Chapter 8

Detection Theory

In this chapter we start our investigation of detection theory, also referred to as hypothesis testing or decision theory. Our goal in these problems is to estimate or infer the value of an unknown "state of nature" based on noisy observations. A general model of this process is shown in Figure 8.1. Nature generates an unknown output H. By convention, we call this output a *hypothesis*. This outcome generated by nature then probabilistically affects the quantities Y that we are allowed to observe. Based on the uncertain observation Y, we must design a rule to decide what the unknown hypothesis was. In the theory of detection the set of possible hypotheses is taken to be discrete. When the set of possibilities is continuous we are in the realm of estimation, which is discussed in Chapter 10. From Figure 8.1 we see that we will need three components in our model:

1. A model of generation processes that creates H – i.e. a model of nature.
2. A model of the observation process.
3. A decision rule $D(y)$ that maps each possible observation y to an associated decision.

In general, the first two elements are set by "nature" or the restrictions of the physical data gathering situation. For example, if we are trying to decide whether a tumor is cancerous or not, the true state of the tumor is decreed by processes outside of our control and the uncertainty or noise in the observations may arise from the physical processes in generating an X-Ray image. It is generally the last element, decision rule design, where the engineer plays the strongest role. Such decision rules can be of two types: deterministic and random. A deterministic decision rule always assigns the same decision or estimate to the same observation – i.e. when a given observation is seen the same decision is always made. In particular, deterministic decision rules can be viewed as a simple partitioning or labeling of the space of observations into disjoint regions marked with the decision corresponding to each observation, as shown in Figure 8.2. In the case of random decision rules, different decisions may arise from the same observation – i.e. when the same observation is made twice, two different decision outcomes are possible. Such random decision rules play an important role when the observed quantity y is discrete in nature. In general, however, our emphasis will be on the design of deterministic decision rules.

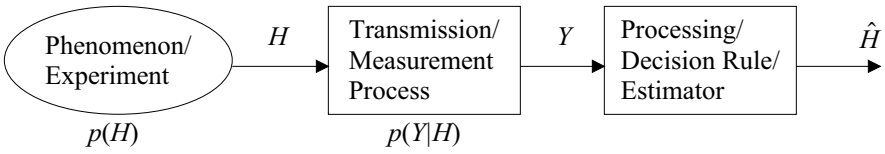

Figure 8.1: Detection problem components.

In Section 8.1 we discuss in detail the case that arises when there are only two possible hypotheses, termed binary hypothesis testing. In Section 8.4 we discuss the more general case of M hypotheses. Throughout

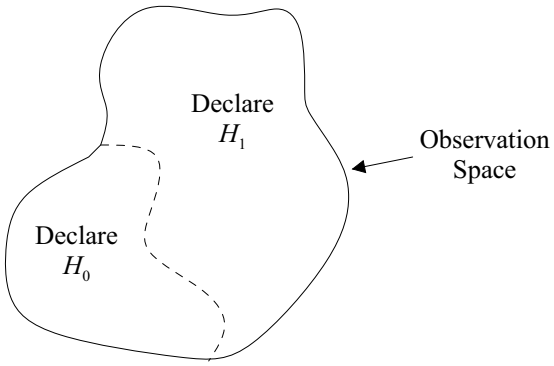

Figure 8.2: Illustration of a deterministic decision rule as a division of the observation space into disjoint regions, illustrated here for the case of two possibilities.

this chapter we focus on the case of detection based on observations of random variables. In Chapter 9 we examine the more complicated case of detection based on observations of random processes.

8.1 Bayesian Binary Hypothesis Testing

In this section we consider the simplest case when there are only two possible states of nature or hypotheses, which by convention we label as H_0 and H_1. This situation is termed "binary hypothesis testing" and the H_0 hypothesis is usually termed the "null hypothesis," due to its typical association with the absence of some quantity of interest. The binary case is of considerable practical importance, as well as having a long and rich history. To give a flavor of the possibilities, let us examine a few examples before proceeding to more detailed developments.

Example 8.1 (Communications)
Consider the following simplified version of a communication system, where a source broadcasts one bit, (either 0 or 1) The transmitter encodes this bit by a voltage, which is either 0 or E, depending on the bit. The receiver observes a noisy version of the transmitted signal, where the noise is additive, and is represented by a random variable w with zero-mean, variance σ^2, and Gaussian distribution. The receiver knows the nature of the signal E, the statistics of the noise σ^2, and the apriori probability $p(k)$ that the bit sent was k, where $k = 0, 1$. The receiver must take the received signal, y, and map this using a rule $D(y)$ into either 0 or 1, depending on the value of r. The problem is to determine the decision rule for which the probability of receiver error is minimized.

