STOCHASTIC PROCESSES, DETECTION AND ESTIMATION

6.432 Course Notes

Alan S. Willsky, Gregory W. Wornell, and Jeffrey H. Shapiro

Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology
Cambridge, MA 02139

Fall 2003

2

Detection Theory, Decision Theory, and Hypothesis Testing

A wide variety of engineering problems involve making decisions based on a set of measurements. For instance, suppose that in a digital communications system, during a particular interval of time one of two possible waveforms is transmitted to signal a 0-bit or a 1-bit. The receiver then obtains a noisy version of the transmitted waveform, and from this data must determine the bit. Of course, the presence of noise means in general that the decision will not always be correct. However, we would like to use a decision process that is as good as possible in an appropriate sense.

As another example, this time involving air traffic control, suppose that a radar system is set up to detect the presence of an aircraft in the sky. During a particular time interval, a suitably designed radar pulse is transmitted, and if an aircraft is present, this pulse reflects off the aircraft and is received back at the ground. Hence, the presence or absence of such a "return pulse" determines whether an aircraft (or other target) is present. Again the presence of noise in the received signal means that perfect detection is generally not possible.

Still other examples, sometimes quite elaborate, arise in voice and face recognition systems. Given a segment of voice waveform known to come from one of a finite set of speakers, one is often interested in identifying the speaker. Similarly, the problem of identifying a face from an image (i.e., spatial waveform) is also important in a number of applications.

Addressing problems of this type is the aim of detection and decision theory, and a natural framework for setting up such problems is in terms of a hypothesis test. In this framework, each of the possible scenarios corresponds to a hypothesis. When there are M hypotheses, we denote the set of possible hypotheses us-

ing $\{H_0, H_1, \ldots, H_{M-1}\}$.[1] For each of the possible hypotheses, there is a different model for the observed data, and this is what we will exploit to distinguish among the hypotheses.

In this chapter, we will restrict our attention to the case in which the observed data can be represented as a K-dimensional random vector

$$\mathbf{y} = \begin{bmatrix} y_1 & y_2 & \cdots & y_K \end{bmatrix}^{\mathrm{T}}, \tag{2.1}$$

with scalar observations corresponding to the special case $K = 1$. As will become apparent, this case is sufficiently general for a wide range of applications, and allows the key concepts and perspectives to be developed. Extensions to observations that take the form of (infinite-length) random sequences and (continuous-time) random waveforms we postpone until Chapter 6.

In many cases the valid hypothesis can be viewed as a (discrete-valued) random variable, and thus we denote it using H. That is, we can associate *a priori* probabilities

$$P_m = \Pr[H = H_m]$$

with the hypothesis. Our binary communication example fits within this class, with $M = 2$ and the *a priori* probabilities typically being equal. In this case, the model for the observed data under each hypothesis takes the form of a conditional probability density, i.e., $p_{\mathbf{y}|H}(\mathbf{y}|H_m)$ for $m = 0, 1, \ldots, M-1$. As we'll see, in practice these conditional probabilities are often specified implicitly rather than explicitly, and must be inferred from the available information.

In other cases, it is more appropriate to view the valid hypothesis not as a random variable, but as a deterministic but unknown quantity, which we denote simply by H. In these situations, *a priori* probabilities are not associated with the various hypotheses. The radar detection problem mentioned above is one that is often viewed this way, since there is typically no natural notion of the *a priori* probability of an aircraft being present. For these tests, while the valid hypothesis is nonrandom, the observations still are, of course. In this case, the probability density model for the observations is *parameterized* by the valid hypothesis rather than conditioned on it, so these models are denoted using $p_{\mathbf{y}}(\mathbf{y}; H_m)$, for $m = 0, 1, \ldots, M-1$. As in the random hypothesis case, these densities are also often specified implicitly.

This chapter explores methods applicable to both random and nonrandom hypothesis tests. However, we begin by focusing on random hypotheses to develop the key ideas, and restrict attention to the binary ($M = 2$) case.

---

[1] Note that H_0 is sometimes referred to as the "null" hypothesis, particularly in asymmetric problems where it has special significance.

2.1 BINARY RANDOM HYPOTHESIS TESTING: A BAYESIAN APPROACH

In solving a Bayesian binary hypothesis testing problem, two pieces of information are used. One is the set of *a priori* probabilities

$$\begin{aligned} P_0 &= \Pr[\mathsf{H} = H_0] \\ P_1 &= \Pr[\mathsf{H} = H_1] = 1 - P_0. \end{aligned} \tag{2.2}$$

These summarize our state of knowledge about the applicable hypothesis before any observed data is available.

The second is the measurement model, corresponding to the probability densities for $\mathbf{y}$ conditioned on each of the hypotheses, i.e.,

$$\begin{aligned} H_0 &: p_{\mathbf{y}|\mathsf{H}}(\mathbf{y}|H_0) \\ H_1 &: p_{\mathbf{y}|\mathsf{H}}(\mathbf{y}|H_1). \end{aligned} \tag{2.3}$$

The observation densities in (2.3) are often referred to as *likelihood functions*. Our choice of notation suggests that $\mathbf{y}$ is continuous-valued; however, $\mathbf{y}$ can equally well be discrete-valued, in which case the corresponding probability mass functions take the form

$$p_{\mathbf{y}|\mathsf{H}}[\mathbf{y}|H_m] = \Pr[\mathbf{y} = \mathbf{y} \mid \mathsf{H} = H_m]. \tag{2.4}$$

For simplicity of exposition, we start by restricting our attention to the continuous case. Again, it is important to emphasize in many problems this measurement model information is provided indirectly, as the following example illustrates.

Example 2.1

As a highly simplified scenario, suppose a single bit of information $m \in \{0, 1\}$ is to be sent over a communication channel by transmitting the scalar s_m, where s_0 and s_1 are both deterministic, known quantities. Let's further suppose that the channel is noisy; specifically, what is received is

$$\mathsf{y} = s_m + \mathsf{w},$$

where w is, independent of m, a zero-mean Gaussian random variable with variance σ^2. From this information, we can readily construct the probability density for the observation under each of the hypotheses, obtaining:

$$\begin{aligned} p_{\mathsf{y}|\mathsf{H}}(y|H_0) &= N(y; s_0, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y-s_0)^2/(2\sigma^2)} \\ p_{\mathsf{y}|\mathsf{H}}(y|H_1) &= N(y; s_1, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y-s_1)^2/(2\sigma^2)}. \end{aligned} \tag{2.5}$$

In addition, if 0's and 1's are equally likely to be transmitted we would set the *a priori* probabilities to

$$P_0 = P_1 = 1/2.$$

2.1.1 Optimum Decision Rules: The Likelihood Ratio Test

The solution to a hypothesis test is specified in terms of a *decision rule*. We will focus for the time being on *deterministic* decision rules. Mathematically, such a decision rule is a function $\hat{H}(\cdot)$ that uniquely maps every possible K-dimensional observation $\mathbf{y}$ to one of the two hypotheses, i.e., $\hat{H} : \mathbb{R}^K \rightarrow \{H_0, H_1\}$. From this perspective, we see that choosing the function $\hat{H}(\cdot)$ is equivalent to partitioning the observation space $\mathcal{Y} = \{\mathbf{y}\}$ into two disjoint "decision" regions, corresponding to the values of $\mathbf{y}$ for which each of the two possible decisions are made. Specifically, we use $\mathcal{Z}_m$ to denote those values of $\mathbf{y}$ for which our rule decides H_m, i.e.,

$$\begin{aligned} \mathcal{Z}_0 &= \{\mathbf{y} \mid \hat{H}(\mathbf{y}) = H_0\} \\ \mathcal{Z}_1 &= \{\mathbf{y} \mid \hat{H}(\mathbf{y}) = H_1\}. \end{aligned} \tag{2.6}$$

These regions are depicted schematically in Fig. 2.1.

Our goal, then, is to design this bi-valued function (or equivalently the associated decision regions $\mathcal{Z}_0$ and $\mathcal{Z}_1$) in such a way that the best possible performance is obtained. In order to do this, we need to be able to quantify the notion of "best." This requires that we have a well-defined objective function corresponding to a suitable measure of quality. For Bayesian problems we use an objective function taking the form of an expected cost function. Specifically, we use

$$\tilde{C}(H_j, H_i) \triangleq C_{ij} \tag{2.7}$$

to denote the "cost" of deciding that the hypothesis is $\hat{H} = H_i$ when the correct hypothesis is $H = H_j$. Then the optimum decision rule takes the form

$$\hat{H}(\cdot) = \arg\min_{f(\cdot)} J(f) \tag{2.8}$$

where the average cost, which is referred to as the "Bayes risk," is

$$J(f) = E\left[\tilde{C}(H, f(\mathbf{y}))\right], \tag{2.9}$$

and where the expectation in (2.9) is over both $\mathbf{y}$ and H, and $f(\cdot)$ is a generic decision rule.

Often, the context of the specific problem suggests how to choose the costs C_{ij}. For example, a symmetric cost function of the form $C_{ij} = 1 - \delta[i - j]$, i.e.,

$$\begin{aligned} C_{00} &= C_{11} = 0 \\ C_{01} &= C_{10} = 1 \end{aligned} \tag{2.10}$$

corresponds to seeking a decision rule that minimizes the probability of a decision error. However, there are many applications where such symmetric cost functions are not well-matched. For example, in a medical diagnosis problem where H_0 denotes the hypotheses that the patient does not have a particular disease and H_1

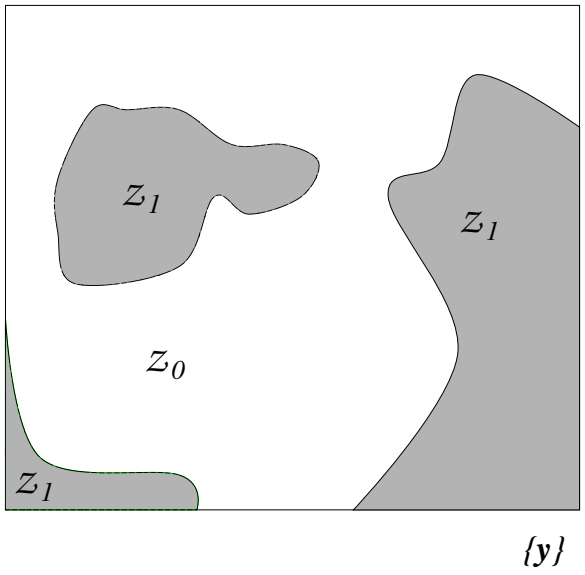


Figure 2.1. The regions $\mathcal{Z}_0$ and $\mathcal{Z}_1$ as defined in (2.6) corresponding to an example decision rule $\hat{H}(\cdot)$.

that he or she does, we would typically want to select cost assignments such that $C_{01} \gg C_{10}$.[2]

Having chosen suitable cost assignments, we proceed to our solution by considering an arbitrary but fixed decision rule $f(\cdot)$. In terms of this generic $f(\cdot)$, the Bayes risk can be expanded in the form

$$\begin{aligned} J(f) &= E\left[\tilde{C}(\mathsf{H}, f(\mathbf{y}))\right] \\ &= E\Big[E\left[\tilde{C}(\mathsf{H}, f(\mathbf{y})) \mid \mathbf{y} = \mathbf{y}\right]\Big] \\ &= \int \tilde{J}(f(\mathbf{y}), \mathbf{y})\, p_{\mathbf{y}}(\mathbf{y})\, d\mathbf{y}, \end{aligned} \tag{2.11}$$

with

$$\tilde{J}(H, \mathbf{y}) = E\left[\tilde{C}(\mathsf{H}, H) \,\Big|\, \mathbf{y} = \mathbf{y}\right], \tag{2.12}$$

and where to obtain the second equality in (2.11) we have used iterated expectation.

From the last equality in (2.11) we obtain a key insight: since $p_{\mathbf{y}}(\mathbf{y})$ is nonnegative, it is clear that we will minimize J if we minimize $\tilde{J}(f(\mathbf{y}), \mathbf{y})$ *for each particular value of* $\mathbf{y}$. The implication here is that we can determine the optimum decision rule $\hat{H}(\cdot)$ on a point by point basis, i.e., $\hat{H}(\mathbf{y})$ for each $\mathbf{y}$.

Let's consider a particular (observation) point $\mathbf{y} = \mathbf{y}_*$. For this point, if we choose the assignment

$$\hat{H}(\mathbf{y}_*) = H_0,$$

[2]In still other problems, it is difficult to make meaningful cost assignments at all. In this case, the Neyman-Pearson framework developed later in the chapter is more natural than the Bayesian framework we develop in this section.

then our conditional expectation (2.12) takes the value

$$\tilde{J}(H_0, \mathbf{y}_*) = C_{00} \Pr[\mathsf{H} = H_0 \mid \mathbf{y} = \mathbf{y}_*] + C_{01} \Pr[\mathsf{H} = H_1 \mid \mathbf{y} = \mathbf{y}_*]. \tag{2.13}$$

Alternatively, if we choose the assignment

$$\hat{H}(\mathbf{y}_*) = H_1,$$

then our conditional expectation (2.12) takes the value

$$\tilde{J}(H_1, \mathbf{y}_*) = C_{10} \Pr[\mathsf{H} = H_0 \mid \mathbf{y} = \mathbf{y}_*] + C_{11} \Pr[\mathsf{H} = H_1 \mid \mathbf{y} = \mathbf{y}_*]. \tag{2.14}$$

Hence, the optimum assignment for the value $\mathbf{y}_*$ is simply the choice corresponding to the smaller of (2.13) and (2.14). It is convenient to express this optimum decision rule using the following notation (now replacing our particular observation $\mathbf{y}_*$ with a generic observation $\mathbf{y}$):

$$\begin{array}{l} C_{00} \Pr[\mathsf{H} = H_0 \mid \mathbf{y} = \mathbf{y}] \\ \quad + C_{01} \Pr[\mathsf{H} = H_1 \mid \mathbf{y} = \mathbf{y}] \end{array} \underset{\hat{H}(\mathbf{y}) = H_0}{\overset{\hat{H}(\mathbf{y}) = H_1}{\gtrless}} \begin{array}{l} C_{10} \Pr[\mathsf{H} = H_0 \mid \mathbf{y} = \mathbf{y}] \\ \quad + C_{11} \Pr[\mathsf{H} = H_1 \mid \mathbf{y} = \mathbf{y}]. \end{array} \tag{2.15}$$

Note that when the two sides of (2.15) are equal, then either assignment is equally good—both have the same effect on the objective function (2.11).

A minor rearrangement of the terms in (2.15) results in

$$(C_{01} - C_{11}) \Pr[\mathsf{H} = H_1 \mid \mathbf{y} = \mathbf{y}] \underset{\hat{H}(\mathbf{y}) = H_0}{\overset{\hat{H}(\mathbf{y}) = H_1}{\gtrless}} (C_{10} - C_{00}) \Pr[\mathsf{H} = H_0 \mid \mathbf{y} = \mathbf{y}]. \tag{2.16}$$

Evidently, the *a posteriori* probabilities, i.e., the probabilities for each of the two hypotheses conditioned on having observed the sample value $\mathbf{y}$ of the random vector $\mathbf{y}$, play an important role in the optimum decision rule (2.16). These probabilities can be readily computed from our measurement models (2.3) together with the *a priori* probabilities (2.2). This follows from a simple application of Bayes' Rule, viz.,

$$\Pr[\mathsf{H} = H_m \mid \mathbf{y} = \mathbf{y}] = \frac{p_{\mathbf{y}|\mathsf{H}}(\mathbf{y}|H_m)\, P_m}{p_{\mathbf{y}|\mathsf{H}}(\mathbf{y}|H_0)\, P_0 + p_{\mathbf{y}|\mathsf{H}}(\mathbf{y}|H_1)\, P_1}. \tag{2.17}$$

Since for any reasonable choice of cost function the cost of an error is higher than the cost of being correct, the terms in parentheses in (2.16) are both nonnegative, so we can equivalently write (2.16) in the form[3]

$$\frac{\Pr[\mathsf{H} = H_1 \mid \mathbf{y} = \mathbf{y}]}{\Pr[\mathsf{H} = H_0 \mid \mathbf{y} = \mathbf{y}]} \underset{\hat{H}(\mathbf{y}) = H_0}{\overset{\hat{H}(\mathbf{y}) = H_1}{\gtrless}} \frac{(C_{10} - C_{00})}{(C_{01} - C_{11})}. \tag{2.18}$$

[3]Technically, we have to be careful about dividing by zero here. To simplify our exposition, however, as we discuss in Section 2.1.2, we will generally restrict our attention to the case where this does not happen.

When we then substitute (2.17) into (2.18) and multiply both sides by P_0/P_1, we obtain the decision rule in its final form, directly in terms of the measurement densities:

$$L(\mathbf{y}) \triangleq \frac{p_{\mathbf{y}|H}(\mathbf{y}|H_1)}{p_{\mathbf{y}|H}(\mathbf{y}|H_0)} \underset{\hat{H}(\mathbf{y})=H_0}{\overset{\hat{H}(\mathbf{y})=H_1}{\gtrless}} \frac{P_0(C_{10}-C_{00})}{P_1(C_{01}-C_{11})} \triangleq \eta. \tag{2.19}$$

The left side of (2.19) is a function of the observed data $\mathbf{y}$ referred to as the *likelihood ratio*—which we denote using $L(\mathbf{y})$—and is constructed from the measurement model. The right side of (2.19)—which we denote using η—is a precomputable threshold which is determined from the *a priori* probabilities and costs. The overall decision rule then takes the form of what is referred to as a *likelihood ratio test* (LRT).

2.1.2 Properties of the Likelihood Ratio Test

Several observations lend valuable insights into the optimum decision rule (2.19). First, note that the likelihood ratio $L(\cdot)$ is a scalar-valued function, i.e., $L : \mathbb{R}^K \to \mathbb{R}$, regardless of the dimension K of the data. In fact, $L(\mathbf{y})$ is an example of what is referred to as a *sufficient statistic* for the problem: it summarizes everything we need to know about the observation vector in order to make a decision. Phrased differently, in terms of our ability to make the optimum decision (in the Bayesian sense in this case), knowledge of $L(\mathbf{y})$ is as good as knowledge of the full data vector $\mathbf{y}$ itself.

We will develop the notion of a sufficient statistic in more detail in subsequent chapters; however, at this point it suffices to make two observations with respect to our detection problem. First, (2.19) tells us an explicit construction for a scalar sufficient statistic for the Bayesian binary hypothesis testing problem. Second, sufficient statistics are not unique. For example, the data $\mathbf{y}$ itself is a sufficient statistic, albeit a trivial one. More importantly, any invertible function of $L(\mathbf{y})$ is also a sufficient statistic. In fact, for the purposes of implementation or analysis it is often more convenient to rewrite the likelihood ratio test in the form

$$\ell(\mathbf{y}) = g(L(\mathbf{y})) \underset{\hat{H}(\mathbf{y})=H_0}{\overset{\hat{H}(\mathbf{y})=H_1}{\gtrless}} g(\eta) = \gamma, \tag{2.20}$$

where $g(\cdot)$ is some suitably chosen, monotonically increasing function. An important example is the case corresponding to $g(\cdot) = \ln(\cdot)$, which simplifies many tests involving densities with exponential factors, such as Gaussians.

It is also important to emphasize that while $L(\mathbf{y})$ is a scalar, $L = L(\mathbf{y})$ is a random variable—i.e., it takes on a different value in each experiment. As such, we will frequently be interested in its probability density function—or at least statistics such as its mean and variance—under each of H_0 and H_1. Such densities can be derived using the method of events discussed in Section 1.5.2 of the previous chapter, and are often used in calculating system performance.

