1)} \\ C_1^{(2)} & C_2^{(2)} \end{bmatrix},\ \boldsymbol{\gamma} = [\gamma_1, \gamma_2]. \tag{1.10.3}$$

now with $C_1^{(1)} < C_2^{(1)};\ C_2^{(2)} < C_1^{(2)}$ to ensure again that "failure" is more expensive than "success" cf. (1.6.6a), and where the upper index as before designates the true state of affairs and the lower the associated decision. The average risk (1.6.7) is accordingly modified to

$$\begin{aligned} R(\sigma, \delta) = \int_{\Gamma} \Big\{ &\Big[p_1 C_1^{(1)} \langle F_J(\mathbf{X}|\mathbf{S}_1) \rangle_1 + p_2 C_1^{(2)} \langle F_J(\mathbf{X}|\mathbf{S}_2) \rangle_2 \Big] \delta(\gamma_1|\mathbf{X}) \\ &+ \Big[p_2 C_2^{(1)} \langle F_J(\mathbf{X}|\mathbf{S}_1) \rangle_1 + p_2 C_2^{(2)} \langle F_J(\mathbf{X}|\mathbf{S}_2) \rangle_2 \Big] \delta(\gamma_2|\mathbf{X}) \} \mathbf{dX} \end{aligned} \tag{1.10.4}$$

in which

$$p_i \langle F_J(\mathbf{X}|\mathbf{S}_i) \rangle_i = \int_{\Gamma} \sigma(\mathbf{S}_i) F_J(\mathbf{X}|\mathbf{S}_i) \mathbf{dS}_i = p_i \int_{\mathbf{S}_i \text{ or } \boldsymbol{\theta}_i} w_i(\mathbf{S}_i \text{ or } \boldsymbol{\theta}_i) F_J(\mathbf{X}|\mathbf{S}_i) \mathbf{dS}_i (\text{or } \mathbf{d}\boldsymbol{\theta}_i),\ i = 1, 2, \tag{1.10.4a}$$

for averages over signal waveform $\mathbf{S}_i$, or parameters $\boldsymbol{\theta}_i$ in $\mathbf{S}_i(\boldsymbol{\theta}_i)$.

The error probabilities are similarly modified:

$$\left.\begin{aligned} \alpha \to \beta_2^{(1)} \equiv \beta_2^{(1)}(\gamma_2|H_1) &= \text{conditional probability of incorrectly deciding that a Class 2} \\ &\quad \text{signal is present, when actually a Class 1 signal occurs;} \\ \beta \to \beta_1^{(2)} \equiv \beta_1^{(2)}(\gamma_1|H_2) &= \text{the reverse of the above.} \end{aligned}\right\} \tag{1.10.5a}$$

Similarly, the respective conditional probabilities of correct decisions are

$$\beta_2^{(2)} = \beta_2^{(2)}\Big(\gamma_2|H_2\Big); \quad \beta_1^{(1)} = \beta_1^{(1)}(\gamma_1|H_1), \tag{1.10.5b}$$

with $p_2\beta_2^{(2)}$, $p_1\beta_1^{(1)}$, $p_2\beta_1^{(2)}$, $p_1\beta_2^{(1)}$ the corresponding total probabilities of correct and incorrect decisions.

The average risk takes the compact forms:

$$R(\sigma,\delta) = \Big(p_1C_1^{(1)} + p_2C_2^{(2)}\Big) + p_2\Big(C_2^{(1)} - C_1^{(2)}\Big)\beta_2^{(1)} + p_2\Big(C_1^{(2)} - C_2^{(2)}\Big)\beta_1^{(2)}, \text{ or} \tag{1.10.6a}$$

$$R(\sigma,\delta) = \Big(p_1C_2^{(1)} + p_2C_1^{(2)}\Big) - p_1\Big(C_2^{(1)} - C_1^{(1)}\Big)\beta_1^{(1)} - p_2\Big(C_1^{(2)} - C_2^{(2)}\Big)\beta_2^{(2)}, \tag{1.10.6b}$$

the former in terms of the *error* probabilities, the latter in terms of the probabilities of correct decision.