In the above example there are two possible hypotheses, H_0 and H_1, only one of which can be true. These hypotheses correspond to whether the transmitted bit was 0 or 1. There is a probabilistic relationship between the observed variable y and the hypotheses H_i. In particular, the observed variable is $y = w$ for hypothesis H_0, and $y = E + w$ for hypothesis H_1. The decision rule divides the space of possible observations into two disjoint decision regions, Z_0 and Z_1, such that, whenever an observation falls into Z_i, the decision that H_i is the correct hypothesis is made. In the example, these regions correspond to the values of y for which $D(y) = 0$ and the values of y for which $D(y) = 1$. These decision regions are established to maximize an appropriate criterion of performance, corresponding to the probability of a correct decision.

Consider other examples:

Example 8.2 (Radar)
A simple radar system makes a scalar observation y to determine the absence or presence of a target at a given range and heading. If a target is present (hypothesis H_1), the observed signal is $y = E + w$, where E is a known signal level, and $w \sim N(0, \sigma^2)$. If no target is present (hypothesis H_0), then only noise is received $y = w$. Find the decision rule for maximizing the probability of detecting the target, given a bound on the probability of false alarm.

Example 8.3 (Quality Control)
At a factory, an automatic quality control device is used to determine whether a manufactured unit is satisfactory (hypothesis H_0) or defective (hypothesis H_1), by measuring a simple quality factor q. Past statistics indicate that one out of every 10 units is defective. For satisfactory units, $q \sim N(2, \sigma^2)$, whereas for defective units, $q \sim N(1, \sigma^2)$. The quality control device is set to remove all units for which $q < t$, where t is a threshold to be designed. The problem is to determine the optimal threshold setting in order to maximize the probability of detecting a defect, subject to the constraint that the probability of removing a satisfactory unit is at most 0.005.

All of the above examples illustrate the problem of binary hypothesis testing. We will develop the relevant theory next.

8.1.1 Bayes Risk Approach and the Likelihood Ratio Test

We are now interested in obtaining "good" decision rules for the binary hypothesis testing case. A rational and common approach is to minimize a cost function given our models of the situation. Building on the development of the introduction, the elements of this approach in the binary case are:

1. **Model of Nature:** In the binary case there are only two possibilities, denoted as H_0 and H_1. Our knowledge of these possibilities is captured by the *prior probabilities* $P_i = \Pr(H = H_i)$. Note that $P_1 = 1 - P_0$.

2. **Observation Model:** As figure 8.1 indicates, the observation model captures the relationship between the observed quantity y and the unknown hypothesis H. This relationship is given by the *conditional densities* $p_{Y|H}(y \mid H_i)$.

3. **Decision Rule:** Our decision rule $D(y)$ is obtained by minimizing the average cost, called the "Bayes risk." Let C_{ij} denote the cost of deciding hypothesis $D(y) = H_i$ when hypothesis H_j is true, then the Bayes risk of the decision rule is given by:

$$E\left[C_{D(y),H}\right] = \sum_{i=0}^{1}\sum_{j=0}^{1} C_{ij}\Pr\left(D(y) = H_i, H_j \text{ true}\right) \tag{8.1}$$

Note that the outcome of deciding H_i in (8.1) is *random*, even if the decision rule is deterministic, because y itself is random. Thus the expectation in (8.1) averages over both the randomness in the true hypothesis H_j (i.e. the randomness in the state of nature) as well as the randomness in the observation, and thus decision outcome (i.e. the randomness in the data).

There are two key assumptions in the Bayes risk approach to the hypothesis testing problem which is formulated above. First, apriori probabilities of each hypothesis occurring P_i can be determined. Second, decision costs C_{ij} can be meaningfully assigned. Under these two assumptions, the Bayes risk hypothesis testing problem above is well posed. Clearly, the key is the minimization of the Bayes risk $E\left[C_{D(y)}\right]$.

Let us now focus on finding the decision rule that minimizes the Bayes risk. Recall (Figure 8.2) that a deterministic decision rule $D(y)$ is nothing more than a division of the observation space $\mathcal{R}^n$ into disjoint decision regions Z_0 and Z_1 such that when $y \in Z_i$ our decision is H_i. Thus finding a deterministic decision rule in the binary case is simply a matter of figuring out which region to assign each observation to. Combining this insight with Bayes rule we proceed by rewriting the Bayes risk as follows:

$$E\left[C_{D(y)}\right] \quad = \quad E\left[E\left[C_{D(y)} \mid y\right]\right] = \int E\left[C_{D(y)} \mid y\right]\, p_Y(y)\, dy \tag{8.2}$$

Now $p_Y(y)$ is always non-negative and the value of $E\left[C_{D(y)} \mid y\right]$ only depends on the decision region to which we assign the particular value y, so we can minimize (8.2) by minimizing $E\left[C_{D(y)} \mid y\right]$ for each value of y. Thus, the optimal decision is to choose the hypothesis that gives the smallest value of the conditional expected cost $E\left[C_{D(y)} \mid y\right]$ for the given value of y.