It follows immediately from the definition in (2.19) that the likelihood ratio is a nonnegative quantity. Furthermore, depending on the problem, some values of $\mathbf{y}$ may lead to $L(\mathbf{y})$ being zero or infinite. In particular, the former occurs when $p_{\mathbf{y}|H}(\mathbf{y}|H_1) = 0$ but $p_{\mathbf{y}|H}(\mathbf{y}|H_0) > 0$, which is an indication that values in a neighborhood of $\mathbf{y}$ effectively cannot occur under H_1 but can under H_0. In this case, there will be values of $\mathbf{y}$ for which we'll effectively know with certainty that the correct hypothesis is H_0. When the likelihood ratio is infinite, corresponding a division by zero scenario, an analogous situation exists, but with the roles of H_0 and H_1 reversed. These cases where such perfect decisions are possible are referred to as *singular detection* scenarios. In some practical problems, these scenarios do in fact occur. However, in other cases they suggest a potential lack of robustness in the data modeling, i.e., that some source of inherent uncertainty may be missing from the model. In any event, to simplify our development for the remainder of the chapter we will largely restrict our attention to the case where $0 < L(\mathbf{y}) < \infty$ for all $\mathbf{y}$.

While the likelihood ratio focuses the observed data into a single scalar for the purpose of making an optimum decision, the threshold η for the test plays a complementary role. In particular, from (2.19) we see that η focuses the relevant features of the cost function and *a priori* probabilities into a single scalar. Furthermore, this information is combined in a manner that is intuitively satisfying. For example, as (2.19) also reflects, an increase in P_0 means that H_0 is more likely, so that η is increased to appropriately bias the test toward deciding H_0 for any particular observation. Similarly, an increase in C_{10} means that deciding H_1 when H_0 is true is more costly, so η is increased to appropriately bias the test toward deciding H_0 to offset this risk. Finally, note that adding a constant to the cost function (i.e., to all C_{ij}) has, as we would anticipate, no effect on the threshold. Hence, without loss of generality we may set at least one of the correct decision costs—i.e., C_{00} or C_{11}—to zero.

Finally, it is important to emphasize that the likelihood ratio test (2.19) indirectly determines the decision regions (2.6). In particular, we have

$$\begin{aligned} \mathcal{Z}_0 &= \{\mathbf{y} \mid \hat{H}(\mathbf{y}) = H_0\} = \{\mathbf{y} \mid L(\mathbf{y}) < \eta\} \\ \mathcal{Z}_1 &= \{\mathbf{y} \mid \hat{H}(\mathbf{y}) = H_1\} = \{\mathbf{y} \mid L(\mathbf{y}) > \eta\}. \end{aligned} \tag{2.21}$$

As Fig. 2.1 suggests, while a decision rule expressed in the measurement data space $\{\mathbf{y}\}$ can be complicated,[4] (2.19) tells us that the observations can be transformed into a one-dimensional space defined via $L = L(\mathbf{y})$ where the decision regions have a particularly simple form: the decision $\hat{H}(L) = H_0$ is made whenever L lies to the left of some point on the line, and $\hat{H}(L) = H_1$ whenever L lies to the right.

[4]Indeed, the respective sets $\mathcal{Z}_0$ and $\mathcal{Z}_1$ are not even connected in general, even for the case $K = 1$.

2.1.3 Maximum A Posteriori and Maximum Likelihood Detection

An important cost assignment for many problems is that given by (2.10), which as we recall corresponds to a minimum probability-of-error ($\Pr(e)$) criterion. Indeed, in this case, we have

$$J(\hat{H}) = \Pr\left[\hat{H}(\mathbf{y}) = H_0, \mathsf{H} = H_1\right] + \Pr\left[\hat{H}(\mathbf{y}) = H_1, \mathsf{H} = H_0\right] = \Pr(e).$$

The corresponding decision rule in this case can be obtained by simply specializing (2.19) to obtain

$$\frac{p_{\mathbf{y}|\mathsf{H}}(\mathbf{y}|H_1)}{p_{\mathbf{y}|\mathsf{H}}(\mathbf{y}|H_0)} \underset{\hat{H}(\mathbf{y})=H_0}{\overset{\hat{H}(\mathbf{y})=H_1}{\gtrless}} \frac{P_0}{P_1}. \tag{2.22}$$

Alternatively, we can obtain additional insight by specializing the equivalent test (2.16), from which we obtain a form of the minimum probability-of-error test expressed in terms of the *a posteriori* probabilities for the problem, viz.,

$$\Pr\left[\mathsf{H} = H_1 \mid \mathbf{y} = \mathbf{y}\right] \underset{\hat{H}(\mathbf{y})=H_0}{\overset{\hat{H}(\mathbf{y})=H_1}{\gtrless}} \Pr\left[\mathsf{H} = H_0 \mid \mathbf{y} = \mathbf{y}\right]. \tag{2.23}$$

From (2.23) we see that to minimize the probability of a decision error, we should choose the hypothesis corresponding to the largest *a posteriori* probability, i.e.,

$$\hat{H}(\mathbf{y}) = \underset{A\in\{H_0,H_1\}}{\arg\max} \Pr\left[\mathsf{H} = A \mid \mathbf{y} = \mathbf{y}\right]. \tag{2.24}$$

For this reason, we refer to the test associated with this cost assignment as the *maximum a posteriori* (MAP) decision rule.

Still further simplification is possible when the hypotheses are equally likely ($P_0 = P_1 = 1/2$). In this case, (2.22) becomes simply

$$p_{\mathbf{y}|\mathsf{H}}(\mathbf{y}|H_1) \underset{\hat{H}(\mathbf{y})=H_0}{\overset{\hat{H}(\mathbf{y})=H_1}{\gtrless}} p_{\mathbf{y}|\mathsf{H}}(\mathbf{y}|H_0),$$

and thus we see that our optimum decision rule chooses the hypothesis for which the corresponding likelihood function is largest, i.e.,

$$\hat{H}(\mathbf{y}) = \underset{A\in\{H_0,H_1\}}{\arg\max}\, p_{\mathbf{y}|\mathsf{H}}(\mathbf{y}|A). \tag{2.25}$$

This special case is referred to as the *maximum likelihood* (ML) decision rule. Maximum likelihood detection plays an important role in a large number of applications, and in particular is widely used in the design of receivers for digital communication systems.

Example 2.2

Continuing with Example 2.1, we can obtain from (2.5) that the likelihood ratio test for this problem takes the form

$$L(y) = \frac{\frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y-s_1)^2/(2\sigma^2)}}{\frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y-s_0)^2/(2\sigma^2)}} \underset{\hat{H}(y)=H_0}{\overset{\hat{H}(y)=H_1}{\gtrless}} \eta. \tag{2.26}$$

As (2.26) suggests—and as is generally the case in Gaussian problems—the natural logarithm of the likelihood ratio is a more convenient sufficient statistic to work with in this example. In this case, taking logarithms of both sides of (2.26) yields

$$\ell(y) = \frac{1}{2\sigma^2}\left[(y-s_0)^2 - (y-s_1)^2\right] \underset{\hat{H}(y)=H_0}{\overset{\hat{H}(y)=H_1}{\gtrless}} \ln\eta. \tag{2.27}$$

Expanding the quadratics and cancelling terms in (2.27) we obtain the test in its simplest form, which for $s_1 > s_0$ is given by

$$y \underset{\hat{H}(y)=H_0}{\overset{\hat{H}(y)=H_1}{\gtrless}} \frac{s_1+s_0}{2} + \frac{\sigma^2 \ln\eta}{s_1-s_0} \triangleq \gamma. \tag{2.28}$$

We also remark that with a minimum probability-of-error criterion, if $P_0 = P_1$ then $\ln\eta = 0$ and we see immediately from (2.27) that the optimum test takes the form

$$|y-s_0| \underset{\hat{H}(y)=H_0}{\overset{\hat{H}(y)=H_1}{\gtrless}} |y-s_1|,$$

which corresponds to a "minimum-distance" decision rule, i.e.,

$$\hat{H}(y) = H_{\hat{m}}, \qquad \hat{m} = \arg\min_{m\in\{0,1\}} |y-s_m|.$$

As we'll see later in the chapter, this minimum-distance property holds in multidimensional Gaussian problems as well.

Note too that in this problem the decisions regions on the y-axis have a particularly simple form; for example, for $s_1 > s_0$ we obtain

$$\begin{aligned} \mathcal{Z}_0 &= \{y \mid y < \gamma\} \\ \mathcal{Z}_1 &= \{y \mid y > \gamma\}. \end{aligned} \tag{2.29}$$

In other problems—even Gaussian ones—the decision regions can be more complicated, as our next example illustrates.

Example 2.3

Suppose that a zero-mean Gaussian random variable has one of two possible variances, σ_1^2 or σ_0^2, where $\sigma_1^2 > \sigma_0^2$. Let the costs and prior probabilities be arbitrary. Then the likelihood ratio test for this problem takes the form

$$L(y) = \frac{\frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-y^2/(2\sigma_1^2)}}{\frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-y^2/(2\sigma_0^2)}} \underset{\hat{H}(y)=H_0}{\overset{\hat{H}(y)=H_1}{\gtrless}} \eta.$$

In this problem, it is a straightforward exercise to show that the test simplifies to one of the form

$$|y| \underset{\hat{H}(y)=H_0}{\overset{\hat{H}(y)=H_1}{\gtrless}} \sqrt{2\,\frac{\sigma_0^2\sigma_1^2}{\sigma_1^2-\sigma_0^2}\,\ln\left(\eta\frac{\sigma_1}{\sigma_0}\right)} \triangleq \gamma.$$

Hence, the decision region $\mathcal{Z}_1$ is the union of two disconnected regions in this case, i.e.,

$$\mathcal{Z}_1 = \{y \mid y > \gamma\} \cup \{y \mid y < -\gamma\}.$$

2.1.4 The Operating Characteristic of the Likelihood Ratio Test

In this section, we develop some additional perspectives on likelihood ratio tests that will provide us with further insight on Bayesian hypothesis testing. These perspectives will also be important in our development of nonBayesian tests later in the chapter.

We begin by observing that the performance of any decision rule[5] $\hat{H}(\cdot)$ may be fully specified in terms of two quantities

$$\begin{aligned} P_D &= \Pr\left[\hat{H}(\mathbf{y}) = H_1 \;\middle|\; \mathsf{H} = H_1\right] = \int_{\mathcal{Z}_1} p_{\mathbf{y}|\mathsf{H}}(\mathbf{y}|H_1)\,d\mathbf{y} \\ P_F &= \Pr\left[\hat{H}(\mathbf{y}) = H_1 \;\middle|\; \mathsf{H} = H_0\right] = \int_{\mathcal{Z}_1} p_{\mathbf{y}|\mathsf{H}}(\mathbf{y}|H_0)\,d\mathbf{y}, \end{aligned} \tag{2.30}$$

where $\mathcal{Z}_0$ and $\mathcal{Z}_1$ are the decision regions defined via (2.6). Using terminology that originated in the radar community where H_1 refers to the presence of a target and H_0 the absence, the quantities P_D and P_F are generally referred to as the "detection" and "false-alarm" probabilities, respectively (and, hence, the choice of notation). In the statistics community, by contrast, P_F is referred to as the *size* of the test and P_D as the *power* of the test.

It is worth emphasizing that the characterization in terms of (P_D, P_F) is not unique, however. For example, any invertible linear or affine transformation of the pair (P_D, P_F) is also complete. For instance, the pair of "probabilities of error of the first and second kind" defined respectively via

$$\begin{aligned} P_E^1 &= \Pr\left[\hat{H}(\mathbf{y}) = H_1 \;\middle|\; \mathsf{H} = H_0\right] = P_F \\ P_E^2 &= \Pr\left[\hat{H}(\mathbf{y}) = H_0 \;\middle|\; \mathsf{H} = H_1\right] = 1 - P_D \triangleq P_M \end{aligned} \tag{2.31}$$

constitute such a characterization, and are preferred in some communities that make use of decision theory. As (2.31) indicates, from the radar perspective the

[5]Note that the arbitrary decision rule we consider here need not be optimized with respect to any particular criterion—it might be, but it might also be a heuristically reasonable rule, or even a bad rule.

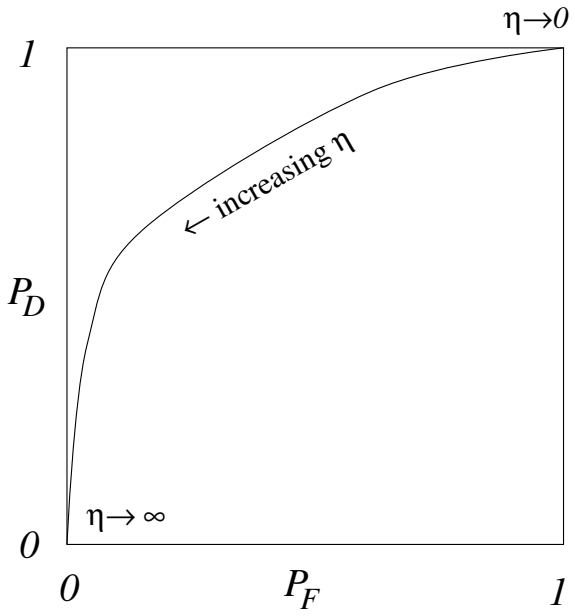


Figure 2.2. Operating characteristic associated with a likelihood ratio test.

probability of error of the first kind is the probability of a false alarm, while probability of error of the second kind is the probability of a miss, which is denoted by P_M.

In general, a good decision rule (detector) is one with a large P_D and a small P_F (or equivalently, small P_E^1 and P_E^2). However, ultimately these are competing objectives. As an illustration of this behavior, let us examine the performance of a likelihood ratio test (2.19) when the threshold η is varied. Note that each choice of η completely specifies a decision rule, with which is associated a particular (P_D, P_F) operating point. Hence, each value of η is associated with a single point in the P_D–P_F plane. Moreover, as η is varied from 0 to ∞, a curve is traced out in this plane as illustrated in Fig. 2.2. This curve is referred to as the *operating characteristic* of the likelihood ratio test.

As Fig. 2.2 suggests, good P_D is generally obtained at the expense of high P_F, and so choosing a threshold η for a particular problem involves making an acceptable tradeoff. Indeed, as $\eta \to 0$ we have $(P_D, P_F) \to (1, 1)$, while as $\eta \to \infty$ we have $(P_D, P_F) \to (0, 0)$. From this perspective, the Bayesian test represents a particular tradeoff, and corresponds to a single point on this curve. To obtain this tradeoff, we effectively selected as our objective function a linear combination of P_D and P_F. More specifically, we performed the optimization (2.8) using

$$J(f) = \alpha P_F - \beta P_D + \gamma,$$

where the choice of α and β is, in turn, determined by the cost assignment (C_{ij}'s) and the *a priori* probabilities (P_m's). In particular, rewriting (2.9) in the form

$$J(f) = \sum_{i,j} C_{ij} \Pr\left[\hat{H}(\mathbf{y}) = H_i \,\middle|\, \mathsf{H} = H_j\right] P_j$$

we obtain

$$\alpha = (C_{10} - C_{00})P_0 \qquad \beta = (C_{01} - C_{11})P_1 \qquad \gamma = (C_{00}P_0 + C_{01}P_1).$$

Let us explore a specific example to gain further insight.

Example 2.4

Let us consider the following special case of our simple scalar Gaussian detection problem from Example 2.1:

$$\begin{aligned} H_0 &: y \sim N(0, \sigma^2) \\ H_1 &: y \sim N(m, \sigma^2), \qquad m \geq 0, \end{aligned} \tag{2.32}$$

which corresponds to choosing $s_0 = 0$ and $s_1 = m$. Specializing (2.28), we see that the optimum decision rule takes the form

$$y \underset{\hat{H}(y)=H_0}{\overset{\hat{H}(y)=H_1}{\gtrless}} \frac{m}{2} + \frac{\sigma^2 \ln \eta}{m} \triangleq \gamma,$$

so that

$$P_D = \int_\gamma^\infty p_{\mathsf{y}|\mathsf{H}}(y|H_1)\, dy \tag{2.33a}$$

$$P_F = \int_\gamma^\infty p_{\mathsf{y}|\mathsf{H}}(y|H_0)\, dy. \tag{2.33b}$$

The expressions (2.33a) and (2.33b) each correspond to tail probabilities in a Gaussian distribution, which are useful to express in the "standard form" described in Section 1.6.2. In particular, we have, in terms of $\mathcal{Q}$-function notation,

$$P_D = \Pr[\mathsf{y} > \gamma \mid \mathsf{H} = H_1] = \Pr\left[\frac{\mathsf{y} - m}{\sigma} > \frac{\gamma - m}{\sigma} \;\middle|\; \mathsf{H} = H_1\right] = \mathcal{Q}\left(\frac{\gamma - m}{\sigma}\right) \tag{2.34a}$$

$$P_F = \Pr[\mathsf{y} > \gamma \mid \mathsf{H} = H_0] = \Pr\left[\frac{\mathsf{y}}{\sigma} > \frac{\gamma}{\sigma} \;\middle|\; \mathsf{H} = H_0\right] = \mathcal{Q}\left(\frac{\gamma}{\sigma}\right). \tag{2.34b}$$

From (2.34a) and (2.34b) we see that as γ is varied, a curve is traced out in the P_D–P_F plane. Moreover, this curve is parameterized by $d = m/\sigma$. The quantity d^2 can be viewed as a "signal-to-noise ratio," so that d is a normalized measure of "distance" between the hypotheses. Several of these curves are plotted in Fig. 2.3. Note that when $m = 0$ ($d = 0$), the hypotheses (2.32) are indistinguishable, and the curve $P_D = P_F$ is obtained. As d increases, better performance is obtained—e.g., for a given P_F, a larger P_D is obtained.[6] Finally, as $d \to \infty$, the P_D–P_F curve approaches the ideal operating characteristic: $P_D = 1$ for all $P_F > 0$.

We will explore generalizations of these results in multidimensional Gaussian problems later in the chapter.

[6]For any reasonable performance criterion, a curve above and to the left of another is always preferable.

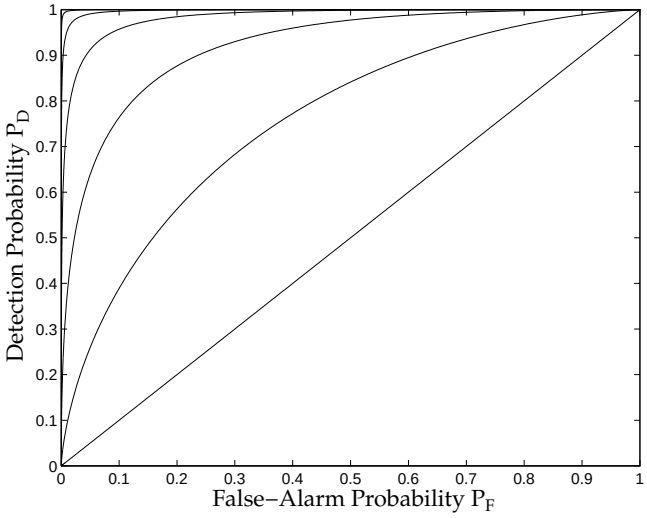


Figure 2.3. Operating characteristic of the likelihood ratio test for the scalar Gaussian detection problem. The successively higher curves correspond to $d = m/\sigma = 0, 1, \ldots, 5$.