1.10.2 Overlapping Hypothesis Classes (F. C. Ogg Jr. [44])

When the signal classes S_1, S_2 are not disjoint but overlap (i.e., $S_1 \cup S_2 \neq 0$), the usual definitions of correct and incorrect decisions are no longer valid, since it is no longer certain whether or not an error has been made. Let us suppose, for example, that signal class S_1 consists of deterministic signals of the type $S(\boldsymbol{\theta}_1)$ and signal class S_2 of the type $S(\boldsymbol{\theta}_2)$, where the waveforms (S) of the two classes are the same and each has the same type of random parameter(s); for example, $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 = \boldsymbol{\theta}$ in both instances represent a common set of random parameters, but with different distribution densities, $w_1(\boldsymbol{\theta}) \neq w_2(\boldsymbol{\theta})$. Any given signal S may belong to either signal class, but S will usually belong to one class with greater probability than to the other. It is reasonable to assign to the more probable decision a lesser cost. Thus, if $\boldsymbol{\theta} = a$ represents a random amplitude, for instance, and if the amplitude a of a particular S lies close to the mean value of $w_1(a)$ but well out on the "tail" of $w_2(a)$, a larger value is assigned to the loss function $\mathsf{F}(S, \gamma_2)$ than to the loss function $\mathsf{F}(S, \gamma_1)$ for the more probably correct decision.

Accordingly, it is clear that the cost assignment should be related to the probability that the signal belongs to *each* of the classes. This can be accomplished in a variety of ways, but the simplest is to require specifically that (1) $\mathsf{F}(S, \gamma)$ be continuous in the prior probabilities (p_1, p_2, w_1, w_2) and (2) that $F(S, \gamma)$ reduce to the usual cost assignments whenever the signal belongs to a disjoint signal class $(S_1 \cap S_2 = 0)$. For the systems considered here, based on constant preset costs, an *extension* of the "constant" cost function F_1, Eq. (1.4.3), satisfying

these conditions is [44]

$$C(S,\gamma_i)=\left[C_1^{(i)}p_1w_1(\boldsymbol{\theta})+C_2^{(i)}p_2w_2(\boldsymbol{\theta})\right]/\sigma(\boldsymbol{\theta})\qquad i=1,\ 2 \tag{1.10.7}$$

where $\sigma(\boldsymbol{\theta})=p_1w_1(\boldsymbol{\theta})+p_2w_2(\boldsymbol{\theta})$ is the prior of $\boldsymbol{\theta}$ for deterministic signals $S(\boldsymbol{\theta})$. In general, we have for signal waveforms

$$C(S,\gamma_i)=\left[C_1^{(i)}p_1w_1(\mathbf{S})+C_2^{(i)}p_2w_2(\mathbf{S})\right]/\sigma(\mathbf{S}) \tag{1.10.8}$$

with $\sigma(\mathbf{S})$ given by Eq. (1.10.1). Thus, by a similar argument Eq. (1.10.8) applies for the case of completely stochastic signals $\mathbf{S}$, where now $w_1(\mathbf{S})\neq w_2(\mathbf{S})$. With Eq. (1.10.7) and (1.10.8) the average risk $R(\sigma,\delta)$ reduces to the original expression (1.10.4) for disjoint classes. Overlapping classes that involve the null signal ($\mathbf{S}=$ noise alone) are handled in the same way, now with

$$C(\mathbf{S},\gamma_i)=\left[C_0^{(i)}qw_0(\mathbf{S})+C_1^{(i)}pw_1(\mathbf{S})\right]/\sigma(\mathbf{S})\qquad i=0,1, \tag{1.10.9}$$

where $\sigma(\mathbf{S})$ is given by (1.6.5) and $R(\sigma,\delta)$ by (1.6.7), and so on. In this way we unite the treatment of overlapping and nonoverlapping signal classes, employing the formalism of the latter as before but now including all the signal types of practical interest.