Now the conditional expected cost is given by:

$$E\left[C_{D(y)} \mid y\right] = \sum_{i=0}^{1}\sum_{j=0}^{1} C_{ij}\Pr\left(D(y) = H_i, H_j \text{ true} \mid y\right) \tag{8.3}$$

But $\Pr(D(y) = H_i, H_j \text{ true} \mid y)$ will either equal 0 or $\Pr(H_j \text{ true} \mid y)$ for a deterministic decision rule, since the decision outcome *given* y is non-random! In particular, for a given observation value y, the expected value of the conditional cost if we choose to assign the observation to H_0 is given by:

$$\text{If } D(y) = H_0: \qquad E\left[C_{D(y)=H_0} \mid y\right] = C_{00}\, p_{H|Y}(H_0 \mid y) + C_{01}\, p_{H|Y}(H_1 \mid y) \tag{8.4}$$

where $p_{H|Y}(H_i \mid y)$ denotes $\Pr(H_i \text{ true} \mid Y = y)$. Similarly, the expected value of the conditional cost if we assign this value of y to H_1 is given by:

$$\text{If } D(y) = H_1: \qquad E\left[C_{D(y)=H_1} \mid y\right] = C_{10}\, p_{H|Y}(H_0 \mid y) + C_{11}\, p_{H|Y}(H_1 \mid y) \tag{8.5}$$

Given the discussion above, the optimal thing to do is to make the decision that results in the smaller of the two conditional costs. We can compactly represent this comparison and its associated decision rule as follows:

$$C_{00}p_{H|Y}(H_0 \mid y) + C_{01}p_{H|Y}(H_1 \mid y) \underset{H_0}{\overset{H_1}{\gtrless}} C_{10}p_{H|Y}(H_0 \mid y) + C_{11}p_{H|Y}(H_1 \mid y) \tag{8.6}$$

where $\underset{H_0}{\overset{H_1}{\gtrless}}$ denotes choosing H_1 is the inequality is $>$ and choosing H_0 if the inequality is $<$. The decision rule given in (8.6) represents the optimal Bayes risk decision rule in its most fundamental form.

Now from Bayes rule we have that:

$$p_{H|Y}(H_i \mid y) = \frac{p_{Y|H}(y \mid H_i)p_H(H_i)}{p_Y(y)} \tag{8.7}$$

Substituting (8.7) into (8.6) and dividing through by $p_Y(y)$ we obtain:

$$(C_{01} - C_{11})\, P_1 p_{Y|H}(y \mid H_1) \underset{H_0}{\overset{H_1}{\gtrless}} (C_{10} - C_{00})\, P_0 p_{Y|H}(y \mid H_0) \tag{8.8}$$

which expresses the optimum Bayes risk decision rule in terms of the prior probabilities P_i and the data "likelihoods" $p_{Y|H}(y \mid H_i)$. Note that the expressions (8.7) and (8.8) are valid for any assignments of the costs C_{ij}.

If we further make the reasonable assumption that errors are more costly than correct decisions, so that

$$(C_{01} - C_{11}) \;>\; 0 \tag{8.9}$$

$$(C_{10} - C_{00}) \;>\; 0 \tag{8.10}$$

we can rewrite the optimum Bayes risk decision rule $D(y)$ in (8.8) as follows:

$$\mathcal{L}(y) = \left[\frac{p_{Y|H}(y \mid H_1)}{p_{Y|H}(y \mid H_0)}\right] \underset{H_0}{\overset{H_1}{\gtrless}} \frac{(C_{10} - C_{00})\, P_0}{(C_{01} - C_{11})\, P_1} \equiv \eta \tag{8.11}$$

The consequences of (8.11) are considerable and we will take some time to discuss them. First, examining (8.11) we see that the form of the optimal Bayes risk decision rule is to compare the ratio $\mathcal{L}(y)$, which is termed the *likelihood ratio*, to a threshold, which is given by η. The value of this threshold is determined, in general, by both the prior probabilities and the assigned cost structure, both of which are known at the outset of the problem (i.e. involve prior knowledge). The test (8.11) is called a *likelihood ratio test* or LRT, for obvious reasons, and thus *all* optimal decision rules (in the Bayes risk sense) are LRTs (with perhaps different thresholds). Thus, while as engineers we may disagree on such details as the assignment of costs and prior probabilities, the *form* of the optimal test (i.e. the data processing) is always the same and given by the LRT. Indeed, while the threshold η can be set by choosing costs and prior probability assignments, it is also possible to view it simply as a tunable parameter.

Second, examining (8.11) we see that the data or observations enter the decision *only* through the likelihood ratio $\mathcal{L}(y)$. Because it is a function of the uncertain observation y, it is itself a random variable.