One property of the operating characteristic of the likelihood ratio test is that it is monotonically nondecreasing. This follows rather immediately from the structure of these tests. In particular, let $P_D(\eta)$ and $P_F(\eta)$ be the detection and false-alarm probabilities, respectively, for the deterministic test associated with a generic threshold η. Then for any η_1 and η_2 such that $\eta_2 > \eta_1$ we have

$$P_D(\eta_2) \leq P_D(\eta_1) \tag{2.35a}$$

$$P_F(\eta_2) \leq P_F(\eta_1). \tag{2.35b}$$

Hence,

$$\frac{P_D(\eta_1) - P_D(\eta_2)}{P_F(\eta_1) - P_F(\eta_2)} \geq 0.$$

Later in the chapter, we will explore additional properties of the operating characteristic associated with the likelihood ratio test. For example, we'll see that the structure of the likelihood ratio imposes important constraints on the shape of this operating characteristic. Furthermore, we'll relate the operating characteristic to the performance of other kinds of decision rules. For example, to every possible decision rule for a given problem there corresponds an operating point in the P_D–P_F plane. We will develop ways of using the operating characteristic of the likelihood ratio test to bound where the operating points of arbitrary rules may lie. To obtain these results, we need to take a look at decision theory from some additional perspectives. We begin by considering a variation on the Bayesian hypothesis testing framework.

2.2 MIN-MAX HYPOTHESIS TESTING

As we have seen, the Bayesian approach to hypothesis testing is a natural one when we can meaningfully assign not only costs C_{ij} but *a priori* probabilities P_m

as well. However, in a number of applications it may be difficult to determine appropriate *a priori* probabilities. Moreover, we don't want to choose these probabilities arbitrarily—if we use incorrect values for the P_m in designing our optimum decision rule, our performance will suffer.

In this section, we develop a method for making Bayesian hypothesis testing robust with respect to uncertainty in the *a priori* probabilities. Our approach is to construct a decision rule that yields the best possible worst-case performance. As we will see, this corresponds to an optimization based on what is referred to as a "min-max" criterion, which is a powerful and practical strategy for a wide range of detection and estimation problems.

Example 2.5

As motivation, let us reconsider our simple radar or communications scenario (2.32) of Example 2.4. For this problem we showed that the decision rule that minimizes the Bayes risk for a cost assignment $\{C_{ij}\}$ and a set of *a priori* probabilities $\{P_m\}$ reduces to

$$y \underset{\hat{H}(y)=H_0}{\overset{\hat{H}(y)=H_1}{\gtrless}} \frac{m}{2} + \frac{\sigma^2}{m} \ln \left[\frac{(C_{10} - C_{00})P_0}{(C_{01} - C_{11})P_1} \right]. \tag{2.36}$$

In a communication scenario the prior probabilities P_0, P_1 may depend on the characteristics of the information source and may not be under the control of the engineer who is designing the receiver. In the radar problem, it may be difficult to accurately determine the *a priori* probability P_1 of target presence.

For scenarios such as that in Example 2.5, let us assess the impact on performance of using a test optimized for the *a priori* probabilities $\{1-p, p\}$ when the corresponding *true a priori* probabilities are $\{P_0, P_1\}$. The Bayes risk for this "mismatched" system is

$$J(p, P_1) = C_{00}(1-P_1)+C_{01}P_1+(C_{10}-C_{00})(1-P_1)P_F(p)-(C_{01}-C_{11})P_1P_D(p), \tag{2.37}$$

where we have explicitly included the dependence of P_F and P_D on p to emphasize that these conditional probabilities are determined from a likelihood ratio test whose threshold is computed using the incorrect prior p.

The system performance in this situation has a convenient geometrical interpretation, as we now develop. With the notation (2.37), $J(P_1, P_1)$ denotes the Bayes risk when the likelihood ratio test corresponding to the correct priors is used. By the optimality properties of this latter test, we know that

$$J(p, P_1) \geq J(P_1, P_1). \tag{2.38}$$

with, of course, equality if $p = P_1$. Moreover, for a fixed p, the mismatch Bayes risk $J(p, P_1)$ is a linear function of the true prior P_1. Hence, we can conclude, as depicted in Fig. 2.4, that when plotted as functions of P_1 for a particular p, the mismatch risk $J(p, P_1)$ is a line tangent to $J(P_1, P_1)$ at $P_1 = p$. Furthermore, since

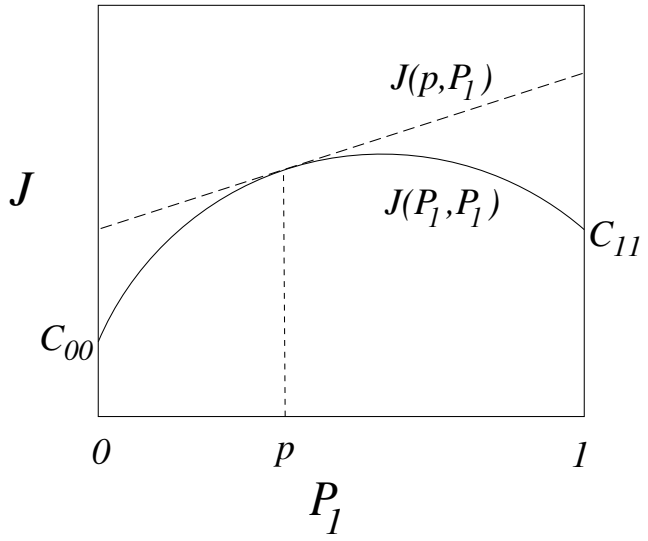


Figure 2.4. Bayes risk objective functions for min-max tests. The solid curve indicates the objective function as a function of the true prior probability for a correctly matched detector—i.e., the detector has correct knowledge of the prior probabilities. The dashed curve indicates the objective function with mismatched detector; for this curve the assumed prior is p.

this must be true for all choices of p, we can further conclude that the optimum risk $J(P_1, P_1)$ must be a concave function of P_1 as Fig. 2.4 also reflects.

From the geometric picture of Fig. 2.4, it is clear that the performance $J(p, P_1)$ obtained using a fixed assumed prior p when the correct prior is P_1, depends on the value of P_1. Moreover, because the mismatch risk is a linear function, we see that the poorest performance, corresponding to worst case mismatch, is obtained either when $P_1 = 0$ or $P_1 = 1$, depending on the sign of the slope of the mismatch risk function. For example, for the p shown in Fig. 2.4, this worst case performance takes place when the true prior is $P_1 = 1$.

Given this behavior, a conservative design strategy is to use a likelihood ratio test based on an assumed prior p chosen so that the worst-case performance is as good as possible. Mathematically, this corresponds to choosing the assumed prior according to a "min-max" criterion: the prior $\hat{P}_1$ obtained in this manner is given by

$$\hat{P}_1 = \arg\min_p \Big\{ \max_{P_1} J(p, P_1) \Big\}. \tag{2.39}$$

This minimizes the sensitivity of $J(p, P_1)$ to variations in P_1, and hence leads to a decision rule that is robust with respect to uncertainty in the prior P_1.

The solution to this min-max problem follows readily from the geometrical picture of Fig. 2.4. Our solution depends on on the details of the shape of $J(P_1, P_1)$ over the range $0 \leq P_1 \leq 1$. There are three cases, which we consider separately.

Case 1: $J(P_1, P_1)$ Monotonically Nonincreasing

An example of this case is depicted in Fig. 2.5(a). In this case, the slope of any line tangent to $J(P_1, P_1)$ is nonpositive, so for any p the maximum of $J(p, P_1)$ lies

at $P_1 = 0$. Hence, as the solution to (2.39) we obtain $\hat{P}_1 = 0$. This corresponds to using a likelihood ratio test with threshold

$$\hat{\eta} = \frac{1 - \hat{P}_1}{\hat{P}_1} \frac{(C_{10} - C_{00})}{(C_{01} - C_{11})} = \infty.$$

This detector makes the decision $\hat{H} = H_0$ regardless of the observed data, and therefore corresponds to the operating point $(P_D, P_F) = (0, 0)$.

Case 2: $J(P_1, P_1)$ Monotonically Nondecreasing

An example of this case is depicted in Fig. 2.5(b). In this case, the slope of any line tangent to $J(P_1, P_1)$ is nonnegative, so for any p the maximum of $J(p, P_1)$ lies at $P_1 = 1$. Hence, as the solution to (2.39) we obtain $\hat{P}_1 = 1$. This corresponds to using a likelihood ratio test with threshold

$$\hat{\eta} = \frac{1 - \hat{P}_1}{\hat{P}_1} \frac{(C_{10} - C_{00})}{(C_{01} - C_{11})} = 0.$$

This detector makes the decision $\hat{H} = H_1$ regardless of the observed data, and therefore corresponds to the operating point $(P_D, P_F) = (1, 1)$.

Case 3: $J(P_1, P_1)$ Nonmonotonic

An example of this case is depicted in our original Fig. 2.4. In this case, $J(P_1, P_1)$ has a point of zero slope (and hence a maximum) at an interior point ($0 < P_1 < 1$), so that $\hat{P}_1$ is the value of p for which the slope of $J(p, P_1)$ is zero.

The corresponding point on the operating characteristic (and in turn, implicitly, the threshold $\hat{\eta}$) can be determined geometrically. In particular, substituting (2.37) with our zero-slope condition, it follows that $\hat{P}_1$ satisfies

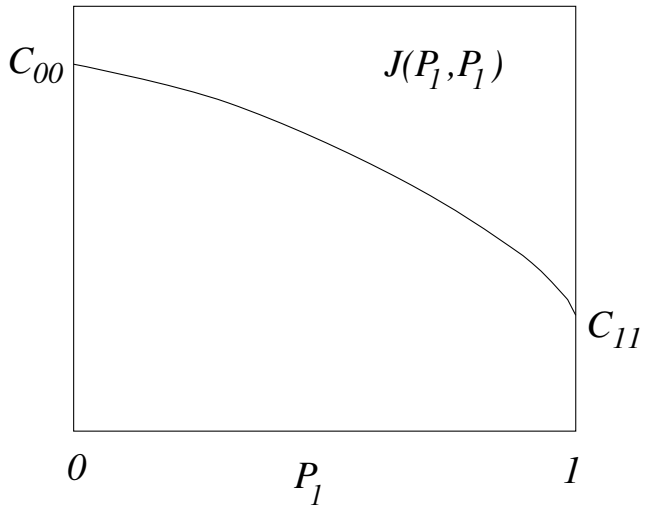
$$\frac{d}{dP_1} J(p, P_1)\bigg|_{p=\hat{P}_1} = (C_{01} - C_{00}) - (C_{10} - C_{00}) P_F(\hat{P}_1) - (C_{01} - C_{11}) P_D(\hat{P}_1) = 0. \quad (2.40)$$

Then (2.40) describes the following line in the P_D–P_F plane

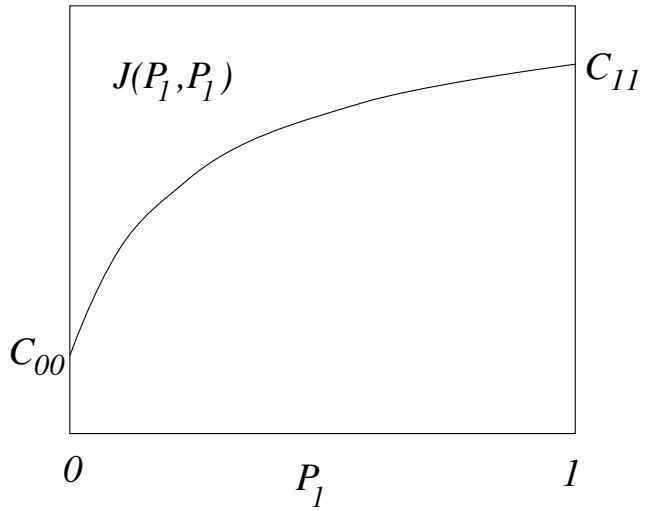
$$P_D = \frac{C_{01} - C_{00}}{C_{01} - C_{11}} - \frac{C_{10} - C_{00}}{C_{01} - C_{11}} P_F, \quad (2.41)$$

which has negative slope as we would expect provided correct decisions are always preferable to incorrect ones, i.e., $C_{ij} > C_{ii}$ for all $i \neq j$,

Hence, its intersection with the operating characteristic for likelihood ratio tests then determines the desired $(\hat{P}_D, \hat{P}_F)$ operating point, and implicitly the corresponding threshold $\hat{\eta}$. Note that to implement the test we never need to explicitly calculate $\hat{P}_1$, just $\hat{\eta}$. Furthermore, that a unique intersection (and hence operating point) exists for such problems follows from the monotonicity of the operating characteristic which we established at the end of Section 2.1.4.



(a)



(b)

Figure 2.5. Examples of possible $J(P_1, P_1)$ curves without points of zero slope.

As a final comment, examining the endpoints of the curve in Fig. 2.4 we see that for some cost assignments we can guarantee that this Case 3 will apply. For example, this happens when $C_{00} = C_{11} = 0$. In this case, we see that the particular operating point on the likelihood ratio test operating characteristic is defined as the point for which the cost of a miss is the same as the cost of a false alarm, i.e.,

$$C_{01}(1 - P_D) = C_{10}P_F. \tag{2.42}$$

As C_{01} and C_{10} are varied relative to one another, the likelihood ratio test threshold $\hat{\eta}$ is varied accordingly.

2.3 NEYMAN-PEARSON BINARY HYPOTHESIS TESTING

Both our basic Bayesian and min-max hypothesis testing formulations require that we choose suitable cost assignments C_{ij}. As we saw, these cost assignments directly influence the (P_D, P_F) operating point of the optimum decision rule. However, in many applications there is no obvious set of cost assignments.

In this kind of situation, an optimization criterion that is frequently more natural is to choose the decision rule so as to maximize P_D subject to a constraint on the maximum allowable P_F, i.e.,

$$\max_{\hat{H}(\cdot)} P_D \text{ such that } P_F \leq \alpha.$$

This is referred to as the Neyman-Pearson criterion. Interestingly, as we will see the likelihood ratio test also plays a key role in the solution for Neyman-Pearson problems. For example, for problems involving continuous-valued data, the optimum decision rule is again a likelihood ratio test with the threshold chosen so that $P_F = \alpha$.

This result can be obtained via the following straightforward Lagrange multiplier approach. As in the Bayesian case, we restrict our attention to deterministic decision rules for the time being. To begin, let $P_F = \alpha' \leq \alpha$, and let us consider minimizing

$$J(\hat{H}) = (1 - P_D) + \lambda(P_F - \alpha')$$

with respect to the choice of $\hat{H}(\cdot)$, where λ is the Lagrange multiplier. To obtain our solution, it is convenient to expand $J(\hat{H})$ in the following form

$$\begin{aligned} J(\hat{H}) &= \int_{\mathcal{Z}_0} p_{\mathbf{y}|H}(\mathbf{y}|H_1)\, d\mathbf{y} + \lambda \left[\int_{\mathcal{Z}_1} p_{\mathbf{y}|H}(\mathbf{y}|H_0)\, d\mathbf{y} - \alpha' \right] \\ &= \int_{\mathcal{Z}_0} p_{\mathbf{y}|H}(\mathbf{y}|H_1)\, d\mathbf{y} + \lambda \left[1 - \int_{\mathcal{Z}_0} p_{\mathbf{y}|H}(\mathbf{y}|H_0)\, d\mathbf{y} - \alpha' \right] \\ &= \lambda(1 - \alpha') + \int_{\mathcal{Z}_0} \Big[p_{\mathbf{y}|H}(\mathbf{y}|H_1) - \lambda p_{\mathbf{y}|H}(\mathbf{y}|H_0) \Big]\, d\mathbf{y}. \end{aligned} \tag{2.43}$$

Now recall from our earlier discussion that specifying $\mathcal{Z}_0$ fully determines $\hat{H}(\cdot)$, so we can view our problem as one of determining the optimum $\mathcal{Z}_0$. From this perspective it is clear we want to choose $\mathcal{Z}_0$ so that it contains precisely those values of $\mathbf{y}$ for which the term in brackets inside the integral in (2.43) is negative, since this choice makes $J(\hat{H})$ smallest. This statement can be expressed in the form

$$p_{\mathbf{y}|H}(\mathbf{y}|H_1) - \lambda p_{\mathbf{y}|H}(\mathbf{y}|H_0) \underset{\hat{H}(\mathbf{y})=H_0}{\overset{\hat{H}(\mathbf{y})=H_1}{\gtrless}} 0,$$

which, in turn, corresponds to a likelihood ratio test; specifically,

$$\frac{p_{\mathbf{y}|H}(\mathbf{y}|H_1)}{p_{\mathbf{y}|H}(\mathbf{y}|H_0)} \underset{\hat{H}(\mathbf{y})=H_0}{\overset{\hat{H}(\mathbf{y})=H_1}{\gtrless}} \lambda \tag{2.44}$$

where λ is chosen so that $P_F = \alpha'$. It remains only to determine α'. However, since the operating characteristic of the likelihood ratio test P_D is a monotonically increasing function of P_F, the best possible P_D is obtained when we let $\alpha' = \alpha$.

In summary, we have seen that the optimum deterministic decision rules for Bayesian, Min-max, and Neyman-Pearson hypothesis testing all take the form of likelihood ratio tests with suitably chosen thresholds. Because such tests arise as the solution to problems with such different criteria, a popular folk theorem is that all reasonable optimization criteria lead to decision rules that can be described in terms of likelihood ratio tests. While this statement is a difficult one to make precise, for most criteria that have been of interest in practice this statement has been true.

And although we have restricted our attention to the case of continuous-valued data, the likelihood ratio test also plays a central role in problems involving discrete-valued data as we will see beginning in Section 2.5. Moreover, for M-ary hypothesis testing problems with $M > 2$, the solutions can be described in terms of natural generalizations of our likelihood ratio tests. Before we proceed explore these topics, however, let us examine the structure of likelihood ratio tests for a number of basic detection problems involving Gaussian data.

2.4 GAUSSIAN HYPOTHESIS TESTING

A large number of detection and decision problems take the form of binary hypothesis tests in which the data are jointly Gaussian under each hypothesis. In this section, we explore some of the particular properties of likelihood ratio tests for these problems. The general scenario we consider takes the form

$$\begin{aligned} H_0 &: \mathbf{y} \sim N(\mathbf{m}_0, \mathbf{\Lambda}_0) \\ H_1 &: \mathbf{y} \sim N(\mathbf{m}_1, \mathbf{\Lambda}_1). \end{aligned} \tag{2.45}$$

Note that $\mathbf{y}$ could be a random vector obtained from some array of sensors, or it could be a collection of samples obtained from a random discrete-time signal $y[n]$, e.g.,

$$\mathbf{y} = \begin{bmatrix} y[0] & y[1] & \cdots & y[N-1] \end{bmatrix}^{\mathrm{T}}. \tag{2.46}$$

For the hypotheses (2.45), and provided $\mathbf{\Lambda}_0$ and $\mathbf{\Lambda}_1$ are nonsingular, the likelihood ratio test takes the form

$$L(\mathbf{y}) = \frac{\frac{1}{(2\pi)^{N/2}|\mathbf{\Lambda}_1|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{y}-\mathbf{m}_1)^{\mathrm{T}}\mathbf{\Lambda}_1^{-1}(\mathbf{y}-\mathbf{m}_1)\right]}{\frac{1}{(2\pi)^{N/2}|\mathbf{\Lambda}_0|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{y}-\mathbf{m}_0)^{\mathrm{T}}\mathbf{\Lambda}_0^{-1}(\mathbf{y}-\mathbf{m}_0)\right]} \underset{\hat{H}(\mathbf{y})=H_0}{\overset{\hat{H}(\mathbf{y})=H_1}{\gtrless}} \eta. \tag{2.47}$$

Some algebraic manipulation allows a simpler sufficient statistic to be obtained for the problem, which corresponds to the test

$$\begin{aligned} \frac{1}{2}\mathbf{y}^{\mathrm{T}}\left[\mathbf{\Lambda}_0^{-1} - \mathbf{\Lambda}_1^{-1}\right]\mathbf{y} & \underset{\hat{H}(\mathbf{y})=H_0}{\overset{\hat{H}(\mathbf{y})=H_1}{\gtrless}} \ln\eta + \frac{1}{2}\ln\left(|\mathbf{\Lambda}_1|/|\mathbf{\Lambda}_0|\right) \\ + \mathbf{y}^{\mathrm{T}}\left[\mathbf{\Lambda}_1^{-1}\mathbf{m}_1 - \mathbf{\Lambda}_0^{-1}\mathbf{m}_0\right] & \qquad + \frac{1}{2}\left[\mathbf{m}_1^{\mathrm{T}}\mathbf{\Lambda}_1^{-1}\mathbf{m}_1 - \mathbf{m}_0^{\mathrm{T}}\mathbf{\Lambda}_0^{-1}\mathbf{m}_0\right]. \end{aligned} \tag{2.48}$$

Further simplification of (2.48) can be obtained in various important special cases. We illustrate two of these via examples.