Let us next calculate the average risk, based on (1.6.7), and determine the minimum average (i.e., Bayes) risk. We observe first that the average risk (1.6.7) must first contain the component exhibited in (1.6.7), here for the two original cases obeying (1.10.7), namely,

$$\begin{aligned}R(\sigma,\delta)=\int_\Gamma&\left[\left\{p_1C_1^{(1)}\left\langle F_J\left(\mathbf{X}\middle|\mathbf{S}^{(1)}\right)\right\rangle+(\,)\right\}+\left\{p_2C_2^{(2)}\left\langle F_J\left(\mathbf{X}\middle|\mathbf{S}^{(2)}\right)\right\rangle+(\,)\right\}\right]\delta(\gamma_1|\mathbf{X})\mathbf{dX}\\&+\left[\left\{p_1C_2^{(1)}\left\langle F_J\left(\mathbf{X}\middle|\mathbf{S}^{(1)}\right)\right\rangle+(\,)\right\}+\left\{p_2C_2^{(2)}\left\langle F_J\left(\mathbf{X}\middle|\mathbf{S}^{(2)}\right)\right\rangle+(\,)\right\}\right]\delta(\gamma_2|\mathbf{X})\mathbf{dX},\end{aligned} \tag{1.10.10a}$$

where the components of the disjoint (i.e., nonoverlapping) component are explicitly given. The quantities $\langle F_J(\mathbf{X}|\mathbf{S}^{(i)})\rangle=\int_\Omega w_i(\boldsymbol{\theta})F_J(\mathbf{X}|\mathbf{S}^{(i)}(\boldsymbol{\theta}))\mathbf{d\theta}$, $i=1,\ 2$, here. The overlap contributions are seen to be from (1.10.7):

$$\left.\begin{aligned}\text{coefficient of }\delta(\gamma_1|\mathbf{X}):\ &C_2^{(1)}p_2\left\langle F_J(\mathbf{X}|\mathbf{S}^{(2)})\right\rangle,\quad C_2^{(1)}p_1\left\langle F_J(\mathbf{X}|\mathbf{S}^{(1)})\right\rangle\\ \text{coefficient of }\delta(\gamma_2|\mathbf{X}):\ &C_1^{(2)}p_2\left\langle F_J(\mathbf{X}|\mathbf{S}^{(2)})\right\rangle,\ C_1^{(2)}p_1\left\langle F_J(\mathbf{X}|\mathbf{S}^{(1)})\right\rangle\end{aligned}\right\}. \tag{1.10.10b}$$

Using the relation $\delta(\gamma_1|\mathbf{X})=1-\delta(\gamma_2|\mathbf{X})$: a decision is always made, and dividing (and multiplying) each term in (1.10.10a) by $qF(\mathbf{X}|\mathbf{0})$, we obtain

$$\begin{aligned}R(\sigma,\delta)=&\int_\Gamma qF_J(\mathbf{X}|\mathbf{0})\left[C_1^{(1)}\Lambda^{(1)}+C_2^{(1)}\Lambda^{(2)}+C_1^{(2)}\Lambda^{(2)}+C_2^{(1)}\Lambda^{(1)}\right]\mathbf{dX}\\&+\int_\Gamma qF_J(\mathbf{X}|\mathbf{0})\left[\left(C_2^{(1)}\Lambda^{(1)}+C_1^{(2)}\Lambda^{(2)}+C_2^{(2)}\Lambda^{(2)}+C_1^{(2)}\Lambda^{(1)}\right)\right.\\&\left.-\left(C_1^{(1)}\Lambda^{(1)}+C_2^{(1)}\Lambda^{(2)}+C_1^{(2)}\Lambda^{(2)}+C_2^{(1)}\Lambda^{(1)}\right)\right]\cdot\delta(\gamma_2|\mathbf{X})\mathbf{dX},\end{aligned} \tag{1.10.11a}$$

where $\Lambda^{(1)} = p_1 \int_\Omega F_J\big(\mathbf{X}\big|\mathbf{S}^{(1)}\big) w_1(\boldsymbol{\theta})\mathbf{d}\boldsymbol{\theta}$, $\Lambda^{(2)} = p_2 \int_\Omega F_J\big(\mathbf{X}\big|\mathbf{S}^{(2)}\big) w_2(\boldsymbol{\theta})\mathbf{d}\boldsymbol{\theta}$. The first term of (1.10.11a) reduces to the irreducible risk

$$\mathsf{R}_{02} \equiv p_1 C_1^{(1)} + p_2 C_2^{(1)} + p_2 C_1^{(2)} + p_1 C_2^{(1)} = p_1\big(C_1^{(1)} + C_2^{(1)}\big) + p_2\big(C_2^{(1)} + C_1^{(2)}\big) > 0, \tag{1.10.11b}$$