Since this scalar function of the data is all that is needed to perform the optimal test, it is a *sufficient statistic* for the detection problem. That is, instead of making a decision based on the original observations y, it is sufficient to make the decision based only on the likelihood ratio, which is a function of y.

Finally, note that the sufficient statistic $\mathcal{L}(y)$ is a scalar. Thus the LRT is a scalar test, independent of the dimension of the observation space. This means we can make a decision in the binary case by making a single comparison, independent of whether we have 1 observation or 1 million.

Before moving on to look at special cases we note that there is another form of (8.11) that is sometimes used. In particular, taking logarithms of both sides of (8.11) does not change the inequality and results in the following equivalent test:

$$\ln\left[\mathcal{L}(y)\right] \underset{H_0}{\overset{H_1}{\gtrless}} \ln\left[\frac{(C_{10}-C_{00})\,P_0}{(C_{01}-C_{11})\,P_1}\right] \tag{8.12}$$

The quantity on the left hand side of (8.12) is called the *log-likelihood ratio*, and as we will see, is conveniently used in Gaussian problems.

8.1.2 Special Cases

Let us now consider some common special cases of the Bayes risk and the associated decision rules corresponding to them.

MPE cost assignment and the MAP rule

Suppose we use the following cost assignment:

$$C_{ij} = 1-\delta_{ij} \tag{8.13}$$

where $\delta_{ij}=1$ if $i=j$ and $\delta_{ij}=0$ if $i\neq j$. Then the cost of all errors ($C_{10}=C_{01}=1$) are the same and there is no cost for correct decisions ($C_{00}=C_{11}=0$). In this case, the Bayes risk is given by:

$$\begin{aligned} E\left[C_{D(y)}\right] &= C_{00}\Pr\left[\text{Decide } H_0,\, H_0 \text{ true}\right] \\ &\quad + C_{01}\Pr\left[\text{Decide } H_0,\, H_1 \text{ true}\right] \\ &\quad + C_{10}\Pr\left[\text{Decide } H_1,\, H_0 \text{ true}\right] \\ &\quad + C_{11}\Pr\left[\text{Decide } H_1,\, H_1 \text{ true}\right] \qquad (8.14) \\ &= \Pr\left[\text{Decide } H_0,\, H_1 \text{ true}\right] + \Pr\left[\text{Decide } H_1,\, H_0 \text{ true}\right] \qquad (8.15) \\ &= \Pr\left[\text{Error}\right] \end{aligned}$$

Thus the optimal detector for this cost assignment minimizes the probability of error. The corresponding decision rule is termed the minimum probability of error (MPE) decision rule and is given by:

$$\frac{p_{Y|H}(y \mid H_1)}{p_{Y|H}(y \mid H_0)} \underset{H_0}{\overset{H_1}{\gtrless}} \frac{P_0}{P_1} \tag{8.16}$$

Since $p_{Y|H}(y \mid H_i)P_i = p_{H|Y}(H_i \mid y)p_Y(y)$ we can rewrite the MPE decision rule (8.16) in the following form:

$$p_{H|Y}(H_1 \mid y) \underset{H_0}{\overset{H_1}{\gtrless}} p_{H|Y}(H_0 \mid y) \tag{8.17}$$

This decision rule says that for minimum probability of error choose the hypothesis whose posterior probability is higher. This is termed the *Maximum aposteriori probability or MAP rule*. Thus we see that the MAP rule is also the MPE rule independent of prior probabilities.

The ML rule

Now suppose we again use the MPE cost criterion with $C_{ij} = 1 - \delta_{ij}$, but also have both hypotheses equally likely apriori so that $P_0 = P_1 = 1/2$. In this case we essentially have no prior preference for one hypothesis over the other. With these assignments we can see that the threshold in (8.11) is given by $\eta = 1$ so that the decision rule becomes:

$$p_{Y|H}(y \mid H_1) \underset{H_0}{\overset{H_1}{\gtrless}} p_{Y|H}(y \mid H_0) \tag{8.18}$$

In this case the decision rule is to choose the hypothesis that gives the higher likelihood of the observation. For this reason this rule is called the *maximum likelihood or ML rule*

Scalar Gaussian Detection

Here we consider the problem of deciding which of two possible Gaussian distributions a single scalar observation comes from. In particular, under hypothesis H_i the observation is distributed according to:

$$p_{Y|H}(y \mid H_i) \quad = \quad N(y; m_i, \sigma_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{1}{2}\frac{(y-m_i)^2}{\sigma_i^2}} \tag{8.19}$$

These two possibilities are depicted in Figure 8.3. The likelihood ratio for this case is given by:

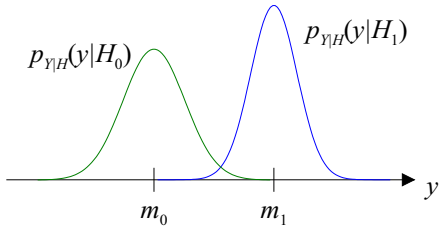

Figure 8.3: General scalar Gaussian case

$$\mathcal{L}(y) \quad = \quad \left[\frac{\left(\frac{1}{\sqrt{2\pi\sigma_1^2}}\right) e^{-\frac{(y-m_1)^2}{2\sigma_1^2}}}{\left(\frac{1}{\sqrt{2\pi\sigma_0^2}}\right) e^{-\frac{(y-m_0)^2}{2\sigma_0^2}}} \right] \underset{H_0}{\overset{H_1}{\gtrless}} \eta \tag{8.20}$$

Now taking natural logs of both sides as in (8.12) and rearranging terms results in the following form of the optimal decision rule:

$$-\frac{(y-m_1)^2}{2\sigma_1^2} + \frac{(y-m_0)^2}{2\sigma_0^2} \underset{H_0}{\overset{H_1}{\gtrless}} \ln\left(\frac{\sigma_1}{\sigma_0}\eta\right) \tag{8.21}$$

Same Variances, Different Means: Let us consider some special sub-cases. First, suppose $\sigma_0 = \sigma_1 = \sigma$ and $m_1 > m_0$. In this case the Gaussian distributions have the same variance but different means and the task is to decide whether the observation came from the Gaussian with the greater or lesser mean. After simplification, (8.21) can be reduced to the following form:

$$y \underset{H_0}{\overset{H_1}{\gtrless}} \frac{m_0 + m_1}{2} + \frac{\sigma^2 \ln(\eta)}{(m_1 - m_0)} \equiv \Gamma \tag{8.22}$$

This situation is depicted in Figure 8.4. There are some interesting things to note about this result. First there are two decision regions separated by Γ. In general, the boundary between the decision regions is an

adjusted threshold, which takes into account both the costs and the prior probabilities. For example, if we consider the ML rule (i.e. the MPE cost structure with equally likely hypotheses), then $\Gamma = (m_0 + m_1)/2$ and the boundary between decision regions is halfway between the means. In particular, in this case $\eta = 1$ and we can write the decision rule in the form:

$$\|y - m_0\|^2 \underset{H_0}{\overset{H_1}{\gtrless}} \|y - m_1\|^2 \tag{8.23}$$

which says to choose the hypothesis "closest" to the corresponding mean. If, however, instead, we use the MPE cost structure, but $P_1 > P_0$ the decision boundary will move closer to m_0, since we expect to see the H_1 case more frequently. In any case, the data processing is linear. This will not always be the case.

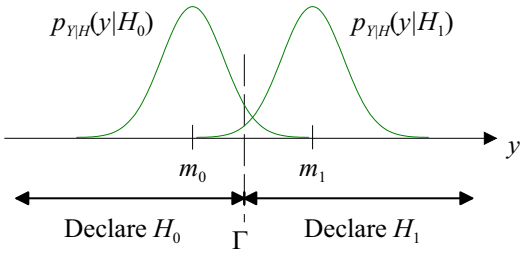

Figure 8.4: Scalar Gaussian case with equal variances

Different Variances, Same Means: Now consider what happens if we instead suppose $\sigma_0 < \sigma_1$ and $m_1 = m_0 = 0$. In this case the Gaussian distributions have the same mean, but different variances and the task is to decide whether the observation came from the Gaussian with the greater or lesser variance. After simplification, (8.21) can be reduced to the following form:

$$y^2 \underset{H_0}{\overset{H_1}{\gtrless}} 2\left(\frac{\sigma_1^2\sigma_0^2}{\sigma_1^2 - \sigma_0^2}\right) \ln\left(\frac{\sigma_1}{\sigma_0}\eta\right) \equiv \Gamma' \tag{8.24}$$

This situation is depicted in Figure 8.5. Note that the decision regions are no longer simple connected segments of the real line. Further, the decision rule is a nonlinear function of the observation y.

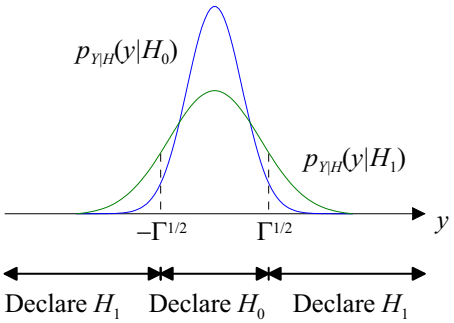

Figure 8.5: Scalar Gaussian case with equal means

8.1.3 Examples

Let us consider some examples.