Example 2.6

Suppose in a spread-spectrum binary communication system, each bit is represented by a code sequence of length K for transmission. Specifically, a 1-bit is signaled via the sequence $m_1[n]$, and a 0-bit via $m_0[n]$, where $n = 0, 1, \ldots, K-1$. At the detector we obtain the following noise corrupted version of the transmitted sequence

$$\begin{aligned} H_0 &: y[n] = m_0[n] + w[n] \\ H_1 &: y[n] = m_1[n] + w[n], \end{aligned} \tag{2.49}$$

where under each hypothesis the noise $w[n]$ is a sequence of independent, identically distributed (IID) Gaussian random variables with mean zero and variance σ^2. Collecting the observed data $y[n]$ for $n = 0, 1, \ldots, N-1$ into a vector $\mathbf{y}$ according to (2.46), we obtain

$$\begin{aligned} H_0 &: \mathbf{y} \sim N(\mathbf{m}_0, \sigma^2\mathbf{I}) \\ H_1 &: \mathbf{y} \sim N(\mathbf{m}_1, \sigma^2\mathbf{I}), \end{aligned} \tag{2.50}$$

where

$$\begin{aligned} \mathbf{m}_0 &\triangleq \begin{bmatrix} m_0[0] & m_0[1] & \cdots & m_0[K-1] \end{bmatrix}^{\mathrm{T}} \\ \mathbf{m}_1 &\triangleq \begin{bmatrix} m_1[0] & m_1[1] & \cdots & m_1[K-1] \end{bmatrix}^{\mathrm{T}}. \end{aligned} \tag{2.51}$$

Defining

$$\Delta\mathbf{m} = \begin{bmatrix} \Delta m[0] & \Delta m[1] & \cdots & \Delta m[K-1] \end{bmatrix}^{\mathrm{T}} = \mathbf{m}_1 - \mathbf{m}_0, \tag{2.52}$$

we see from simplifying (2.48) that a sufficient statistic for making the optimum Bayesian decision at the detector (regardless of the cost and *a priori* probability assignments) is

$$\ell(\mathbf{y}) = \mathbf{y}^{\mathrm{T}}\,\Delta\mathbf{m} = \sum_{n=0}^{K-1} y[n]\,\Delta m[n], \tag{2.53}$$

and the test is

$$\ell(\mathbf{y}) \underset{\hat{H}(\mathbf{y})=H_0}{\overset{\hat{H}(\mathbf{y})=H_1}{\gtrless}} \eta' \triangleq \sigma^2 \ln \eta + \frac{1}{2}\left[\mathbf{m}_1^{\mathrm{T}}\mathbf{m}_1 - \mathbf{m}_0^{\mathrm{T}}\mathbf{m}_0\right]. \tag{2.54}$$

Note that the correlation computation (2.53) that defines $\ell(\mathbf{y})$ is equivalent to a "convolution and sample" operation. Specifically, it is a straightforward exercise to verify that ℓ can be expressed in the form

$$\ell = (y[n] * h[n])\Big|_{n=K-1}, \tag{2.55}$$

where the filter unit-sample response is defined via

$$h[n] = \begin{cases} 0 & n \geq K \\ \Delta m[K-1-n] & 0 \leq n \leq K-1 \\ 0 & n \leq -1. \end{cases} \tag{2.56}$$

This detector is an example of a *matched filter*, a concept we will explore in detail in Chapter 6. Whether the direct computation of ℓ as the correlation (2.53) or its computation via (2.55) is more efficient depends on the particular implementation. It is also straightforward to obtain expressions for the performance of this detector; we leave these as an exercise.

To develop further insight, let us consider the generalization of Example 2.6 to the case of an arbitrary noise covariance, so our hypotheses are

$$\begin{aligned} H_0 &: \mathbf{y} \sim N(\mathbf{m}_0, \mathbf{\Lambda}) \\ H_1 &: \mathbf{y} \sim N(\mathbf{m}_1, \mathbf{\Lambda}), \end{aligned} \tag{2.57}$$

where $\mathbf{\Lambda} = \mathbf{\Lambda}_1 = \mathbf{\Lambda}_0$ is the covariance matrix. In this case the likelihood ratio test simplifies to

$$(\mathbf{y}-\mathbf{m}_0)^{\mathrm{T}}\mathbf{\Lambda}^{-1}(\mathbf{y}-\mathbf{m}_0) \underset{\hat{H}(\mathbf{y})=H_0}{\overset{\hat{H}(\mathbf{y})=H_1}{\gtrless}} (\mathbf{y}-\mathbf{m}_1)^{\mathrm{T}}\mathbf{\Lambda}^{-1}(\mathbf{y}-\mathbf{m}_1) + 2\ln\eta \tag{2.58}$$

or

$$\ell'(\mathbf{y}) \triangleq (\mathbf{m}_1-\mathbf{m}_0)^{\mathrm{T}}\mathbf{\Lambda}^{-1}\mathbf{y} \underset{\hat{H}(\mathbf{y})=H_0}{\overset{\hat{H}(\mathbf{y})=H_1}{\gtrless}} \frac{1}{2}\left(2\ln\eta + \mathbf{m}_1^{\mathrm{T}}\mathbf{\Lambda}^{-1}\mathbf{m}_1 - \mathbf{m}_0^{\mathrm{T}}\mathbf{\Lambda}^{-1}\mathbf{m}_0\right) \triangleq \eta'. \tag{2.59}$$

Useful insights are obtained from a geometrical interpretation of our results. To develop this geometry, we begin by defining an inner product[7]

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^{\mathrm{T}}\mathbf{\Lambda}^{-1}\mathbf{y} \tag{2.60}$$

and the associated induced norm

$$\|\mathbf{y}\| = \sqrt{\langle \mathbf{y}, \mathbf{y} \rangle} = \sqrt{\mathbf{y}^{\mathrm{T}}\mathbf{\Lambda}^{-1}\mathbf{y}}. \tag{2.61}$$

[7]We leave it as an exercise for you to verify that we have defined a valid inner product.

Then, our optimum decision rule (2.59) corresponds to comparing a projection

$$\ell'(\mathbf{y}) = \langle \mathbf{y}, \Delta\mathbf{m} \rangle \tag{2.62}$$

against a threshold, where, again, $\Delta\mathbf{m} = \mathbf{m}_1 - \mathbf{m}_0$.

Specializing further, note that in the case of a minimum probability of error cost assignment and equally likely hypotheses, it is clear from using (2.61) with (2.58) that the maximum likelihood decision rule is equivalent to a minimum distance rule, i.e.,

$$\hat{H}(\mathbf{y}) = H_{\hat{m}} \quad \text{where} \quad \hat{m} = \underset{m \in \{0,1\}}{\arg\min} \|\mathbf{y} - \mathbf{m}_m\|,$$

where our distance metric is defined via our norm (2.61). In addition, via (2.59), when $\|\mathbf{m}_0\| = \|\mathbf{m}_1\|$, solving for the minimum distance is equivalent to solving for the maximum projection; in particular, specializing (2.59) we obtain

$$\hat{H}(\mathbf{y}) = H_{\hat{m}} \quad \text{where} \quad \hat{m} = \underset{m \in \{0,1\}}{\arg\max} \langle \mathbf{y}, \mathbf{m}_m \rangle ,$$

where our inner product is that defined in (2.62). In our vector space $\{\mathbf{y}\}$, the corresponding decisions regions are separated by a hyperplane equidistant from $\mathbf{m}_0$ and $\mathbf{m}_1$ and perpendicular to the line connecting them.

The performance of these likelihood ratio tests can be readily calculated since $\ell'(\mathbf{y})$ is a linear function of the Gaussian random vector $\mathbf{y}$ under each hypothesis. As a result, $\ell'(\mathbf{y})$ is Gaussian under each hypothesis. Specifically,

$$\begin{aligned} H_0 &: \ell' \sim N(\mathbf{m}'_0, \sigma^2_{\ell'}) \\ H_1 &: \ell' \sim N(\mathbf{m}'_1, \sigma^2_{\ell'}), \end{aligned} \tag{2.63}$$

where

$$\mathbf{m}'_m = \langle \Delta\mathbf{m}, \mathbf{m}_m \rangle \tag{2.64}$$

$$\sigma_{\ell'} = \|\mathbf{m}_1 - \mathbf{m}_0\|. \tag{2.65}$$

From here our performance evaluation is identical to the scalar case. In particular, we have

$$\begin{aligned} P_D &= \Pr\left[\ell' \geq \eta' \mid \mathsf{H} = H_1\right] \\ &= \Pr\left[\frac{\ell' - \mathbf{m}'_1}{\sigma_{\ell'}} \geq \frac{\eta' - \mathbf{m}'_1}{\sigma_{\ell'}} \;\middle|\; \mathsf{H} = H_1\right] = \mathcal{Q}\left(\frac{\eta' - \mathbf{m}'_1}{\sigma_{\ell'}}\right) \qquad (2.66a) \\ P_F &= \Pr\left[\ell' \geq \eta' \mid \mathsf{H} = H_0\right] \\ &= \Pr\left[\frac{\ell' - \mathbf{m}'_0}{\sigma_{\ell'}} \geq \frac{\eta' - \mathbf{m}'_0}{\sigma_{\ell'}} \;\middle|\; \mathsf{H} = H_0\right] = \mathcal{Q}\left(\frac{\eta' - \mathbf{m}'_0}{\sigma_{\ell'}}\right). \qquad (2.66b) \end{aligned}$$

While computing actual values of P_D and P_F requires that we evaluate the $\mathcal{Q}(\cdot)$ function numerically, we can obtain bounds on these quantities that are explicitly computable. For example, using (1.129) and (1.128) with (2.66) we obtain, for

thresholds in the range $\mathbf{m}_0' < \eta' < \mathbf{m}_1'$,

$$P_D \geq 1 - \frac{1}{2}\exp\left[-\frac{1}{2}\left(\frac{\eta' - \mathbf{m}_1'}{\sigma_{\ell'}}\right)^2\right] \tag{2.67a}$$

$$P_F \leq \frac{1}{2}\exp\left[-\frac{1}{2}\left(\frac{\eta' - \mathbf{m}_0'}{\sigma_{\ell'}}\right)^2\right]. \tag{2.67b}$$

As in Example 2.6, the sufficient statistic for this problem can be implemented either directly as a correlator, or as a matched filter. In the latter case, the filter $h[n]$ captures the relevant information about the signal and the noise, taking the form

$$h[n] = \begin{cases} 0 & n \geq K \\ \Delta m'[K-1-n] & 0 \leq n \leq K-1 \\ 0 & n \leq -1, \end{cases} \tag{2.68}$$

where

$$\Delta\mathbf{m}' = \begin{bmatrix}\Delta m'[0] & \Delta m'[1] & \cdots & \Delta m'[K-1]\end{bmatrix}^{\mathrm{T}} = \mathbf{\Lambda}^{-1}(\mathbf{m}_1 - \mathbf{m}_0). \tag{2.69}$$

Let us next consider an example of a different class of Gaussian detection problems.

Example 2.7

Suppose that we are trying to detect a random signal $x[n]$ at an antenna based on observations of the samples $n = 0, 1, \ldots K-1$. Depending on whether the signal is absent or present, the noisy observations $y[n]$ take the form

$$\begin{aligned} H_0 &: y[n] = w[n] \\ H_1 &: y[n] = x[n] + w[n], \end{aligned} \tag{2.70}$$

where under both hypotheses $w[n]$ is an IID sequence of $N(0, \sigma^2)$ random variables that is independent of the sequence $x[n]$.

If $x[n]$ was a known signal $x[n]$, then our optimum detector would be that determined in Example 2.6 with $\Delta m[n] = x[n]$; i.e., our sufficient statistic is obtained by correlating our received signal $y[n]$ with $x[n]$. However, in this example we assume we know only the statistics of $x[n]$—specifically, we know that

$$\mathbf{x} = \begin{bmatrix}x[0] & x[1] & \cdots & x[K-1]\end{bmatrix}^{\mathrm{T}} \sim N(\mathbf{0}, \mathbf{\Lambda}_{\mathbf{x}})$$

where $\mathbf{\Lambda}_{\mathbf{x}}$ is the signal covariance matrix.

In this case, we have that the likelihood ratio test (2.48) simplifies to the test

$$\ell(\mathbf{y}) \triangleq \mathbf{y}^{\mathrm{T}}\hat{\mathbf{x}}(\mathbf{y}) \underset{\hat{H}(\mathbf{y})=H_0}{\overset{\hat{H}(\mathbf{y})=H_1}{\gtrless}} 2\sigma^2 \ln\left(\eta\frac{|\sigma^2\mathbf{I} + \mathbf{\Lambda}_{\mathbf{x}}|^{1/2}}{\sigma^K}\right) \triangleq \gamma \tag{2.71}$$

where

$$\hat{\mathbf{x}}(\mathbf{y}) = \mathbf{\Lambda}_{\mathbf{x}}\left[\sigma^2\mathbf{I} + \mathbf{\Lambda}_{\mathbf{x}}\right]^{-1}\mathbf{y} = \begin{bmatrix}\hat{x}[0] & \hat{x}[1] & \cdots & \hat{x}[K-1]\end{bmatrix}^{\mathrm{T}}. \tag{2.72}$$

This detector has a special interpretation as our notation suggests. In particular, as will become apparent in Chapter 3, (2.72) is, in fact, what will be referred to as the Bayes least-squares estimate of $\mathbf{x}$ based on an observation of $\mathbf{y}$. Hence, in this example the sufficient statistic is obtained by correlating the received signal $y[n]$ with an estimate $\hat{x}[n]$ of the unknown signal $x[n]$. Note, however, that the threshold γ in (2.71) is different from that in the known signal case. Furthermore, it should be emphasized that sufficient statistics for more general detection problems involving unknown or partially known signals do not always have such interpretations. We explore such issues further in Chapter 6 in the context of joint detection and estimation.

Note that evaluation of the performance of the optimum decision rule is less straightforward than was the case in Example 2.6. As is apparent from (2.71) and (2.72), the sufficient statistic in this problem is no longer has a Gaussian distribution but rather a chi-squared distribution of degree K, i.e., $\ell \sim \chi^2_K$. As a result, we cannot get simple expressions for P_D and P_F in terms of the $\mathcal{Q}(\cdot)$ function. It is tempting to obtain approximations to P_D and P_F by directly exploiting a central limit theorem argument to approximate ℓ as Gaussian when K is reasonably large. However, a problem with this strategy is that the resulting Gaussian approximations are poorest in the tails of the distribution, yet this is typically the regime of interest for P_F and P_D calculations. However, useful bounds and approximations for these quantities are obtained via a technique referred to as the Chernoff bound. Such bounds have proven useful in a wide range of engineering problems over the last several decades.

2.5 TESTS WITH DISCRETE-VALUED OBSERVATIONS

Thus far we have focussed in this chapter on the case of observation vectors $\mathbf{y}$ that are specifically continuous-valued. However, there are many important decision and detection problems that involve inherently discrete-valued observations. Conveniently, the preceding theory carries over to the discrete case in a largely straightforward manner, with the likelihood ratio test continuing to play a central role, as we now develop. However, there are at least some important differences, which we will emphasize.

2.5.1 Bayesian Tests

With the notation (2.4), the optimum Bayesian decision rule takes a form analogous to that for the case of continuous-valued data; specifically,

$$L(\mathbf{y}) \triangleq \frac{p_{\mathbf{y}|H}[\mathbf{y}|H_1]}{p_{\mathbf{y}|H}[\mathbf{y}|H_0]} \underset{\hat{H}(\mathbf{y})=H_0}{\overset{\hat{H}(\mathbf{y})=H_1}{\gtrless}} \frac{P_0(C_{10}-C_{00})}{P_1(C_{01}-C_{11})} \triangleq \eta. \tag{2.73}$$

The derivation of this result closely mimics that for the case of continuous-valued data, the verification of which we leave as an exercise for the reader.

A couple of special issues arise in the case of likelihood ratio tests of the form (2.73). In particular, when the observations are discrete-valued, the likelihood function $L(\mathbf{y})$ is also discrete-valued; for future convenience, we denote these values by $0 \leq \eta_0 < \eta_1 < \eta_2 < \cdots$. This property has some significant consequences. First, it means that for many sets of costs and *a priori* probability assignments, the resulting η formed in (2.73) will often not coincide with one of the possible values of $L = L(\mathbf{y})$, i.e., we will often have $\eta \neq \eta_i$ for all i. In such cases, the case of equality in the likelihood ratio test will not arise, and the minimum Bayes risk is achieved by a likelihood ratio test that corresponds, as usual, to a unique (P_D, P_F) point on the associated operating characteristic.

However, for some choices of the costs and *a priori* probabilities, the resulting threshold η will satisfy $\eta = \eta_i$ for some particular i. This means that, unlike with continuous-valued data, in this case equality in the likelihood ratio test will occur with nonzero probability. Nevertheless, it can be readily verified that the Bayes risk is the same no matter how a decision is made in this event. Hence, when equality occurs, the decision can still be made arbitrarily, but these choices will correspond to different points on the operating characteristic.

2.5.2 The Operating Characteristic of the Likelihood Ratio Test, and Neyman-Pearson Tests

Let us more generally examine the form of the operating characteristic associated with the likelihood ratio test in the case of discrete-valued observations. In particular, let us begin by sweeping the threshold η in (2.73) from 0 to ∞ and examining the (P_D, P_F) values that are obtained. As our development in the last section revealed, in order to ensure that each threshold η maps to a unique (P_D, P_F), we need to choose an arbitrary but fixed convention for handling the case of equality in (2.73). For this purpose let us associate equality with the decision $\hat{H}(\mathbf{y}) = H_1$, expressing the likelihood ratio test in the form

$$L(\mathbf{y}) = \frac{p_{\mathbf{y}|H}[\mathbf{y}|H_1]}{p_{\mathbf{y}|H}[\mathbf{y}|H_0]} \underset{\hat{H}(\mathbf{y})=H_0}{\overset{\hat{H}(\mathbf{y})=H_1}{\gtrless}} \eta. \tag{2.74}$$

Before developing further results, let's explore a specific example to illustrate some of the key ideas.