Since $F_J\big(\mathbf{X}\big|\mathbf{0}\big) \geq 0$, the second term is clearly minimized when $\delta\big(\gamma_2\big|\mathbf{X}\big) = 1$, that is a (nonrandom) decision is made that signal $\mathbf{S}^{(2)}$ is present, where the expression in [ ] is set equal to zero. This latter gives us the result (collecting $\Lambda^{(2)}$s and $\Lambda^{(1)}$s)

$$\Lambda^{(2)}\big(C_1^{(2)} + C_2^{(2)}\big) + \Lambda^{(1)}\big(C_2^{(1)} + C_1^{(2)}\big) \leq \Lambda^{(1)}\big(C_1^{(1)} + C_2^{(1)}\big) + \Lambda^{(2)}\big(C_2^{(1)} + C_1^{(2)}\big),$$
$$\therefore \Lambda^{(2)}\big(C_2^{(2)} - C_2^{(1)}\big) \leq \Lambda^{(1)}\big(C_1^{(1)} - C_1^{(2)}\big),$$

and since "failure" is more expensive than "success," that is, $C_2^{(1)} - C_2^{(2)}$, $C_1^{(2)} - C_1^{(1)} > 0$, we have finally

$$\delta\big(\gamma_2|\mathbf{X}\big) = 1,\ \therefore \delta\big(\gamma_1|\mathbf{X}\big) = 0, \text{ if } \quad \Lambda^{(2)} > \left(\frac{C_1^{(2)} - C_1^{(1)}}{C_2^{(1)} - C_2^{(2)}}\right)\Lambda^{(1)} \text{or } \Lambda^{(2)} \geq K_{12}\Lambda^{(1)};\ K_{12} > 0 \tag{1.10.12}$$

For the decisions $\delta\big(\gamma_1\big|\mathbf{X}\big) = 1$, we have $\delta\big(\gamma_2\big|0\big) = 0$ and $\Lambda^{(2)} \leq K_{12}\Lambda^{(1)}$. It is to be noted that the effects of overlap Eq. (1.10.10b) leave the decision process unchanged: they are the same as for the nonoverlapping cases. This is a direct consequence of the choice of cost function (1.10.7), which now unites the treatment of both types of signal class (overlapping and nonoverlapping), as stated above. We observe, however, that the irreducible risk R_{02} (1.10.11b) contains four terms rather than two $\big(p_1 C_1^{(1)} + p_2 C_1^{(2)}\big)$. Figure 1.14 shows the decision regions for (1.10.12), in logarithm forms, that is, $\log\Lambda_2$ versus $\log\Lambda_1 + \log K_{12}$.

The case of overlapping classes involving the null signal (1.10.9) follows at once. Setting 2 equal to 1, and 1 equal to 0 in the above gives the result

$$\left.\begin{array}{l} \text{decide } \delta\big(\gamma_1|\mathbf{X}\big) = 1 : \Lambda^{(1)} \geq \big(\dfrac{C_0^{(1)} - C_0^{(0)}}{C_1^{(0)} - C_1^{(1)}}\big), \text{ or } \Lambda^{(2)} \geq K_{01}, K_{01} > 0 \\ \quad \big(\text{and } \delta\big(\gamma_0|\mathbf{X}\big) = 0\big) \\ \text{decide } \delta\big(\gamma_0|\mathbf{X}\big) = 1 : \ \Lambda^{(1)} < K_{01} \\ \quad \big(\text{and } \delta\big(\gamma_1|\mathbf{X}\big) = 0\big) \end{array}\right\}. \tag{1.10.13}$$

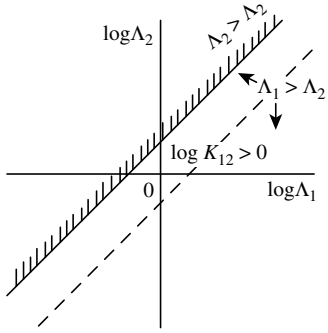


FIGURE 1.14 Decision regions for Λ_1 and Λ_2, for $K_{12} > 1$; for $0 < K_{12} \leq 1$ the boundary lies below the dotted line.