Example 8.4 (Radar)
Consider the radar example, Example 8.2, discussed earlier. This is really just a scalar Gaussian detection problem. The likelihood ratio for this example is given by:

$$\mathcal{L}(y) = \frac{e^{\frac{-(y-E)^2}{2\sigma^2}}}{e^{\frac{-(y)^2}{2\sigma^2}}} = e^{\frac{2Ey-E^2}{2\sigma^2}} \tag{8.25}$$

Thus, the optimal decision rule is given by:

$$e^{\frac{2Ey-E^2}{2\sigma^2}} \underset{H_0}{\overset{H_1}{\gtrless}} \eta \tag{8.26}$$

Taking logarithms of both sides means that the new decision rule can be restated as:

$$y \underset{H_0}{\overset{H_1}{\gtrless}} \frac{E}{2} + \frac{\sigma^2 \ln(\eta)}{E} \tag{8.27}$$

In the case that the cost criterion is minimum probability of error (MPE) so that $C_{00} = C_{11} = 0, C_{01} = C_{10} = 1$, and the probability of each hypothesis is apriori equal ($P_0 = P_1 = 1/2$), we have that $\eta = 1$. Note that the optimal detection test in this case is to compute which mean the measurement is closer to! This is just an example of the scalar Gaussian detection problem treated above.

Example 8.5 (Multiple Observations)
Consider the radar detection example, except that N independent pulses are sent out, so that a vector of measurements is collected. This is the typical situation in radar systems, where multiple pulses are processed to improve the signal-to-noise ratio and thus obtain better detection performance. We assume that each pulse provides a measurement y_i, where

$$y_i = \begin{cases} n_i & \text{if hypothesis } H_0 \text{ is true (no target present)} \\ E + n_i & \text{if hypothesis } H_1 \text{ is true (target present)} \end{cases}$$

and n_i is a set of independent, identically distributed $N(0, \sigma^2)$ random variables. In this case, the likelihood ratio is given by:

$$\mathcal{L}(y) = \frac{p_{Y_1,\cdots Y_N|H}(y_1, \cdots, y_N \mid H_1)}{p_{Y_1,\cdots Y_N|H}(y_1, \cdots, y_N \mid H_0)} = \prod_{i=1}^{N} \frac{e^{\frac{-(y_i-E)^2}{2\sigma^2}}}{e^{\frac{-(y_i)^2}{2\sigma^2}}} = \prod_{i=1}^{N} e^{\frac{2Ey_i-E^2}{2\sigma^2}} = e^{\frac{2E\left(\sum_{i=1}^{N} y_i\right) - NE^2}{2\sigma^2}} \tag{8.28}$$

By again taking logs of both sides the decision rule can be reduced to:

$$\frac{1}{N}\sum_{i=1}^{N} y_i \underset{H_0}{\overset{H_1}{\gtrless}} \frac{E}{2} + \frac{\sigma^2 \ln(\eta)}{NE} \tag{8.29}$$

Comparing with (8.27), the effect of using the extra measurements is to reduce the measurement covariance by a factor of $N^{1/2}$.

Before, we said that the likelihood ratio was a sufficient statistic. It may not be the simplest sufficient statistic however. Whenever there is a function of the data, $g(y)$ such that the likelihood ratio can be computed strictly from $g(y)$, this value is *also* a sufficient statistic. Thus sufficient statistics are not unique. In the above example, it is clear that the sample mean, $\frac{1}{N}\sum_{i=1}^{N} y_i$, is a sufficient statistic for the detection problem; note that this is a linear function of the measurement vector y and much simpler than the likelihood ratio $\mathcal{L}(y_i)$ in (8.28).

Example 8.6
Assume that, under hypothesis H_0, we have a vector of N observations y, with independent, identically distributed $N(0, \sigma_0^2)$ components y_i. Under hypothesis H_1, we have a vector of N observations y, with independent, identically distributed $N(0, \sigma_1^2)$ components y_i. Thus, the two hypothesis correspond to multiple observations of independent identically

distributed random variables with the same mean but different covariances. The likelihood ratio is given by:

$$\begin{aligned}\mathcal{L}(y) &= \frac{\frac{e^{-\frac{\sum_{i=1}^{N} y_i^2}{2\sigma_1^2}}}{(2\pi\sigma_1^2)^{N/2}}}{\frac{e^{-\frac{\sum_{i=1}^{N} y_i^2}{2\sigma_0^2}}}{(2\pi\sigma_0^2)^{N/2}}} \\ &= \frac{\sigma_1^N}{\sigma_2^N} e^{-\frac{\sum_{i=1}^{N} y_i^2}{2\sigma_1^2} + \frac{\sum_{i=1}^{N} y_i^2}{2\sigma_0^2}} \end{aligned} \tag{8.30}$$

Again, after taking logs the optimal decision rule can be rewritten in terms of a simpler test, as:

$$\frac{1}{N}\sum_{i=1}^{N} y_i^2 \underset{H_0}{\overset{H_1}{\gtrless}} 2\frac{\sigma_0^2\sigma_1^2}{\sigma_1^2 - \sigma_0^2} \ln\left(\eta^{1/N}\frac{\sigma_1}{\sigma_0}\right) \tag{8.31}$$

Clearly, a sufficient statistic for this problem is the quadratic function of the measurements: $\frac{1}{N}\sum_{i=1}^{N} y_i^2$.