Example 2.8

Suppose in an optical communication system bits are signaled by turning on and off a laser. At the detector, the measured photon arrival rate is used to determine whether the laser is on or off. When the laser is off (0-bit), the photons arrive according to a Poisson process with average arrival rate m_0; when the laser is on (1-bit), the rate is m_1 with $m_1 > m_0$. Suppose during a bit period we count the number

of photons y that arrive and use this observed data to make a decision. Then the likelihood functions for this decision problem are

$$p_{y|\mathsf{H}}[y|H_i] = \Pr[y = y \mid \mathsf{H} = H_i] = \begin{cases} \dfrac{m_i^y e^{-m_i}}{y!} & y = 0, 1, \ldots \\ 0 & \text{otherwise} \end{cases}.$$

In this example the likelihood ratio (2.74) leads to the test

$$L(y) = \left(\frac{m_1}{m_0}\right)^y e^{-(m_1 - m_0)} \underset{\hat{H}(y)=H_0}{\overset{\hat{H}(y)=H_1}{\gtrless}} \eta,$$

which further simplifies to

$$y \underset{\hat{H}(y)=H_0}{\overset{\hat{H}(y)=H_1}{\gtrless}} \frac{\ln\eta + (m_1 - m_0)}{\ln(m_1/m_0)} \triangleq \gamma. \tag{2.75}$$

The discrete nature of the this hypothesis testing problem means that the operating characteristic associated with the likelihood ratio test (2.75) is a discrete collection of points rather than a continuous curve of the type we encountered in an earlier example involving Gaussian data. Indeed, while the left-hand side of (2.75) is integer-valued, the right side is not in general. As a result, we have that P_D and P_F are given in terms of γ by the expressions

$$P_D = \Pr[y \geq \gamma \mid \mathsf{H} = H_1] = \Pr[y \geq \lceil\gamma\rceil \mid \mathsf{H} = H_1] = \sum_{y \geq \lceil\gamma\rceil} \frac{m_1^y e^{-m_1}}{y!} \tag{2.76a}$$

$$P_F = \Pr[y \geq \gamma \mid \mathsf{H} = H_0] = \Pr[y \geq \lceil\gamma\rceil \mid \mathsf{H} = H_0] = \sum_{y \geq \lceil\gamma\rceil} \frac{m_0^y e^{-m_0}}{y!}. \tag{2.76b}$$

The resulting operating characteristic is depicted in Fig. 2.6. In Figure 2.6, only the isolated (P_D, P_F) points indicated by circles are achievable by simple likelihood ratio tests. For example, the uppermost point is achieved for all $\gamma \leq 0$, the next highest point for all $0 < \gamma \leq 1$, the next for $1 < \gamma \leq 2$, and so on. This behavior is representative of such discrete decision problems.

As we discussed in Section 2.5.1, for Bayesian problems the specific cost and *a priori* probability assignments determine a threshold η in (2.74), which in turn corresponds to one of the isolated points comprising the operating characteristic.

For Neyman-Pearson problems with discrete observations, it is also straightforward to show that a likelihood ratio test of the form (2.74) is the optimum deterministic decision rule. Moreover, as we might expect, for this rule the threshold η is chosen so as to achieve the largest P_D subject to our contraint on the maximum allowable P_F (i.e., α). When α corresponds to at least one of the discrete points of the operating characteristic, then the corresponding P_D indicates the achievable detection probability. More typically, however, α will lie strictly between the P_F

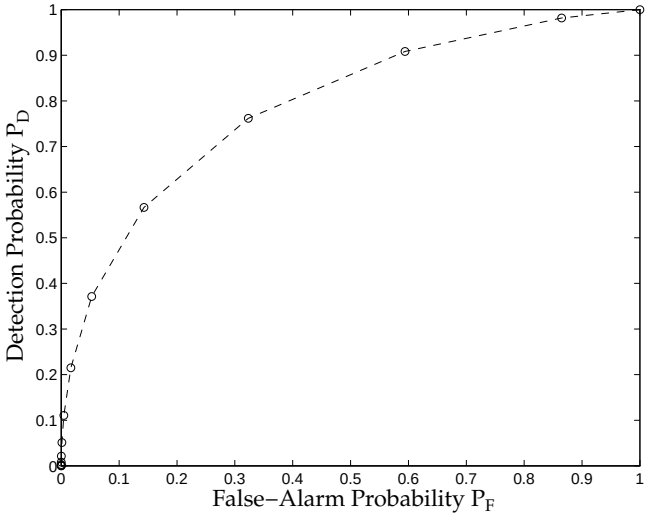


Figure 2.6. Operating characteristic associated with the likelihood ratio test for the Poisson detection problem. The photon arrival rates under H_0 and H_1 are $m_0 = 2$ and $m_1 = 4$, respectively. The circles mark points achievable by likelihood ratio tests.

values of points on the operating characteristic. In this case, the appropriate operating point corresponds to the (P_D, P_F) whose P_F is the largest P_F that is smaller than α. This operating point then uniquely specifies the decision rule.

While this decision rule is a straightforward extension of the result for continuous-valued observations, it is possible to develop a more sophisticated decision rule that typically yields at least somewhat better performance as measured by the Neyman-Pearson criterion. To see this, we first note that the straightforward likelihood ratio test defined above was obtained as the optimum decision rule among all possible *deterministic* tests. Specifically, the likelihood ratio test was the best decision rule of the form

$$\hat{H}(\mathbf{y}) = \begin{cases} H_0 & \mathbf{y} \in \mathcal{Z}_0 \\ H_1 & \mathbf{y} \in \mathcal{Z}_1 \end{cases}$$

where $\mathcal{Z}_0$ and $\mathcal{Z}_1$ are mutually exclusive, collectively exhaustive sets in the observation space $\{\mathbf{y}\}$. For these rules, each observation $\mathbf{y}$ maps to a unique decision $\hat{H}(\mathbf{y})$.

When the regular likelihood ratio test cannot meet the false-alarm constraint with equality, better performance can be achieved by employing a decision rule chosen from outside the class of deterministic rules. In particular, if we allow for some randomness in the decision process, so that a particular observation does not always produce the same decision, we can obtain improved detection performance while meeting our false-alarm constraint. This can be accomplished as follows. Consider the sequence of thresholds $\eta_0, \eta_1, \ldots$ that correspond to values that the likelihood ratio can take on, and denote the corresponding operating points by $(P_D(\eta_i), P_F(\eta_i))$ for $i = 0, 1, \ldots$. Determine $\hat{\imath}$ such that $\eta_{\hat{\imath}}$ is the threshold value that results in the likelihood ratio test with the smallest false-alarm probability that is greater than α. Then as illustrated in Fig. 2.7, $P_F(\eta_{\hat{\imath}})$ and $P_F(\eta_{\hat{\imath}+1})$ "bracket" α: using $\eta_{\hat{\imath}+1}$ results in test with the largest false-alarm probability that is less than α.

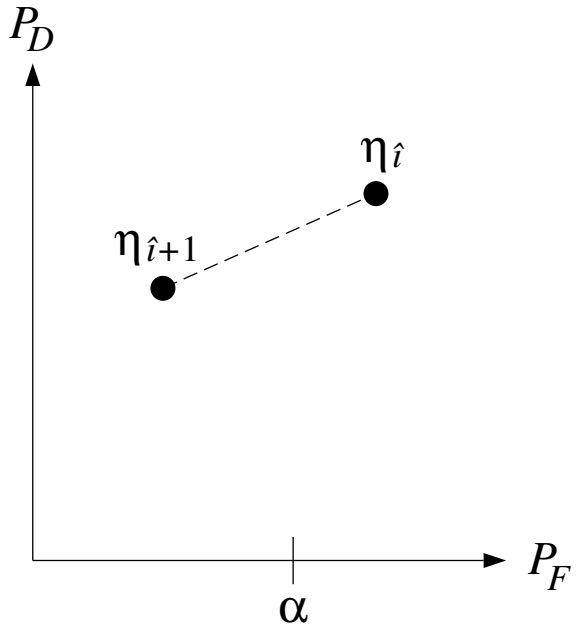


Figure 2.7. Achievable operating points for a randomization between two likelihood ratio tests, one with threshold $\eta_{\hat{\imath}}$ and the other with threshold $\eta_{\hat{\imath}+1}$.

Now consider the following randomized decision rule. We flip a biased coin for which the probability of "heads" is p and that of "tails" is $1-p$. If "heads" turns up, we use the likelihood ratio test with threshold $\eta_{\hat{\imath}}$; if "tails" turns up, we use the likelihood ratio test with threshold $\eta_{\hat{\imath}+1}$. Then the resulting test achieves

$$\begin{aligned} P_D &= pP_D(\eta_{\hat{\imath}}) + (1-p)P_D(\eta_{\hat{\imath}+1}) \\ P_F &= pP_F(\eta_{\hat{\imath}}) + (1-p)P_F(\eta_{\hat{\imath}+1}), \end{aligned} \tag{2.77}$$

and corresponds to a point on the line segment connecting $(P_D(\eta_{\hat{\imath}}), P_F(\eta_{\hat{\imath}}))$ and $(P_D(\eta_{\hat{\imath}+1}), P_F(\eta_{\hat{\imath}+1}))$. This line segment is indicated with the dashed line segment in Fig. 2.7. In particular, as p is varied from 0 to 1, the operating point moves from $(P_D(\eta_{\hat{\imath}+1}), P_F(\eta_{\hat{\imath}+1}))$ to $(P_D(\eta_{\hat{\imath}}), P_F(\eta_{\hat{\imath}}))$. Hence, by choosing p appropriately, we can achieve the false-alarm probability constraint with equality, i.e., there exists a $\hat{p}$ such that

$$P_F = \hat{p}P_F(\eta_{\hat{\imath}}) + (1-\hat{p})P_F(\eta_{\hat{\imath}+1}) = \alpha.$$

Since the operating characteristic is a monotonically increasing function, it follows that $P_D(\eta_{\hat{\imath}}) > P_D(\eta_{\hat{\imath}+1})$, which in turn implies that $\hat{p}$ is the value of p yielding the largest possible P_D.

Our randomization argument allows us to conclude that all points on the operating characteristic as well as all points on the line segments connecting operating characteristic points (adjacent or not) can be achieved via randomized likelihood ratio tests. Referring to Example 2.8 in particular, this means that all operating points on the piecewise linear dashed line in Fig. 2.6 can be achieved, allowing any false alarm probability constraint α to be met with equality.

While we have developed randomized likelihood ratio tests as a random choice among a pair of likelihood ratio tests, they can be implemented in other ways that can sometimes be more convenient. To see this, note that for the randomized rule above, both likelihood ratio tests lead to the decision H_1 when $L(\mathbf{y}) \geq \eta_{i+1}$. Hence, regardless of the outcome of the coin flip, the decision will be H_1 for the associated values of $\mathbf{y}$. Similarly, if $L(\mathbf{y}) \leq \eta_{i-1}$ then both tests lead to the decision H_0, so the randomized rule will decide H_0 regardless of the outcome of the coin flip. Hence, only when $L(\mathbf{y}) = \eta_i$ does the randomization play a role. In this case, if "heads" comes up, the decision will be H_1, while if "tails" comes up, the decision will be H_0. We can summarize this implementation of the test in the following form:

$$\Pr\left[\hat{H}(\mathbf{y}) = H_1 \,\middle|\, \mathbf{y} = \mathbf{y}\right] = \begin{cases} 1 & L(\mathbf{y}) \geq \eta_{i+1} \\ \hat{p} & L(\mathbf{y}) = \eta_i \\ 0 & L(\mathbf{y}) \leq \eta_{i-1} \end{cases}. \tag{2.78}$$

Our results in this section have suggested that at least in some problems involving discrete-valued data and the Neyman-Pearson criterion a randomized test can lead to better performance than a deterministic test. This observation, in turn, raises several natural and important questions. For example, we considered a very particular class of randomized tests—specifically, a simple random choice between the outcomes of two (deterministic) likelihood ratio tests. Would some other type of randomized test be able to perform better still? And could some more general form of randomized test be able to improve performance in the case of continuous-valued data with Bayesian or Neyman-Pearson criteria? To answer these questions, in the next section we develop decision rules optimized over of a broad class of randomized tests.

2.6 RANDOMIZED TESTS

For a randomized test, the decision rule is a *random function* of the data, which we denote using $\hat{H}(\cdot)$. Hence, even for a deterministic argument $\mathbf{y}$, the decision $\hat{H}(\mathbf{y})$ is a random quantity. However, $\hat{H}(\mathbf{y})$ has the property that conditioned on knowledge of $\mathbf{y}$, the function is independent of the hypothesis H. Such a test is fully described by the probabilities

$$\begin{aligned} Q_0(\mathbf{y}) &= \Pr\left[\hat{H}(\mathbf{y}) = H_0 \,\middle|\, \mathbf{y} = \mathbf{y}\right] = \Pr\left[\hat{H}(\mathbf{y}) = H_0 \,\middle|\, \mathbf{y} = \mathbf{y}, H = H_i\right] \\ Q_1(\mathbf{y}) &= \Pr\left[\hat{H}(\mathbf{y}) = H_1 \,\middle|\, \mathbf{y} = \mathbf{y}\right] = \Pr\left[\hat{H}(\mathbf{y}) = H_1 \,\middle|\, \mathbf{y} = \mathbf{y}, H = H_i\right], \end{aligned} \tag{2.79}$$

where, of course, $Q_0(\mathbf{y}) + Q_1(\mathbf{y}) = 1$.

With this notation, we see that deterministic rules are a special case, corresponding to

$$Q_1(\mathbf{y}) = \begin{cases} 1 & \mathbf{y} \in \mathcal{Z}_1 \\ 0 & \mathbf{y} \in \mathcal{Z}_0 \end{cases}. \tag{2.80}$$

Moreover, it also follows immediately that tests formed by a random choice among two likelihood ratio tests, such as were considered in Section 2.5.2 are also special cases. For example, the test described via (2.78) corresponds to

$$Q_1(\mathbf{y}) = \begin{cases} 1 & L(\mathbf{y}) \geq \eta_{i+1} \\ \hat{p} & L(\mathbf{y}) = \eta_i \\ 0 & L(\mathbf{y}) \leq \eta_{i-1}. \end{cases} \tag{2.81}$$

More generally, for a randomized test that corresponds to the random choice between two likelihood ratio tests with respective thresholds η_1 and η_2 such that $\eta_2 > \eta_1$, and where the probability of selecting the first test is p, it is straightforward to verify that

$$Q_1(\mathbf{y}) = \begin{cases} 1 & L(\mathbf{y}) \geq \eta_2 \\ p & \eta_1 \leq L(\mathbf{y}) < \eta_2 \\ 0 & L(\mathbf{y}) < \eta_1. \end{cases} \tag{2.82}$$

From our general characterization (2.79), we see that specifying a randomized decision rule is equivalent to specifying $Q_0(\cdot)$—or $Q_1(\cdot)$. Hence, determining the optimum randomized test for a given performance criterion involves solving for the optimum mapping $Q_0(\cdot)$. Using this approach, we develop the optimum *randomized* tests for Bayesian and Neyman-Pearson hypothesis testing problems in the next two sections, respectively. As we will see this allows us to draw some important conclusions about when randomized tests are—and are not—needed.

2.6.1 Bayesian Case

We begin by establishing a more general version of our Bayesian result for the case of randomized tests. We consider the case of continuous-valued data; the derivation in the discrete case is analogous. We begin by writing our Bayes risk in the form

$$J(Q_0) = \int \tilde{J}(\mathbf{y})\, p_{\mathbf{y}}(\mathbf{y})\, d\mathbf{y},$$

where

$$\tilde{J}(\mathbf{y}) = E\left[C(H, \hat{H}(\mathbf{y})) \,\middle|\, \mathbf{y} = \mathbf{y}\right].$$

Again we see it suffices to minimize $\tilde{J}(\mathbf{y})$ for each $\mathbf{y}$. Applying, in turn, (2.79) and Bayes' Rule, we can write $\tilde{J}(\mathbf{y})$ in the form

$$\begin{aligned}\tilde{J}(\mathbf{y}) &= \sum_{i,j} C_{ij} \Pr\left[H = H_j, \hat{H}(\mathbf{y}) = H_i \;\middle|\; \mathbf{y} = \mathbf{y}\right] \\ &= \sum_{i,j} C_{ij} \Pr\left[H = H_j \mid \mathbf{y} = \mathbf{y}\right] \Pr\left[\hat{H} = H_i \;\middle|\; \mathbf{y} = \mathbf{y}\right] \\ &= \sum_{i,j} C_{ij}\, Q_i(\mathbf{y})\, \frac{P_j\, p_{\mathbf{y}|H}(\mathbf{y}|H_j)}{p_{\mathbf{y}}(\mathbf{y})}, \end{aligned} \tag{2.83}$$

from which, with

$$\Delta(\mathbf{y}) \triangleq C_{10} \frac{P_0 p_{\mathbf{y}|H}(\mathbf{y}|H_0)}{p_{\mathbf{y}}(\mathbf{y})} + C_{11} \frac{P_1 p_{\mathbf{y}|H}(\mathbf{y}|H_1)}{p_{\mathbf{y}}(\mathbf{y})}, \tag{2.84}$$

we obtain

$$\begin{aligned}\tilde{J}(\mathbf{y}) &= \Delta(\mathbf{y}) + Q_0(\mathbf{y}) \frac{p_{\mathbf{y}|H}(\mathbf{y}|H_0)}{p_{\mathbf{y}}(\mathbf{y})} P_1(C_{01} - C_{11}) \left[\frac{p_{\mathbf{y}|H}(\mathbf{y}|H_1)}{p_{\mathbf{y}|H}(\mathbf{y}|H_0)} - \frac{P_0(C_{10} - C_{00})}{P_1(C_{01} - C_{11})} - \right] \\ &= \Delta(\mathbf{y}) + Q_0(\mathbf{y}) \frac{p_{\mathbf{y}|H}(\mathbf{y}|H_0)}{p_{\mathbf{y}}(\mathbf{y})} P_1(C_{01} - C_{11}) \left[L(\mathbf{y}) - \eta\right] \end{aligned} \tag{2.85}$$

using the notation defined in (2.19). From (2.85) we can immediately conclude that $\tilde{J}(\mathbf{y})$ is minimized over $0 \leq Q_0(\mathbf{y}) \leq 1$ by choosing $Q_0(\mathbf{y}) = 0$ when the term in brackets is positive and $Q_0(\mathbf{y}) = 1$ when the term in brackets is negative. But this is precisely the (deterministic) likelihood ratio test (2.19) we developed earlier! Hence, for Bayesian problems, we can conclude that a deterministic test will always suffice. Furthermore, although our derivation has been for the case of continuous-valued data, the same conclusion is reached in the discrete case.