The irreducible risk, however, remains unchanged, namely, (1.10.11b). For stochastic signals (**S**), **S** replaces the deterministic **S**(**θ**) in (1.10.10a)–(1.10.13). Finally, we observe from Section 4.2 following that these results are extendable to the $(K+1)$-ary or K-ary signal classes: the decision process remains the disjoint result, but the irreducible risks and Bayes risks are themselves different.

The error probabilities, average, and Bayes risks, are obtained here from (1.6.8) and Section 1.6.2 generally. The results are

$$\beta^{(1)} = \int_{\Gamma} \left\langle F\big(\mathbf{X}|\mathbf{S}^{(1)}\big)\right\rangle \delta\big(\gamma_2|\mathbf{X}\big)\mathbf{dX}; \quad \beta^{(2)} = \int_{\Gamma} \left\langle F\big(\mathbf{X}|\mathbf{S}^{(2)}\big)\right\rangle \delta\big(\gamma_1|\mathbf{X}\big)\mathbf{dX} \tag{1.10.14}$$

$$\begin{aligned} R\big(\sigma,\delta\big) &= p_1\big(C_1^{(1)} + C_2^{(1)}\big)\big(1-\beta^{(1)}\big) + p_2\big(C_2^{(1)} + C_1^{(2)}\big)\beta^{(2)} \\ &\quad + p_1\big(C_2^{(1)} + C_1^{(2)}\big)\beta^{(1)} + p_2\big(C_1^{(2)} + C_2^{(2)}\big)\big(1-\beta^{(2)}\big) \\ &= p_1\big(C_1^{(2)} - C_1^{(1)}\big)\beta^{(1)} + p_2\big(C_2^{(1)} - C_2^{(2)}\big)\beta^{(2)} + R_{02} \end{aligned} \tag{1.10.15a}$$

and[47] for optimality, we have

$$R\big(\sigma,\delta\big)^{*} = p_1\big(C_1^{(2)} - C_1^{(1)}\big)\beta^{(1^*)} + p_2\big(C_2^{(1)} - C_2^{(2)}\big)\beta^{(2^*)} + R_{02}. \tag{1.10.15b}$$

The evaluation of $\beta^{(1^*)}$, $\beta^{(2^*)}$ associated with the Bayes risk, and in the suboptimum cases, is formally accomplished in Section 1.9. In the case of the null signal, we have

$$R\big(\sigma,\delta\big)^{*} = p_0\big(C_0^{(1)} - C_0^{(0)}\big)\alpha^{*} + p_1\big(C_1^{(0)} - C_1^{(1)}\big)\beta^{*} + R_{01} \tag{1.10.16}$$

with $\big(\alpha^{*},\beta^{*}\big) \to \big(\alpha,\beta\big)$ in the average (nonoptimal) risk.

1.11 CONCLUDING REMARKS

In this first chapter we have obtained some of the principal concepts and techniques used in SCT, based on the fundamental viewpoint of a Bayesian statistical decision theory developed mainly in the mid-twentieth century. A concise topical description of the major elements of both may be gleaned from Sections 1.1–1.10 above, which in turn contain the guiding principles and definitions as well as generic examples. Their implementation is one of the principal aims of the present book, along with the extension to random space-time *fields*. Another is the general use of discrete sampling methods, in conjunction with the physical world of four-dimension, namely space and time, as distinct from earlier analyses devoted to stochastic time processes alone. Thus, to summarize briefly, we employ Sections 1.1–1.5 to provide the formal structure of SDT, basically, a concise description of the fundamental concepts involved. Sections 1.6–1.8 are an illustrative introduction to binary detection and a variety of optimization procedures, with the extension in Section 1.10 to a two-signal binary formulation, in which disjoint and overlapping signal classes are treated.