Before proceeding to another section, consider a problem which does not involve Gaussian random variables.

Example 8.7
Assume that we observe a random variable y which is Poisson distributed with mean m_0 when H_0 is true, and with mean m_1 when H_1 is true. Thus the likelihoods are given by:

$$p_{Y|H}(y \mid H_i) = \frac{m_i^y e^{-m_i}}{k!} \tag{8.32}$$

Note that the measurements are discrete-valued; thus, the likelihood ratios will involve probability distributions rather than densities. The likelihood ratio is given by:

$$\mathcal{L}(y) = \frac{p_{Y|H}(y \mid H_1)}{p_{Y|H}(y \mid H_0)} = \frac{m_1^y e^{-m_1}}{m_2^y e^{-m_2}} \tag{8.33}$$

Thus, the optimal decision rule can be written as:

$$y \underset{H_0}{\overset{H_1}{\gtrless}} \frac{(m_1 - m_0) + \ln(\eta)}{\ln\left(\frac{m_1}{m_0}\right)} \tag{8.34}$$

8.2 Performance and the Receiver Operating Characteristic

In the discussion so far we have focused on the form of the optimal test and on the nature of the data processing involved. We have found that the optimum Bayes risk test is the likelihood ratio test, where a function of the data (the likelihood ratio) is compared to a threshold. Let us now turn our attention to characterizing the performance of decision rules in general and LRT-based decision rules in particular. To aid in this discussion let us define the following standard terminology, arising from classical radar detection theory:

$$\begin{aligned} P_F &\equiv \Pr(\text{Choose } H_1 \mid H_0 \text{ True}) = \text{Probability of False Alarm} \qquad \text{(called a "Type I" Error)} \\ P_D &\equiv \Pr(\text{Choose } H_1 \mid H_1 \text{ True}) = \text{Probability of Detection} \\ P_M &\equiv \Pr(\text{Choose } H_0 \mid H_1 \text{ True}) = \text{Probability of Miss} \qquad \text{(called a "Type II" Error)} \end{aligned}$$

The quantity P_F is the probability that the decision rule will declare H_1 when H_0 is true, while P_D is the probability that the decision rule will declare H_1 when H_1 is true and P_M is the probability that the decision rule will declare H_0 when H_1 is true. Note carefully that these are *conditional* probabilities!

Now there are two natural metrics to evaluate the performance of a decision rule. The first metric is the expected value of the cost $E[C_{D(y)}]$, i.e. the value of the Bayes risk. Let us examine this cost in more

detail. Following (8.14), and using Bayes rule and the definitions of the conditional densities P_F, P_M, and P_D above, the Bayes risk can is given by:

$$\begin{aligned} E\left[C_{D(y)}\right] &= C_{00}\Pr\left[\text{Decide } H_0 \mid H_0\right]P_0 + C_{01}\Pr\left[\text{Decide } H_0 \mid H_1\right]P_1 \\ &\quad + C_{10}\Pr\left[\text{Decide } H_1 \mid H_0\right]P_0 + C_{11}\Pr\left[\text{Decide } H_1 \mid H_1\right]P_1 \\ &= C_{00}(1-P_F)P_0 + C_{01}(1-P_D)P_1 + C_{10}P_FP_0 + C_{11}P_DP_1 \\ &= \underbrace{C_{00}P_0 + C_{01}P_1}_{\text{Fixed Cost}} + \underbrace{(C_{10}-C_{00})\,P_0P_F - (C_{01}-C_{11})\,P_1P_D}_{\text{Varies as function of decision rule}} \end{aligned} \tag{8.35}$$

Note that this cost has two components. The first component is independent of the decision rule used, is based only on the "prior" components of the problem, and represents a fixed cost. The second component varies as a function of the decision rule (e.g. as the threshold η of the LRT is varied). In particular, of the elements in this second component it is P_F and P_D that will vary as the decision rule is changed. Thus, from a performance standpoint, we can say that $E[C_{D(y)}]$ can be expressed purely as a function of P_F and P_D (where we assume C_{ij} and P_i are fixed).

A second natural performance metric of decision rules is the probability of error Pr[error]. Starting from (8.14) and again using Bayes rule and the definitions of P_F, P_M, and P_D we find:

$$\begin{aligned} \Pr\left[\text{Error}\right] &= \Pr\left[\text{Decide } H_0,\, H_1 \text{ true}\right] + \Pr\left[\text{Decide } H_1,\, H_0 \text{ true}\right] \\ &= P_MP_1 + P_FP_0 \\ &= (1-P_D)P_1 + P_FP_0 \end{aligned} \tag{8.36}$$

Again, the parts of this expression that will vary as the decision rule is changed are P_D and P_F. Thus, we can also express Pr[Error] as a function of just P_D and P_F (again, assuming C_{ij} and P_i are fixed).