2.6.2 Neyman-Pearson Case

We next establish a more general version of our Neyman-Pearson result for the case of randomized tests. We begin with the case of continuous-valued data. As in the deterministic case, we follow a Lagrange multiplier approach, expressing our objective function as

$$\begin{aligned} J(Q_0) &= 1 - P_D + \lambda(P_F - \alpha') \\ &= \lambda(1 - \alpha') + \Pr\left[\hat{H}(\mathbf{y}) = H_0 \;\middle|\; H = H_1\right] - \lambda \Pr\left[\hat{H}(\mathbf{y}) = H_0 \;\middle|\; H = H_0\right] \end{aligned} \tag{2.86}$$

for some $\alpha' \leq \alpha$. This time, though, (2.86) expands as

$$J(Q_0) = \lambda(1 - \alpha') + \int \Pr\left[\hat{H}(\mathbf{y}) = H_0 \;\middle|\; \mathbf{y} = \mathbf{y}\right] \left[p_{\mathbf{y}|H}(\mathbf{y}|H_1) - \lambda p_{\mathbf{y}|H}(\mathbf{y}|H_0)\right] d\mathbf{y}$$

from which we obtain, using the definition of $L(\mathbf{y})$,

$$\begin{aligned} J(Q_0) &= \lambda(1-\alpha') + \int Q_0(\mathbf{y}) \left[p_{\mathbf{y}|H}(\mathbf{y}|H_1) - \lambda p_{\mathbf{y}|H}(\mathbf{y}|H_0) \right] d\mathbf{y} \\ &= \lambda(1-\alpha') + \int Q_0(\mathbf{y}) \left[L(\mathbf{y}) - \lambda \right] p_{\mathbf{y}|H}(\mathbf{y}|H_0)\, d\mathbf{y}. \end{aligned} \tag{2.87}$$

Since $0 \le Q_0(\mathbf{y}) \le 1$, the objective function (2.87) is minimized by setting $Q_0(\mathbf{y}) = 0$ for all values of $\mathbf{y}$ such that the term in braces is positive, and $Q_0(\mathbf{y}) = 1$ for all values of $\mathbf{y}$ such that the term in braces is negative; i.e.,

$$Q_0(\mathbf{y}) = \begin{cases} 0 & L(\mathbf{y}) > \lambda \\ 1 & L(\mathbf{y}) < \lambda. \end{cases} \tag{2.88}$$

Hence, we can conclude that our optimum rule is almost a simple (deterministic) likelihood ratio test. In particular, at least except when $L(\mathbf{y}) = \lambda$ we have that

$$\hat{H}(\mathbf{y}) = \begin{cases} H_1 & L(\mathbf{y}) > \lambda \\ H_0 & L(\mathbf{y}) < \lambda. \end{cases} \tag{2.89}$$

It remains only to determine what the nature of the decision (i.e., $Q_0(\mathbf{y})$) when $L(\mathbf{y}) = \lambda$ and the choice of λ. These quantities are determined by meeting the constraint $P_F = \alpha'$. Specifically,

$$\begin{aligned} P_F &= \int \left[1 - Q_0(\mathbf{y}) \right] p_{\mathbf{y}|H}(\mathbf{y}|H_0)\, d\mathbf{y} \\ &= \int_{\{\mathbf{y}|L(\mathbf{y})>\lambda\}} p_{\mathbf{y}|H}(\mathbf{y}|H_0)\, d\mathbf{y} + \int_{\{\mathbf{y}|L(\mathbf{y})=\lambda\}} \left[1 - Q_0(\mathbf{y}) \right] p_{\mathbf{y}|H}(\mathbf{y}|H_0)\, d\mathbf{y} \\ &= \Pr\left[L(\mathbf{y}) > \lambda \mid H = H_0 \right] + \int_{\{\mathbf{y}|L(\mathbf{y})=\lambda\}} \left[1 - Q_0(\mathbf{y}) \right] p_{\mathbf{y}|H}(\mathbf{y}|H_0)\, d\mathbf{y}, \end{aligned} \tag{2.90}$$

where the second equality follows from substituting for $Q_0(\mathbf{y})$ using (2.88). Observe that $P_F = 1$ if $\lambda < 0$, so it suffices to restrict our attention to $\lambda \ge 0$. Furthermore, note that the first term in (2.90) is a nonincreasing function of λ.

Now since $\mathbf{y}$ is a continuous-valued random variable, then the second term in (2.90) is zero and the remaining term, which is a continuous of λ, can be chosen so that it equals α'. In this case, it does not matter how we choose $Q_0(\mathbf{y})$ when $L(\mathbf{y}) = \lambda$, so the optimum randomized rule degenerates to a deterministic likelihood ratio test again. It remains only to show that for optimum P_D we want $\alpha' = \alpha$. Given our preceding results, it suffices to exploit the fact that for likelihood ratio tests P_D is a monotonically nondecreasing function of P_F.

Discrete Case

When the data is discrete, some important differences arise, which we now explore. Proceeding as we did at the outset of Section 2.6.2, we obtain

$$\begin{aligned} J(Q_0) &= \lambda(1-\alpha') + \sum_{\mathbf{y}} Q_0(\mathbf{y}) \left[p_{\mathbf{y}|H}[\mathbf{y}|H_1] - \lambda p_{\mathbf{y}|H}[\mathbf{y}|H_0] \right] \\ &= \lambda(1-\alpha') + \sum_{\mathbf{y}} Q_0(\mathbf{y}) \left[L(\mathbf{y}) - \lambda \right] p_{\mathbf{y}|H}[\mathbf{y}|H_0]. \end{aligned} \tag{2.91}$$

from which we analogously conclude that for $L(\mathbf{y}) \neq \lambda$ we have (2.88). Hence, it remains only to determine λ and the decision when $L(\mathbf{y}) = \lambda$ from the false alarm constraint.

An important difference from the continuous case is that when $\mathbf{y}$ is discrete-valued, so that $L(\mathbf{y})$ is also discrete-valued, the first term in (2.90) is not a continuous function of λ, but is piecewise constant.

As before, let us denote the values that $L(\mathbf{y})$ takes on by $\eta_0, \eta_1, \eta_2, \ldots$, where $0 < \eta_0 < \eta_1 < \eta_2 < \cdots$. And let us choose $\hat{\imath}$ so that $\lambda = \eta_{\hat{\imath}}$ is the smallest threshold such that

$$\Pr\left[L(\mathbf{y}) > \lambda \mid \mathsf{H} = H_0\right] = \Pr\left[L(\mathbf{y}) \geq \eta_{\hat{\imath}+1} \mid \mathsf{H} = H_0\right] \leq \alpha'.$$

Then we can obtain $P_F = \alpha'$ by choosing

$$Q_1(\mathbf{y}) = 1 - Q_0(\mathbf{y})$$

appropriately for all $\mathbf{y}$ such that $L(\mathbf{y}) = \eta_{\hat{\imath}}$. In particular, $0 < Q_1(\cdot) < 1$ in this range must be chosen so that

$$\begin{aligned} \alpha' - \Pr\left[L(\mathbf{y}) \geq \eta_{\hat{\imath}+1} \mid \mathsf{H} = H_0\right] &= \sum_{\{\mathbf{y} \mid L(\mathbf{y}) = \eta_{\hat{\imath}}\}} Q_1(\mathbf{y})\, p_{\mathbf{y}|H}[\mathbf{y}|H_0] \\ &= q \, \Pr\left[L(\mathbf{y}) = \eta_{\hat{\imath}} \mid \mathsf{H} = H_0\right] \end{aligned} \tag{2.92}$$

where

$$q \triangleq E\left[Q_1(\mathbf{y}) \mid L(\mathbf{y}) = \eta_{\hat{\imath}}\right]. \tag{2.93}$$

As we would expect, our decision probabilities $Q_1(\cdot)$ when $L(\mathbf{y}) = \eta_{\hat{\imath}}$ appear in (2.92) only through q, so it suffices to appropriately select the latter.

For $q = 0$, the resulting decision rule is the deterministic test

$$L(\mathbf{y}) \underset{\hat{H}(\mathbf{y})=H_0}{\overset{\hat{H}(\mathbf{y})=H_1}{\gtrless}} \eta_{\hat{\imath}+1}. \tag{2.94}$$

More generally, when $q > 0$, we have that the decision rule involves a random choice between the decision rule (2.94) and

$$L(\mathbf{y}) \underset{\hat{H}(\mathbf{y})=H_0}{\overset{\hat{H}(\mathbf{y})=H_1}{\gtrless}} \eta_{\hat{\imath}}. \tag{2.95}$$

In particular, the test (2.95) is chosen with probability q, and the test (2.94) with probability $1-q$. In terms of the operating characteristic of the likelihood ratio test, this is a point on the line segment connecting the points corresponding to the two deterministic tests (2.94) and (2.95). This, of course, is precisely the form of the heuristically designed randomized test we explored at the end of Section 2.5.2, which we now see is optimal. As a result, when we include randomized tests, it makes sense to *redefine* the operating characteristic for the discrete case as the isolate likelihood ratio test operating points together with the line segments that connect these discrete points.

It remains only to verify that for optimum P_D we want $\alpha' = \alpha$ in the discrete case as well. However, for likelihood ratio tests involving discrete data, the P_D values form a nondecreasing sequence as a function of P_F. Then since the performance of randomized tests corresponds to points on the line segments connecting the performance points associated with deterministic tests, P_D is a nondecreasing function of P_F for our more general class of randomized tests as well.

In summary, optimum Neyman-Pearson decision rules always take the form of either a deterministic rule in the form of a likelihood ratio test or a randomized rule in the form of a simple randomization between two likelihood ratio tests. In both cases, the likelihood ratio test and its associated operating characteristic play a central role. Accordingly, we explore their properties further.

2.7 PROPERTIES OF THE LIKELIHOOD RATIO TEST OPERATING CHARACTERISTIC

By exploiting the special role that the likelihood ratio test plays in both deterministic and randomized optimum decisions rules, we can develop a number of key properties of the P_D–P_F operating characteristic associated with the likelihood ratio test. For future reference, recall that for continuous-valued data the test takes the form

$$L(\mathbf{y}) = \frac{p_{\mathbf{y}|H}(\mathbf{y}|H_1)}{p_{\mathbf{y}|H}(\mathbf{y}|H_0)} \underset{\hat{H}(\mathbf{y})=H_0}{\overset{\hat{H}(\mathbf{y})=H_1}{\gtrless}} \eta, \tag{2.96}$$

while for discrete-valued data the form is

$$L(\mathbf{y}) = \frac{p_{\mathbf{y}|H}[\mathbf{y}|H_1]}{p_{\mathbf{y}|H}[\mathbf{y}|H_0]} \underset{\hat{H}(\mathbf{y})=H_0}{\overset{\hat{H}(\mathbf{y})=H_1}{\gtrless}} \eta. \tag{2.97}$$

We emphasize at the outset that the detailed shape of the operating characteristic is determined by the measurement model for the data—for example, by $p_{\mathbf{y}|H}(\mathbf{y}|H_0)$ and $p_{\mathbf{y}|H}(\mathbf{y}|H_1)$ in the continuous-case—since it is this information that is used to construct the likelihood ratio $L(\mathbf{y})$. However, all operating characteristics share some important characteristics in common, and it is these that we explore in this section. As a simple example, which was mentioned earlier, we have

that the (P_D, P_F) points $(0, 0)$ and $(1, 1)$ always lie on the operating characteristic, and correspond to $\eta \to \infty$ and $\eta \to 0$, respectively.

It is also straightforward to verify that $P_D \geq P_F$, i.e., that the operating characteristic always lies above the diagonal in the P_D–P_F plane. This can be verified using a randomization argument. In particular, suppose our decision rule ignores the data $\mathbf{y}$ and bases its decision solely on the outcome of a biased coin flip, where the probability of "heads" is p. If the coin comes up "heads" we make the decision $\hat{H}(\mathbf{y}) = H_1$, while if it comes up "tails" we make the decision $\hat{H}(\mathbf{y}) = H_0$. Then for this rule we have

$$P_D = \Pr\left[\hat{H}(\mathbf{y}) = H_1 \,\middle|\, \mathsf{H} = H_1\right] = \Pr\left[\hat{H}(\mathbf{y}) = H_1\right] = p$$
$$P_F = \Pr\left[\hat{H}(\mathbf{y}) = H_1 \,\middle|\, \mathsf{H} = H_0\right] = \Pr\left[\hat{H}(\mathbf{y}) = H_1\right] = p,$$

which corresponds to a point on the diagonal in the P_D–P_F plane. Hence, we can achieve all points on the diagonal by varying the bias p in the coin. However, for a given P_F this decision rule cannot provide a better P_D than the corresponding point on the operating characteristic; otherwise, this would contradict our Neyman-Pearson result that this operating characteristic defines the best achievable P_D for a given P_F.[8] Hence, we conclude $P_D \geq P_F$.

Randomization arguments play an important role in establishing other properties of likelihood ratio test operating characteristics as well. For example, we can also use such an argument to establish that the operating characteristic is a concave function. To see this, let $(P_D(\eta_1), P_F(\eta_1))$ and $(P_D(\eta_2), P_F(\eta_2))$ be points on the operating characteristic corresponding to two arbitrary thresholds η_1 and η_2, respectively. Then the operating characteristic is concave if the points on the straight line segment joining these two points invariably lie below (or on) the operating characteristic. However, the points on this straight line segment can be parameterized according to

$$(P_D, P_F) = (pP_D(\eta_1) + (1-p)P_D(\eta_2), pP_F(\eta_1) + (1-p)P_F(\eta_2)), \tag{2.98}$$

where $0 \leq p \leq 1$ is the parameter. Moreover, the points (2.98) can be achieved via the following randomized test: a biased coin is flipped, and if it turns up "heads," the likelihood ratio test with threshold η_1 is used; otherwise, the likelihood ratio test with threshold η_2 is used. Again by varying the bias p in our coin, we can achieve all the points on the line segment. Hence, from our Neyman-Pearson result, no part of the likelihood ratio test operating characteristic between $P_F(\eta_1)$ and $P_F(\eta_2)$ can lie below this line segment. Thus, since the endpoints were arbitrary, we conclude that the operating characteristic is concave.

Let's consider one final property, which applies to likelihood ratio test operating characteristics associated with continuous-valued data. In particular, we

[8] Indeed, we wouldn't expect to obtain better performance by ignoring the data $\mathbf{y}$!

show that at those points where it is defined, the slope of the operating characteristic is numerically equal to the corresponding threshold η, i.e.,

$$\frac{dP_D}{dP_F} = \eta. \tag{2.99}$$

Note that since $\eta \geq 0$, this is another way of verifying that the operating characteristic is nondecreasing. A proof is as follows. With

$$\mathcal{Z}_1(\eta) = \{\mathbf{y} \mid L(\mathbf{y}) > \eta\} \tag{2.100}$$

we have

$$P_D(\eta) = \int_{\mathcal{Z}_1(\eta)} p_{\mathbf{y}|H}(\mathbf{y}|H_1)\, d\mathbf{y},$$

which after applying the definition of the likelihood function (2.96) yields

$$P_D(\eta) = \int_{\mathcal{Z}_1(\eta)} L(\mathbf{y})\, p_{\mathbf{y}|H}(\mathbf{y}|H_0)\, d\mathbf{y}. \tag{2.101}$$

Next, note that with $u(\cdot)$ denoting the unit step function, i.e.,

$$u(x) = \begin{cases} 1 & x > 0 \\ 0 & \text{otherwise} \end{cases},$$

we have

$$u(L(\mathbf{y}) - \eta) = \begin{cases} 1 & \mathbf{y} \in \mathcal{Z}_1(\eta) \\ 0 & \text{otherwise} \end{cases}. \tag{2.102}$$

In turn, using the result (2.102) in (2.101) we obtain

$$P_D(\eta) = E\left[u(L(\mathbf{y}) - \eta)\, L(\mathbf{y}) \mid H = H_0\right] = E\left[u(L - \eta)\, L \mid H = H_0\right] = \int_\eta^\infty L\, p_{L|H}(L|H_0)\, dL \tag{2.103}$$

Finally, differentiating (2.103) with respect to η then yields

$$\frac{dP_D}{d\eta} = -\eta\, p_{L|H}(\eta|H_0). \tag{2.104}$$

However, since

$$P_F = \int_\eta^\infty p_{L|H}(L|H_0)\, dL$$

we know

$$\frac{dP_F}{d\eta} = -p_{L|H}(\eta|H_0). \tag{2.105}$$

Hence, dividing (2.104) by (2.105) we obtain (2.99).

2.7.1 Achievable Operating Points

Having determined some of the structure of the operating characteristic, let us next explore more generally how to use this characteristic to define what operating points can be achieved by any test—deterministic or randomized—in the P_D–P_F plane. First, we note that every point between the operating characteristic and the diagonal ($P_D = P_F$) can be achieved by a simple randomized test. To see this it suffices to recognize that every point in this region lies on some line connecting two points on the operating characteristic. Hence, every such point can be achieved by a simple randomization between two likelihood ratio tests having the corresponding thresholds, using a coin with suitable bias p.

The test that achieves a given operating point in this region need not be unique, however. To illustrate this, let η_0 be the threshold corresponding to a particular point on the operating characteristic. Then if for our decision rule we use a random choice (using a coin with bias p) between the outcome of this likelihood ratio test and that of the "guessing rule" that achieves an arbitrary point on the diagonal (using a different coin with bias q). Hence, by choosing $\eta_0 > 0$, $0 \leq p \leq 1$, and $0 \leq q \leq 1$, appropriately, our "doubly-randomized" test can also achieve any desired point in the region of interest.

Let us next consider which points below the diagonal are achievable. This can be addressed via a simple "rule-reversal" argument. For this we require the following notion of a reversed test. If $\hat{H}(\cdot)$ describes a deterministic decision rule, then the corresponding reversed rule, which we denote using $\overline{\hat{H}}(\cdot)$, is simply one whose decisions are made as follows:

$$\overline{\hat{H}}(\mathbf{y}) = \begin{cases} H_0 & \hat{H}(\mathbf{y}) = H_1 \\ H_1 & \hat{H}(\mathbf{y}) = H_0 \end{cases}.$$

More generally, if $Q_0(\cdot)$ describes a randomized decision rule, then the corresponding decision rule, which we denote using $\overline{Q_0}(\cdot)$, is defined via

$$\overline{Q_0}(\mathbf{y}) = Q_1(\mathbf{y}) = 1 - Q_0(\mathbf{y}).$$

If a deterministic or randomized test achieves the operating point $(P_D, P_F) = (\beta, \alpha)$, then it is easy to verify that the corresponding reversed test achieves the operating point $(P_D, P_F) = (1-\beta, 1-\alpha)$, i.e.,

$$\Pr\left[\overline{\hat{H}}(\mathbf{y}) = H_1 \,\middle|\, H = H_1\right] = \Pr\left[\hat{H}(\mathbf{y}) = H_0 \,\middle|\, H = H_1\right] = 1 - \beta$$
$$\Pr\left[\overline{\hat{H}}(\mathbf{y}) = H_1 \,\middle|\, H = H_0\right] = \Pr\left[\hat{H}(\mathbf{y}) = H_0 \,\middle|\, H = H_0\right] = 1 - \alpha.$$

Using this property, it follows that those points lying on the curve corresponding to the operating characteristic reflected across the $P_D = 1/2$ and $P_F = 1/2$ lines are achievable by likelihood ratio tests whose decisions are reversed. In turn, all points between the diagonal and this "reflected operating characteristic" are

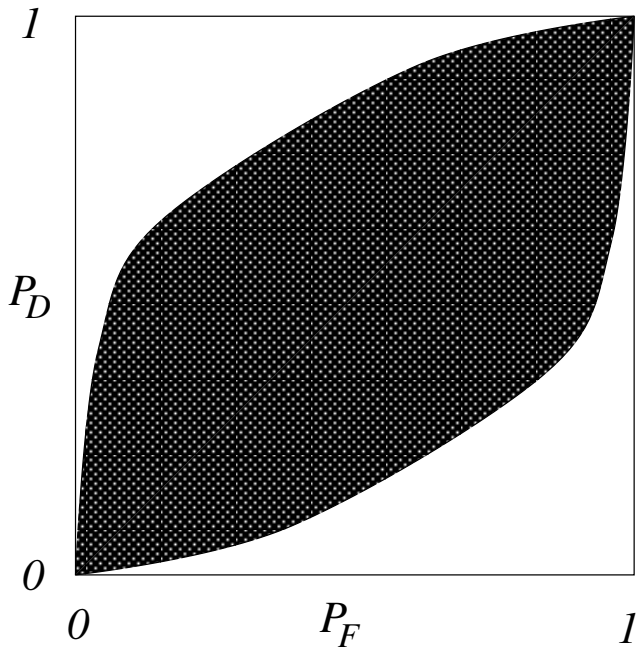


Figure 2.8. Achievable region in the P_D–P_F plane.

achievable using a suitably designed randomized test. Hence, as illustrated in Fig. 2.8 we can conclude that all points within the region bounded by the operating characteristic and reversed operating characteristic curves can be achieved using a suitably designed decision rule.