Optimality, and its approximation, is another goal of the analysis, in conjunction with the Bayesian philosophy used here, with possible constraints imposed by system demands and always subject to the specifics of the physical environment. As will be seen (in Chapters 8, 9

[47] Here we require "failure" always to be more expensive than "success," so that $C_1^{(2)} > C_1^{(1)}, C_2^{(1)} > C_2^{(2)}$.

particularly), the propagation physics needs to be specifically introduced, as it is in many ways a major controlling factor in successful operation. Thus, the physics of the channel in pertinent detail is required. The convenient "black box" approach of additive Gaussian and (often) deterministic interference, with *ad hoc* statistics, loosely based on a postulated random process, in many cases does not represent a full or realistic model of the environment. In Chapter 9, we shall quantitatively describe *non-Gaussain noise* (fields and processes) based on the underlying noise mechanisms, with attention to their spatial as well as their temporal properties. Here our aim is to provide probability distributions (or densities), not just the lower order moments.

In Chapter 2 following, we shall begin this journey from generality to statistical detail by considering first the space–time covariance of a noise field and various conditions that determine its properties. Other relations then follow, in particular the four-dimensional Wiener–Khintchin (W–Kh) theorem.

REFERENCES

1. D. Middleton, *An Introduction to Statistical Communication Theory*, McGraw-Hill, New York, 1960; IEEE Press, Classic Reissue Piscataway, NJ. Chapter 18, 1996. (See also Ref. [19] for other editions.).
2. C. E. Shannon, A Mathematical Theory of Communication, *Bell System Tech. J.*, **27**, 379, 623, 1948.
3. M. G. Kendall, *The Advanced Theory of Statistics*, 5th ed., Vol. II, Griffin, London, 1952.
4. H. Cramér, *Mathematical Methods of Statistics*, Princeton University Press, Princeton, NJ, 1946.
5. R. D. Luce and H. Raiffa, *Games and Decisions*, John Wiley & Sons, Inc., New York, 1957.
6. D. Middleton and D. Van Meter, Detection and Extraction of Signals in Noise from the Point of View of Statistical Decision Theory, *J. Soc. Ind. Appl. Math.*, **3**, 192–253, 1955; **4**, 86–149, 1956.
7. D. Middleton and D. Van Meter, On Optimum Multiple Alternative Detection of Signals in Noise, *IRE Trans. Inform. Theory*, **IT-1**, 1–9, 1955.
8. A. Wald, *Sequential Analysis*, John Wiley & Sons, Inc., New York, 1947.
9. J. Bussgang and D. Middleton, Optimum Sequential Detection of Signals in Noise, *IRE Trans. Inform. Theory*, **IT-1**, 5, 1955.
10. J. L. Doob, *Stochastic Processes*, Chapters 10 and 11, John Wiley & Sons, Inc., New York, 1953.
11. A. Blanc-Lapierre and R. Fortet, *Théorie des Fonctions Aléatoires*, Masson et Cie, Paris, 1953.
12. M. C. Wang and G. E. Uhlenbeck, On the Theory of the Brownian Motion. II, *Rev. Modern Phys.*, **17**, 323, 1945.
13. D. Middleton, Threshold Detection in Correlated Non-Gaussian Noise Fields, *IEEE Trans. Inform. Theory*, **41**(4), 976–1000, 1995.
14. C. W. Helstrom, *Elements of Signal Detection and Estimation*, Prentice-Hall, Englewood Cliffs, NJ, 1995.
15. J. L. Hodges and E. L. Lehmann, Some Problems in Minimax Point Estimation, *Am. Math. Stat.*, **21**, 182, 1950.
16. W. Hoeffding, Optimum Non-parametric Tests, *Proceedings of the 2d Berkeley Symposium on Mathematical Statistics and Probability*, 1950, p. 83.
17. E. L. Lehmann and C. Stein, On the Theory of Some Nonparametric Hypotheses, *Am. Math. Stat.*, **20**, 28, 1949.