We can also establish the converse result: that no decision rule—deterministic or randomized—can achieve points outside this region. That no (P_D, P_F) point above the operating characteristic can be achieved follows from our Neyman-Pearson results. That no point below the reflected operating characteristic can be achieved (including $(P_D, P_F) = (0, 1)$!) follows as well, using a "proof-by-contradiction" argument. In particular, if such a point could be achieved, then so could its reflection via a reversed test. However, this reflection would then lie above the operating characteristic, which would contradict the Neyman-Pearson optimality of the likelihood ratio test.

2.8 M-ARY HYPOTHESIS TESTING

Thus far we have focussed on the case of binary hypothesis testing in this chapter. From this investigation, we have developed important insights that apply to decision problems involving multiple hypotheses more generally. However, some special considerations and issues arise in the more general M-ary hypothesis testing problem. We explore a few of these issues in this section, but emphasize that our treatment is an especially introductory one.

To simplify our discussion, we restrict our attention to the Bayesian problem formulation with continuous-valued data and deterministic decision rules. Accordingly, the scenario we consider is as follows:

1. There are M hypotheses $H_0, H_1, ..., H_{M-1}$ with associated *a priori* probabilities $P_i = \Pr[\mathsf{H} = H_i]$.
2. We have a set of costs of the form C_{ij}, which corresponds to the cost of deciding $\hat{H} = H_i$ when $\mathsf{H} = H_j$ is true.
3. Our (generally vector) measurements $\mathbf{y}$ are characterized by the set of densities $p_{\mathbf{y}|H}(\mathbf{y}|H_i)$.

For this problem, we explore a procedure for choosing one of the M hypotheses based on the observed data $\mathbf{y}$ so as to minimize the associated expected cost. This corresponds to designing an M-valued decision rule $\hat{H}(\cdot) : \mathbb{R}^K \rightarrow \{H_0, H_1, \ldots, H_{M-1}\}$. By analogy to the binary case, we begin by noting that if $\mathbf{y} = \mathbf{y}$ and $\hat{H}(\mathbf{y}) = H_{\hat{m}}$, then the expected cost is

$$\tilde{J}(H_{\hat{m}}, \mathbf{y}) = \sum_{m=0}^{M-1} C_{\hat{m}m} \Pr[\mathsf{H} = H_m \mid \mathbf{y} = \mathbf{y}]. \tag{2.106}$$

Consequently, given the observation $\mathbf{y} = \mathbf{y}$ we want to choose $\hat{m}$ to minimize (2.106), i.e.,

$$\hat{H}(\mathbf{y}) = H_{\hat{m}} \quad \text{where} \quad \hat{m} = \underset{m \in \{0,1,\ldots,M-1\}}{\arg\min} \sum_{j=0}^{M-1} C_{mj} \Pr[\mathsf{H} = H_j \mid \mathbf{y} = \mathbf{y}]. \tag{2.107}$$

Let's explore several aspects of this decision rule. We begin by considering a special case.

2.8.1 Special Case: Minimum Probability-of-Error Decisions

If all possible errors are penalized equally, i.e.,

$$C_{ij} = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases}, \tag{2.108}$$

then (2.107) becomes

$$\hat{H}(\mathbf{y}) = H_{\hat{m}} \quad \text{where} \quad \hat{m} = \underset{m \in \{0,1,\ldots,M-1\}}{\arg\min} \sum_{j \neq m} \Pr[\mathsf{H} = H_j \mid \mathbf{y} = \mathbf{y}], \tag{2.109}$$

Exploiting the fact that

$$\sum_{j \neq m} \Pr[\mathsf{H} = H_j \mid \mathbf{y} = \mathbf{y}] = 1 - \Pr[\mathsf{H} = H_m \mid \mathbf{y} = \mathbf{y}]$$

we obtain that (2.109) can be equivalently described in the form

$$\hat{H}(\mathbf{y}) = H_{\hat{m}} \quad \text{where} \quad \hat{m} = \underset{m\in\{0,1,\ldots,M-1\}}{\arg\max} \Pr\left[H = H_m \mid \mathbf{y} = \mathbf{y}\right]. \tag{2.110}$$

Hence, as in the binary ($M = 2$) case, the minimum probability-of-error decision rule (2.110) chooses the hypothesis having the largest *a posteriori* probability. As a result, for arbitrary M this is referred to as the maximum *a posteriori* (MAP) rule.

Also as in the binary case, rules such as this can typically be manipulated into simpler forms for actual implementation. For example, from Bayes' rule we have that

$$\Pr\left[H = H_m \mid \mathbf{y} = \mathbf{y}\right] = \frac{p_{\mathbf{y}|H}(\mathbf{y}|H_m)\, P_m}{\sum_{j=0}^{M-1} p_{\mathbf{y}|H}(\mathbf{y}|H_j)\, P_j}. \tag{2.111}$$

Since the denominator is, as always, a normalization constant independent of m, we can multiply (2.110) by this constant, yielding

$$\hat{H}(\mathbf{y}) = H_{\hat{m}} \quad \text{where} \quad \hat{m} = \underset{m\in\{0,1,\ldots,M-1\}}{\arg\max} p_{\mathbf{y}|H}(\mathbf{y}|H_m)\, P_m. \tag{2.112}$$

If the P_m are all equal, it follows immediately that (2.112) can be simplified to the maximum likelihood (ML) rule

$$\hat{H}(\mathbf{y}) = H_{\hat{m}} \quad \text{where} \quad \hat{m} = \underset{m\in\{0,1,\ldots,M-1\}}{\arg\max} p_{\mathbf{y}|H}(\mathbf{y}|H_m). \tag{2.113}$$

To illustrate the application of these results, let's explore a minimum probability-of-error rule in the context of simple Gaussian example.

Example 2.9

Suppose that $\mathbf{y}$ is a K-dimensional Gaussian vector under each of the M hypotheses, with

$$p_{\mathbf{y}|H}(\mathbf{y}|H_m) = N(\mathbf{y}; \mathbf{m}_m, \mathbf{I}). \tag{2.114}$$

In this case, applying (2.112)—and recognizing that we may work with the logarithm of the quantity we are maximizing—yields

$$\hat{m} = \underset{m\in\{0,1,\ldots,M-1\}}{\arg\max} \left[-\frac{K}{2}\log(2\pi) - \frac{1}{2}(\mathbf{y}-\mathbf{m}_m)^{\mathrm{T}}(\mathbf{y}-\mathbf{m}_m) + \log P_m\right]. \tag{2.115}$$

This can be simplified to

$$\hat{m} = \underset{m\in\{0,1,\ldots,M-1\}}{\arg\max} \left[\ell_m(\mathbf{y}) - \frac{1}{2}\mathbf{m}_m^{\mathrm{T}}\mathbf{m}_m + \log P_m\right] \tag{2.116}$$

where

$$\ell_m(\mathbf{y}) = \langle \mathbf{m}_m, \mathbf{y}\rangle, \qquad i = 0, 1, \ldots, M-1, \tag{2.117}$$

with the inner product and associated norm defined by

$$\langle \mathbf{x}, \mathbf{y}\rangle = \mathbf{x}^{\mathrm{T}}\mathbf{y}$$

and

$$\|\mathbf{y}\| = \sqrt{\langle \mathbf{y}, \mathbf{y} \rangle} = \sqrt{\mathbf{y}^{\mathrm{T}}\mathbf{y}}.$$

Consequently, the required data processing consists of the correlation computations in (2.117) followed by the comparisons in (2.116). Note that if, in addition, the hypotheses are equally likely and have equal signal energy—i.e., $\|\mathbf{m}_m\|^2$ is the same for all hypotheses—(2.115) and (2.116) simplify further to the minimum-distance rule

$$\begin{aligned} \hat{H}(\mathbf{y}) = H_{\hat{m}} \quad \text{where} \quad \hat{m} &= \underset{m \in \{0,1,\ldots,M-1\}}{\arg\max} \ell_m(\mathbf{y}) \\ &= \underset{m \in \{0,1,\ldots,M-1\}}{\arg\min} \|\mathbf{y} - \mathbf{m}_m\|. \end{aligned} \tag{2.118}$$

2.8.2 Structure of the General Bayesian Decision Rule

The Bayesian decision rule in the M-ary case is a natural generalization of the corresponding binary rule. To see this, we substitute Bayes' rule (2.111) into the general rule (2.107) and multiply through by the denominator in (2.111) to obtain the following rule

$$\hat{H}(\mathbf{y}) = H_{\hat{m}} \quad \text{where} \quad \hat{m} = \underset{m \in \{0,1,\ldots,M-1\}}{\arg\min} \sum_{j=0}^{M-1} C_{mj}\, P_j p_{\mathbf{y}|H}(\mathbf{y}|H_j). \tag{2.119}$$

Simplifying further by dividing by $p_{\mathbf{y}|H}(\mathbf{y}|H_0)$ and defining likelihood ratios

$$L_j(\mathbf{y}) = \frac{p_{\mathbf{y}|H}(\mathbf{y}|H_j)}{p_{\mathbf{y}|H}(\mathbf{y}|H_0)}, \qquad j = 1, 2, \ldots, M-1, \tag{2.120}$$

we obtain the rule

$$\hat{H}(\mathbf{y}) = H_{\hat{m}} \quad \text{where} \quad \hat{m} = \underset{m \in \{0,1,\ldots,M-1\}}{\arg\min} \left[C_{m0} P_0 + \sum_{j=1}^{M-1} C_{mj} P_j L_j(\mathbf{y}) \right]. \tag{2.121}$$

Let us illustrate the resulting structure of the rule for the case of three hypotheses. In this case, minimization inherent in the rule (2.121) requires access to a subset of the results from three comparisons, each of which eliminates one

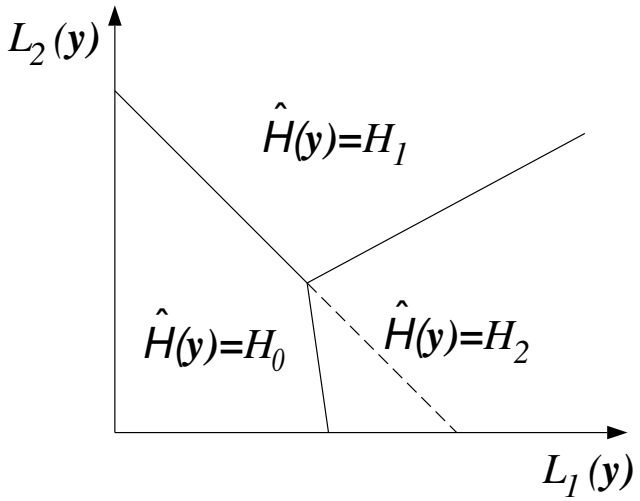


Figure 2.9. Decision Regions for the M-ary Bayesian hypothesis tests. The line segment separating the decision H_0 from H_1 (including its dashed extension) corresponds to the test (2.122a); that separating the decision H_1 from H_2 corresponds to the test (2.122b); and that separating the decision H_0 from H_2 corresponds to the test (2.122c)

hypothesis. Rearranging terms in each of these comparisons yields the following:

$$P_1(C_{01}-C_{11})L_1(\mathbf{y}) \underset{\substack{\hat{H}(\mathbf{y})=H_0 \text{ or } H_2\\ (\text{i.e., } \hat{H}(\mathbf{y})\neq H_1)}}{\overset{\substack{\hat{H}(\mathbf{y})=H_1 \text{ or } H_2\\ (\text{i.e., } \hat{H}(\mathbf{y})\neq H_0)}}{\gtrless}} P_0(C_{10}-C_{00})+P_2(C_{12}-C_{02})L_2(\mathbf{y}) \quad (2.122a)$$

$$P_1(C_{11}-C_{21})L_1(\mathbf{y}) \underset{\substack{\hat{H}(\mathbf{y})=H_1 \text{ or } H_0\\ (\text{i.e., } \hat{H}(\mathbf{y})\neq H_2)}}{\overset{\substack{\hat{H}(\mathbf{y})=H_2 \text{ or } H_0\\ (\text{i.e., } \hat{H}(\mathbf{y})\neq H_1)}}{\gtrless}} P_0(C_{20}-C_{10})+P_2(C_{22}-C_{12})L_2(\mathbf{y}) \quad (2.122b)$$

$$P_1(C_{21}-C_{01})L_1(\mathbf{y}) \underset{\substack{\hat{H}(\mathbf{y})=H_2 \text{ or } H_1\\ (\text{i.e., } \hat{H}(\mathbf{y})\neq H_0)}}{\overset{\substack{\hat{H}(\mathbf{y})=H_0 \text{ or } H_1\\ (\text{i.e., } \hat{H}(\mathbf{y})\neq H_2)}}{\gtrless}} P_0(C_{00}-C_{20})+P_2(C_{02}-C_{22})L_2(\mathbf{y}) \quad (2.122c)$$

The decision rule corresponding to (2.122) has the graphical interpretation depicted in Fig. 2.9. In particular, equality in each of the three relationships (2.122) determines a linear (affine) relationship between $L_1(\mathbf{y})$ and $L_2(\mathbf{y})$. If we plot these lines in the L_1–L_2 plane, values on one side correspond to one inequality direction, and values on the other side to the other inequality. Furthermore, it is straightforward to verify that the three straight lines intersect at a single point—it suffices to sum the three left-hand and right-hand sides of (2.122).

In this three hypothesis case, we have seen that the optimum decision rule (2.121) can be can be reduced to a set of three comparisons (2.122) involving linear combinations of two statistics, $L_1(\mathbf{y})$ and $L_2(\mathbf{y})$, which are computed from the data. More generally, examining (2.121) we can conclude that for M hypotheses with $M > 3$, the minimization involves access to a subset of $\binom{M}{2} = M(M-1)/2$ com-

parisons of linear combinations of the $M-1$ statistics $L_1(\mathbf{y}), L_2(\mathbf{y}), \ldots, L_{M-1}(\mathbf{y})$. Each of these comparisons looks at the relative merits of choosing one particular hypothesis over a second specific hypothesis. For example, in the three hypothesis case, (2.122a) compares H_0 and H_1 (see Fig. 2.9). This comparison establishes whether the point $(L_1(\mathbf{y}), L_2(\mathbf{y}))$ lies above or below the line corresponding to (2.122a) including its extension indicated by a dashed line in the figure. Once this determination is made, one of the other comparisons yields the final decision. For example, if comparison (2.122a) tells us that $(L_1(\mathbf{y}), L_2(\mathbf{y}))$ lies above the associated line, we would then use (2.122b) to decide between H_1 and H_2.

It is important to emphasize that the structure of the decision rule can be viewed as a process of eliminating one hypothesis at a time. Since we need only eliminate $M-1$ of the hypotheses, only $M-1$ of the comparisons need be used. However, we need to have all comparisons available, since the set of $M-1$ comparisons that get used depends upon the actual value of the observation. For example, if (2.122a) tells us that $(L_1(\mathbf{y}), L_2(\mathbf{y}))$ lies below the line, we would then use (2.122c) to decide between H_0 and H_2. Finally, each of these two-way comparisons has something of the flavor of a binary hypothesis testing problem. An important difference, however, is that in deciding between, say, H_0 and H_1, we need to take into account the possibility that the actual hypothesis is H_2. For example, if the cost C_{02} of deciding H_0 when H_2 is correct is much greater than the cost C_{12} of deciding H_1 when H_2 is correct, the decision rule accounts for this through a term (namely the last one in (2.122a)) in the comparison of H_0 and H_1 that favors H_1 over H_0. To be more specific, let us rewrite (2.122a) as

$$L_1(\mathbf{y}) \underset{\hat{H}(\mathbf{y}) = H_0 \text{ or } H_2}{\overset{\hat{H}(\mathbf{y}) = H_1 \text{ or } H_2}{\gtrless}} \frac{P_0(C_{10} - C_{00})}{P_1(C_{01} - C_{11})} + \frac{P_2(C_{12} - C_{02})}{P_1(C_{01} - C_{11})} L_2(\mathbf{y}), \tag{2.123}$$

and note that if the last term in (2.123) were not present, this would be exactly the binary hypothesis test for deciding between H_0 and H_1. Assuming that $C_{01} > C_{11}$ (i.e., that it is always more costly to make a mistake than to be correct) and that $C_{12} < C_{02}$, we see that the last term is negative, biasing the comparison in favor of H_1. Furthermore, this bias increases as $L_2(\mathbf{y})$ increases, i.e., when the data indicates H_2 to be more and more likely.

2.8.3 Performance Analysis

Let us next explore some aspects of the performance of the optimum decision rule for M-ary Bayesian hypothesis testing problems. Generalizing our approach from the binary case, we begin with an expression for the expected cost obtained by enumerating all the possible scenarios (i.e., deciding $\hat{H}(\mathbf{y}) = H_i$ when $\mathsf{H} = H_j$ is correct for all possible values of i and j):

$$E[C] = \sum_{i=0}^{M-1} \sum_{j=0}^{M-1} C_{ij} \Pr\left[\hat{H}(\mathbf{y}) = H_i \,\middle|\, \mathsf{H} = H_j\right] P_j. \tag{2.124}$$

Hence, the key quantities to be computed are the decision probabilities

$$\Pr\left[\hat{H}(\mathbf{y}) = H_i \;\middle|\; \mathsf{H} = H_j\right]. \tag{2.125}$$

However, since these probabilities must sum over i to unity, only $M-1$ probabilities to be calculated for each j. Thus, in total, $M(M-1)$ of these quantities need to be calculated. Note that this is consistent with the binary case ($M=2$) in which only two quantities need to be calculated, viz.,

$$P_F = \Pr\left[\hat{H}(\mathbf{y}) = H_1 \;\middle|\; \mathsf{H} = H_0\right] \quad \text{and} \quad P_D = \Pr\left[\hat{H}(\mathbf{y}) = H_1 \;\middle|\; \mathsf{H} = H_1\right].$$

In the more general case, however, not only does the number of quantities to be computed increase with M, but also the complexity of each such calculation as well. Specifically, as we have seen, reaching the decision $\hat{H}(\mathbf{y}) = H_i$ involves a set of comparisons among $M-1$ statistics. For example, in the 3-hypothesis case $\Pr[\hat{H}(\mathbf{y}) = H_m \mid \mathsf{H} = H_0]$ equals the integral over the region marked "$\hat{H}(\mathbf{y}) = H_m$" in Fig. 2.9 of the joint distribution for $L_1(\mathbf{y})$ and $L_2(\mathbf{y})$ conditioned on $\mathsf{H} = H_0$. Unfortunately, calculating these kinds of multidimensional integrals over such regions are cumbersome and can generally only be accomplished numerically. Indeed, there are few simplifications even in the Gaussian case.

Example 2.10

Let us return to Example 2.9 where the densities of $\mathbf{y}$ under the various hypotheses are given by (2.114). For simplicity, let us assume that the hypotheses are equally likely and the signals have equal energy so that the optimum decision rule is given by (2.118). Let

$$\boldsymbol{\ell}(\mathbf{y}) = \begin{bmatrix} \ell_0(\mathbf{y}) \\ \ell_1(\mathbf{y}) \\ \vdots \\ \ell_{M-1}(\mathbf{y}) \end{bmatrix} = \mathbf{M}\mathbf{y} \tag{2.126}$$

where

$$\mathbf{M} = \begin{bmatrix} \mathbf{m}_0 & \mathbf{m}_1 & \cdots & \mathbf{m}_{M-1} \end{bmatrix}^{\mathrm{T}}. \tag{2.127}$$

Then, since $\mathbf{y}$ is Gaussian under each hypothesis, so is $\boldsymbol{\ell}(\mathbf{y})$. In fact, from (2.113) we see that the second-moment statistics of the log likelihood ratio are

$$E\left[\boldsymbol{\ell} \mid \mathsf{H} = H_j\right] = \mathbf{M}\mathbf{m}_j = \begin{bmatrix} \mathbf{m}_0^{\mathrm{T}}\mathbf{m}_j \\ \mathbf{m}_1^{\mathrm{T}}\mathbf{m}_j \\ \vdots \\ \mathbf{m}_{M-1}^{\mathrm{T}}\mathbf{m}_j \end{bmatrix} \tag{2.128}$$

and

$$\begin{aligned} \boldsymbol{\Lambda}_{\boldsymbol{\ell}|\mathsf{H}=H_j} &= E\left[\left(\boldsymbol{\ell} - E\left[\boldsymbol{\ell} \mid \mathsf{H} = H_j\right]\right)\left(\boldsymbol{\ell} - E\left[\boldsymbol{\ell} \mid \mathsf{H} = H_j\right]\right)^{\mathrm{T}}\right] = \mathbf{M}\mathbf{M}^{\mathrm{T}} \\ &= \begin{bmatrix} \mathbf{m}_0^{\mathrm{T}}\mathbf{m}_0 & \mathbf{m}_0^{\mathrm{T}}\mathbf{m}_1 & \cdots & \mathbf{m}_0^{\mathrm{T}}\mathbf{m}_{M-1} \\ \mathbf{m}_1^{\mathrm{T}}\mathbf{m}_0 & \mathbf{m}_1^{\mathrm{T}}\mathbf{m}_1 & \cdots & \mathbf{m}_1^{\mathrm{T}}\mathbf{m}_{M-1} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{m}_{M-1}^{\mathrm{T}}\mathbf{m}_0 & \mathbf{m}_{M-1}^{\mathrm{T}}\mathbf{m}_1 & \cdots & \mathbf{m}_{M-1}^{\mathrm{T}}\mathbf{m}_{M-1} \end{bmatrix} \end{aligned} \tag{2.129}$$

so that

$$p_{\boldsymbol{\ell}|\mathsf{H}}(\mathbf{l}|H_j) = N(\mathbf{l}; \mathbf{M}\mathbf{m}_j, \mathbf{M}\mathbf{M}^{\mathrm{T}}) \tag{2.130}$$

In this case

$$\Pr\left[\hat{H}(\mathbf{y}) = H_i \;\middle|\; \mathsf{H} = H_j\right] = \Pr\left[\ell_i \geq \ell_m, \quad m = 0, 1, \ldots, M-1 \mid \mathsf{H} = H_j\right] \tag{2.131}$$

which is an M-dimensional integral of the distribution in (2.130) over the set in which the ith coordinate is at least as large as all of the others.

Since exact calculation of performance quantities such as those in Example 2.10 is generally impractical, approximation techniques are often employed. We next consider a class of approximation techniques, based on the union bound mentioned in Section 1.2 and its generalizations, that are frequently useful in such problems.

The Union Bound in Performance Calculations

In order to simply our development, let us restrict our attention to the costs associated with the minimum probability-of-error criterion. Specifically, suppose that the C_{ij} are given by (2.108), and in addition that the hypotheses are equally likely. In this case (2.124) becomes

$$E\left[C\right] = \Pr(e) = \frac{1}{M}\sum_{j=0}^{M-1} \Pr\left[\hat{H}(\mathbf{y}) \neq H_j \;\middle|\; \mathsf{H} = H_j\right] \tag{2.132}$$

Furthermore, from (2.110) we see that the event

$$\mathcal{E}_j = \{\hat{H}(\mathbf{y}) \neq H_j\}$$

is a union, for $k \neq j$, of the events

$$\mathcal{E}_{kj} = \{\Pr\left[\mathsf{H} = H_k \mid \mathbf{y}\right] > \Pr\left[\mathsf{H} = H_j \mid \mathbf{y}\right]\}. \tag{2.133}$$

Therefore

$$\begin{aligned}
\Pr\left[\hat{H}(\mathbf{y}) \neq H_j \;\middle|\; \mathsf{H} = H_j\right] &= \Pr\left[\mathcal{E}_j \mid \mathsf{H} = H_j\right] \\
&= \Pr\left[\bigcup_{k\neq j} \mathcal{E}_{kj} \;\middle|\; \mathsf{H} = H_j\right] \\
&= \sum_{k\neq j} \Pr\left[\mathcal{E}_{kj} \mid \mathsf{H} = H_j\right] \\
&\quad - \sum_{\substack{k\neq j \\ i\neq j \\ i\neq k}} \Pr\left[\mathcal{E}_{kj} \cap \mathcal{E}_{ij} \mid \mathsf{H} = H_j\right] \\
&\quad + \sum_{\substack{k\neq j, \\ k\neq i,}} \sum_{\substack{i\neq j, \\ k\neq n,}} \sum_{\substack{n\neq j \\ i\neq n}} \Pr\left[\mathcal{E}_{kj} \cap \mathcal{E}_{ij} \cap \mathcal{E}_{nj} \mid \mathsf{H} = H_j\right] \\
&\quad - \cdots
\end{aligned} \tag{2.134}$$

where to obtain the last equality in (2.134) we have used the natural generalization of the equality (1.5).[9]

Eq. (2.134) leads us to a natural approximation strategy. Specifically, it follows that

$$\Pr\left[\bigcup_{k\neq j}\mathcal{E}_{kj}\;\middle|\;\mathsf{H}=H_j\right]\leq\sum_{k\neq j}\Pr\left[\mathcal{E}_{kj}\mid\mathsf{H}=H_j\right] \tag{2.135}$$

since the sum on the right-hand side adds in more than once the probabilities of intersection of the $\mathcal{E}_{kj}$.

The bound (2.135) is referred to as the *union bound* and is in fact the simplest (and loosest) of a sequence of possible bounds. Specifically, while (2.135) tells us that the first term on the right of (2.134) is an upper bound to the desired probability, the first two terms together are a lower bound[10], the first three terms again form another (somewhat tighter) upper bound, etc. It is the first of these bounds, however, that leads to the simplest computations and is of primary interest, yielding

$$\Pr(e)\leq\frac{1}{M}\sum_{j=0}^{M-1}\sum_{k\neq j}\Pr\left[\mathcal{E}_{kj}\mid\mathsf{H}=H_j\right]. \tag{2.136}$$

The union bound is widely used in bit error rate calculations for communications applications.

We conclude this section by illustrating the application of this bound in the context of an example.

Example 2.11

Again, we return to Example 2.9 and its continuation 2.10. In this case, via (2.118) or, equivalently, (2.131)), we have

$$\mathcal{E}_{kj}=\{\ell_k>\ell_j\}=\{\ell_k-\ell_j>0\}$$

so that

$$\Pr(e)\leq\frac{1}{M}\sum_{j=0}^{M-1}\sum_{k\neq j}\Pr\left[\ell_k-\ell_j>0\mid\mathsf{H}=H_j\right] \tag{2.137}$$

From our earlier calculations we see that

$$p_{\ell_k-\ell_j|\mathsf{H}}(l|H_j)=N\left(l;(\mathbf{m}_k-\mathbf{m}_j)^{\mathrm{T}}\mathbf{m}_j,(\mathbf{m}_k-\mathbf{m}_j)^{\mathrm{T}}(\mathbf{m}_k-\mathbf{m}_j)\right), \tag{2.138}$$

so that $\Pr\left[\ell_k-\ell_j>0\mid\mathsf{H}=H_j\right]$ is a single error function calculation.

[9]For example, we have

$$\Pr(A\cup B\cup C)=\Pr(A)+\Pr(B)+\Pr(C)-\Pr(A\cap B)-\Pr(A\cap C)-\Pr(B\cap C)+\Pr(A\cap B\cap C).$$

[10]To see this, note that we've subtracted out probabilities of intersections of pairs of the $\mathcal{E}_{kj}$ but in the process have now missed probabilities of intersections of *three* of the $\mathcal{E}_{kj}$ together.

As a specific numerical example, suppose that there are three hypotheses and

$$\mathbf{m}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \qquad \mathbf{m}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \qquad \mathbf{m}_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

In this case, for *any* $k \neq j$ we have

$$p_{\ell_k - \ell_j | \mathsf{H}}(l|H_j) = N(l; -1, 2) \tag{2.139}$$

so that

$$\Pr(e) \leq 2 \int_0^\infty \frac{1}{\sqrt{4\pi}} e^{-(l+1)^2/4}\, dl = 2\mathcal{Q}\left(\frac{1}{\sqrt{2}}\right) \tag{2.140}$$

2.8.4 Alternative Geometrical Interpretations

It is also possible to develop alternative geometrical interpretations of the optimum Bayesian decision rule (2.107). These interpretations lend additional insight into the structure of these tests. In this section, we explore one such alternative interpretation. To begin, we define the conditional probability vector

$$\boldsymbol{\pi}(\mathbf{y}) = \begin{bmatrix} \Pr\left[\mathsf{H} = H_0 \mid \mathbf{y} = \mathbf{y}\right] \\ \Pr\left[\mathsf{H} = H_1 \mid \mathbf{y} = \mathbf{y}\right] \\ \vdots \\ \Pr\left[\mathsf{H} = H_{M-1} \mid \mathbf{y} = \mathbf{y}\right] \end{bmatrix}, \tag{2.141}$$

and note that all of the components of $\boldsymbol{\pi}(\mathbf{y})$ are nonnegative and sum to unity. As depicted in Fig. 2.10, the sets of all such probability vectors form a line when $M = 2$ and a plane when $M = 3$. Let us also define a set of cost *vectors* $\mathbf{c}_i$, each of which consists of the set of possible costs associated with making a particular decision, viz.,

$$\mathbf{c}_i = \begin{bmatrix} C_{i0} \\ C_{i1} \\ \vdots \\ C_{i,M-1} \end{bmatrix}. \tag{2.142}$$

With this new notation, the optimal decision rule (2.107) can be expressed in the form

$$\hat{H}(\mathbf{y}) = H_{\hat{m}} \quad \text{where} \quad \hat{m} = \underset{m \in \{0,1,\ldots,M-1\}}{\arg\min} \mathbf{c}_m^{\mathrm{T}} \boldsymbol{\pi}(\mathbf{y}), \tag{2.143}$$

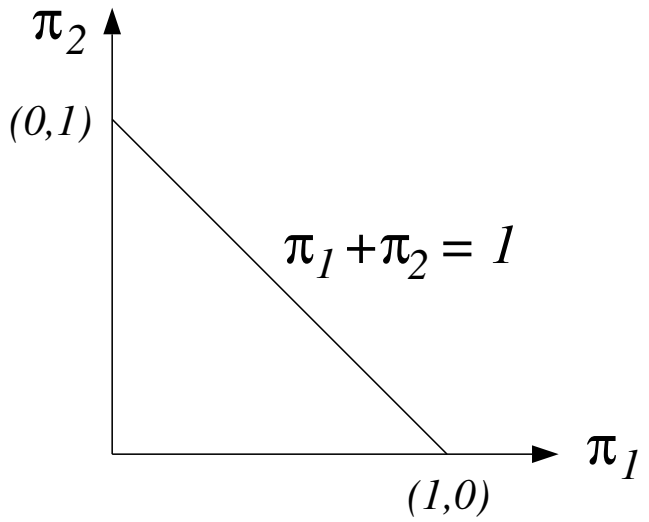
or, equivalently,

$$\hat{H}(\mathbf{y}) = H_{\hat{m}} \quad \text{if for all } m \text{ we have} \quad (\mathbf{c}_{\hat{m}} - \mathbf{c}_m)^{\mathrm{T}} \boldsymbol{\pi}(\mathbf{y}) \leq 0 \tag{2.144}$$

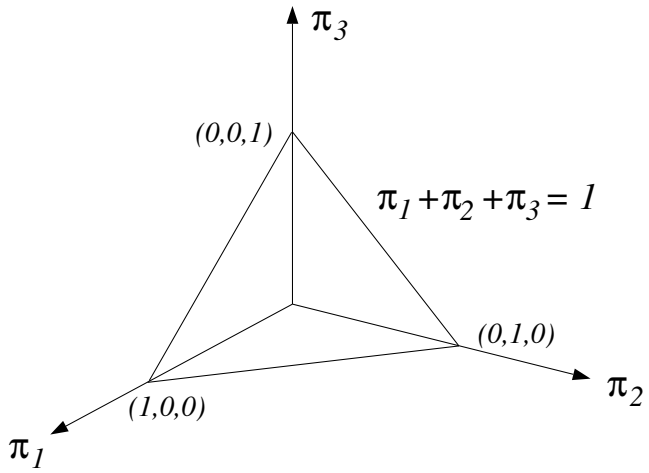
As before, this rule takes the form of a set of comparisons: for each i and k we have

$$(\mathbf{c}_k - \mathbf{c}_i)^{\mathrm{T}} \boldsymbol{\pi}(\mathbf{y}) < 0 \qquad \Longrightarrow \qquad \hat{H}(\mathbf{y}) \neq H_i \tag{2.145a}$$

$$(\mathbf{c}_k - \mathbf{c}_i)^{\mathrm{T}} \boldsymbol{\pi}(\mathbf{y}) > 0 \qquad \Longrightarrow \qquad \hat{H}(\mathbf{y}) \neq H_k \tag{2.145b}$$



(a)



(b)

Figure 2.10. Geometry optimum M-ary Bayesian decision rules in π-space.

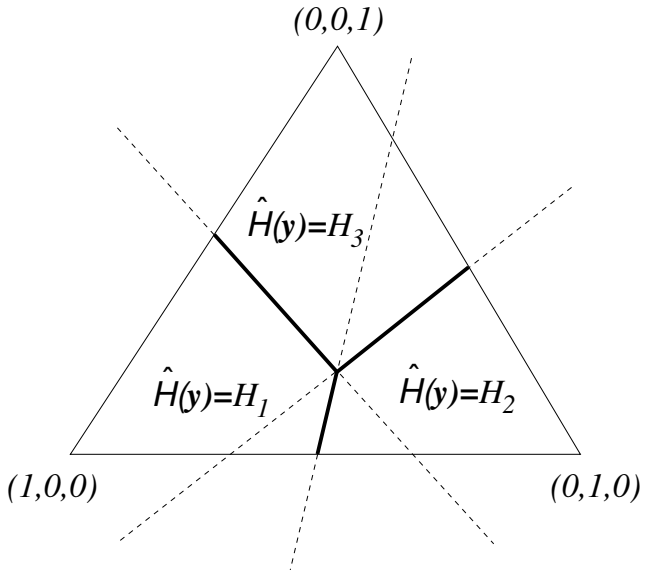


Figure 2.11. Optimum decision regions with triangle of Fig. 2.10(b).

From the vector space geometry developed in Section 1.7, we see that each of the equations $(\mathbf{c}_k - \mathbf{c}_i)^{\mathrm{T}}\boldsymbol{\pi}(\mathbf{y}) = 0$ defines a subspace that separates the space into two half-spaces corresponding to (2.145a) and (2.145b). By incorporating all of these comparisons, the set of subspaces $(\mathbf{c}_k - \mathbf{c}_i)^{\mathrm{T}}\boldsymbol{\pi}(\mathbf{y}) = 0$ for all choices of k and i partition the space into the optimum decision regions. The set of possible probability vectors, which is a subset of the space, is therefore partitioned into decision regions. For example, for $M = 3$, there are three planes through the origin in Fig. 2.10(b) that partition the space. In Fig. 2.11 we illustrate what this partitioning looks like restricted to the triangle in Fig. 2.10(b) of possible probability vectors.

2.9 RANDOM AND NONRANDOM HYPOTHESES, AND SELECTING TESTS

Throughout this chapter, recall that we have assumed that the hypotheses H_m were inherently outcomes of a random variable H. Indeed, we described the probability density for the data $\mathbf{y}$ under each hypothesis H_m as conditional densities of the form $p_{\mathbf{y}|H}(\mathrm{y}|H_m)$. And, in addition, with each H_m we associated an *a priori* probability P_m for the hypothesis.

However, it is important to reemphasize that in many problems it may not be appropriate to view the hypotheses as random—the notion of *a priori* probabilities may be rather unnatural. Rather, as we discussed at the outset of the chapter, the true hypothesis H may be a completely deterministic but unknown quantity. In these situations, it often makes more sense to view the density for the observed data not as being conditioned on the unknown hypothesis but rather as being *parameterized* by the unknown hypothesis. For such tests it is then appropriate to

use the notation

$$\begin{aligned} H_0 &: \mathbf{y} \sim p_{\mathbf{y}}(\mathbf{y}; H_0) \\ H_1 &: \mathbf{y} \sim p_{\mathbf{y}}(\mathbf{y}; H_1), \end{aligned} \tag{2.146}$$

which makes this parameterization explicit.

Hence, we can distinguish between three different classes of hypothesis testing problems:

1. The true hypothesis is random, and the *a priori* probabilities for the possibilities are known.
2. The true hypothesis is random, but the *a priori* probabilities for the possibilities are unknown.
3. The true hypothesis is nonrandom, but unknown.

For the first case, we can reasonably apply the Bayesian method provided we have a suitable means for assigning costs to the various kinds of decision errors. When we cannot meaningfully assign costs, it is often more appropriate to apply a Neyman-Pearson formulation for the problem.

In the second case, we cannot apply the Bayesian method directly. However, provided suitable cost assignments can be made, we can apply the min-max method to obtain a decision rule that is robust with respect to the unknown prior probabilities. When suitable cost assignments cannot be made, we can use the Neyman-Pearson approach in this case as well.

Finally, in the third case, corresponding to hypothesis being nonrandom, the Neyman-Pearson is natural approach. Our development in this case proceeds exactly as it does in the random hypothesis case, except for our modified notation for the densities. As a result, the optimum decision rule is a likelihood ratio test, possibly randomized in the case of discrete data, where the likelihood ratio is now defined in terms of the modified notion, i.e.,

$$L(\mathbf{y}) = \frac{p_{\mathbf{y}}(\mathbf{y}; H_1)}{p_{\mathbf{y}}(\mathbf{y}; H_0)} \tag{2.147}$$

in the continuous-valued observation case, or

$$L(\mathbf{y}) = \frac{p_{\mathbf{y}}[\mathbf{y}; H_1]}{p_{\mathbf{y}}[\mathbf{y}; H_0]} \tag{2.148}$$

in the case of discrete-valued observations.

The distinction between random and nonrandom hypotheses may seem primarily a philosophical one at this juncture in our development. However, making distinctions between random quantities and nonrandom but unknown quantities—and keeping track of the consequences of such distinctions—will make our development in subsequent chapters much easier to follow and provide some important perspectives. With this approach, it will also be easier to understand the practical implications of these distinctions as well as the connections between the various approaches to detection and estimation we develop.