18. J. D. Gibson and J. L. Melsa, *Introduction to Nonparametric Detection with Applications*, Academic Press, New York, 1975; reprinted, IEEE Press, Piscataway, NJ, 1996.
19. D. Middleton, *An Introduction to Statistical Communication Theory*, 1st ed., McGraw-Hill, New York, 1960–1972, 2nd Ed., Peninsula Publishing Co., Los Altos, CA, 1987–1996.
20. D. Middleton, *Topics in Communication Theory*, McGraw-Hill, New York, 1965–1972. Peninsula Publishing, Los Altos, CA, 1987.
21. H. V. Poor, *An Introduction to Signal Processing and Estimation*, Springer-Verlag, New York, 1988.
22. H. Blasbalg, The Relationship of Sequential Filter Theory to Information Theory and Its Application to the Detection of Signal in Noise by Bernoulli Trials, *IRE Trans. Inform. Theory* **IT-3**, 122–131, 1957.
23. M. Basseville and I. V. Nikiferov, *Detection of Abrupt Changes*, Prentice Hall, Englewood Cliffs, NJ, 1993.
24. N. W. McLechdaln, *Complex Variable and Operational Calculus*, Cambridge University Press, London and New York, 1939.
25. (a) A. Wald, *Statistical Decision Functions*, John Wiley & Sons, Inc., New York, 1950; (b) A. Wald, Basic Ideas of a General Theory of Statistical Decision Rules, *Proceedings of the International Congress of Mathmaticians*, 1950, p. 231.
26. P. M. Woodward, *Probability and Information Theory with Applications to Radar*, Chapters 4 and 5, McGraw-Hill, New York, 1953.
27. J. Von Neumann, Zur Theorie der Gesellschaftsspiele, *Math. Ann.*, **100**, 295, 1928.
28. J. Von Neumann and O. Morgenstern, *Theory of Games and Economic Behavior*, 3rd ed., Princeton University Press, Princeton, NJ, 1953.
29. R. Radner and J. Marschak, Note on Some Proposed Decision Criteria, in R. M. Thrall, C. H. Coombs,and R. L. Davis (Eds.), *Decision Processes,* Section V, John Wiley & Sons Inc., New York, 1954.
30. J. L. Hodges and E. L. Lehmann, The Use of Previous Experience in Reaching Statistical Decisions, *Ann. Math. Stat.*, **23**, 296, 1952.
31. J. Kiefer, On Wald's Complete Class Theorem," *Ann. Math. Stat.*, **24**, 70, 1953.
32. A. Dvoretzky, A. Wald, and J. Wolfowitz, Elimination of Randomization in Certain Statistical Decision Procedures and Zero-sum Two Person Games, *Ann. Math. Stat.*, **22**, 1, 1951.
33. E. T. Jaynes, in R. D. Rosenkranz (Ed.) *Papers on Probability*, *Statistics and Statistical Physics, D. Reidel*, Dordrecht, Holland, 1983.
34. J. O. Berger, *Statistical Decision Theory and Bayesian Analysis*, 2nd ed., Springer-Verlag, New York, 1985.
35. W. H. Jeffreys and J. O. Berger, Ockham's Razor and Bayesian Analysis, *Am. Scientist*, **80**, 64–73, 1992.
36. D. Middleton, *Topics in Communication Theory*, McGraw-Hill, New York, 1965.
37. C. W. Helstrom, *Statistical Theory of Signal Detection*, 2nd ed. Pergamon Press, New York, 1968 (also Ref. [14] above).
38. W. B. Davenport and W. L. Root, *An Introduction to the Theory of Random Signals and Noise,* McGraw-Hill, New York, 1958.
39. S. S. Wilks, *Mathematical Statistics*, John Wiley & Sons, Inc., New York, 1962.
40. A. J. F. Siegert, Preface and Section 7.5 of J. L. Lawson and G. E. Uhtenbech, *Threshold Signals*, MIT Radiation Laboratory Series, Vol. **24**, McGraw-Hill, New York, 1950.
41. G. A. Campbell and R. M. Foster, *Fourier Integrals for Practical Applications*, 3rd ed., D. Van Nostrand, New York, 1951.

42. J. T. Cushing, *Applied Analytical Mathematics for Physical Scientists,* John Wiley & Sons, New York, 1975.
43. I. S. Gradshteyn and I. H. Rizhik, *Tables of Integrals*, Series and Products, Corrected and Enlarged Edition (Ed. Alan Jeffrey), Academic Press, New York, 1980.
44. F. C. Ogg, Jr., A Note on Bayes Detection of Signals, *IRE Trans. Inform. Theory*, **IT-10**, 57–60, 1964.
45. V. A. Kotelnikov, *The Theory of Optimum Noise Immunity* (Translated from Russian by R.A. Silverman), McGraw-Hill Book Co, NY, 1959.
46. M. Bykhovskiy, *Pioneers of the Information Era*, Technosphera, Moscow, 2006, p. 19.