Let us summarize the development thus far. Given *any* decision rule we can determine its performance (i.e. either its corresponding Bayes risk $E\left[C_{D(y)}\right]$ or its Pr[Error]) by calculating P_D and P_F for the decision rule. Further, we know that "good" decision rules (i.e. those optimal in the Bayes risk sense) are likelihood ratio test – i.e. they compare the likelihood ratio to a fixed threshold to make their decision. The only undetermined quantity in a LRT is its threshold. Given this discussion it seems reasonable to limit ourselves to consideration of LRT decision rules and to calculate P_D and P_F for every possible value of the threshold η. Given this information, we have essentially characterized every possible "reasonable" decision rule. This information may be conveniently and compactly represented as graph of $P_D(\eta)$ versus $P_F(\eta)$ – that is, a plot of the points (P_F, P_D) as the parameter η is varied. Such an important plot for a decision rule has a special name – it is called the *Receiver Operating Characteristic* or ROC for the detection problem. An illustration of a ROC is given in Figure 8.6.

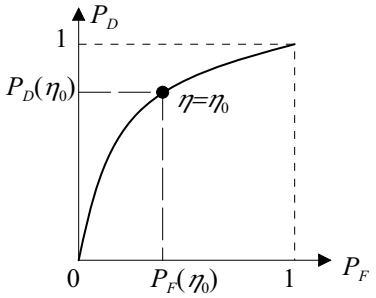

Figure 8.6: Illustration of ROC.

Let us emphasize some features of the ROC. First, note that the threshold η is a parameter along the curve. Thus any one point on the ROC corresponds to a particular choice of threshold (and vice versa). The ROC itself does not depend on the costs C_{ij} or the apriori probabilities P_i. These terms can be used, however, to determine a particular threshold, and thus a particular operating point corresponding to the

optimal Bayes risk detector. Finding appropriate values of these costs and densities can be challenging, and the ROC allows us to characterize the performance of all possible optimal detectors.

The key challenge in generating the ROC for a particular problem is finding the quantities P_D and P_F as a function of a threshold parameter. To this end, note that a general LRT decision rule can always be expressed in the following form:

$$\ell(y) \underset{H_0}{\overset{H_1}{\gtrless}} \Gamma \tag{8.37}$$

where $\ell(y)$ is a sufficient statistic for the detection problem and Γ is a corresponding threshold. The sufficient statistic might be the original likelihood ratio $\mathcal{L}(y) = p_{Y|H}(y \mid H_1)/p_{Y|H}(y \mid H_0)$ or it might be a simpler function of the observations, as we saw in the radar example. The important thing is that it completely captures the influence of the observations. Note that $\ell(y)$ is itself a random variable, since it is a function of y.

Now we can express P_D and P_F as follows:

$$\begin{aligned} P_D &= \Pr(\text{Choose } H_1 \mid H_1 \text{ True}) && (8.38)\\ &= \int_{\{y|\text{Choose } H_1\}} p_{Y|H}(y \mid H_1)\, dy && (8.39)\\ &= \int_{\ell>\Gamma} p_{L|H}(\ell \mid H_1)\, d\ell && (8.40) \end{aligned}$$

$$\begin{aligned} P_F &= \Pr(\text{Choose } H_1 \mid H_0 \text{ True}) && (8.41)\\ &= \int_{\{y|\text{Choose } H_1\}} p_{Y|H}(y \mid H_0)\, dy && (8.42)\\ &= \int_{\ell>\Gamma} p_{L|H}(\ell \mid H_0)\, d\ell && (8.43) \end{aligned}$$

The expressions (8.39) and (8.42) express the probabilities in terms of quantities in the space of the observations, i.e. in terms of the likelihoods. The expressions (8.40) and (8.43) express the probabilities in terms of quantities in the space of the test statistic and its densities. Both expressions are correct, and the choice of which to use is usually based on convenience, as we will see. Note that the region of integration (i.e. the set of values of y or ℓ used in calculation) is the *same* for both P_D and P_F, it is just the densities used that are different. We illustrate these ideas with an example.

Example 8.8 (Scalar Gaussian Detection)
Consider again the problem of determining which of two Gaussian densities of scalar observation comes from. In particular, suppose y is scalar and distributed $N(0, \sigma^2)$ under H_0 and distributed $N(E, \sigma^2)$ under H_1. We have seen in (8.27) that the optimal decision rule was:

$$\ell(y) = y \underset{H_0}{\overset{H_1}{\gtrless}} \frac{E}{2} + \frac{\sigma^2 \ln(\eta)}{E} = \Gamma$$

In this case $\ell(y) = y$ so the observation space is the same as the space of the test statistic and it is easy to see that $\ell(y)$ will be a Gaussian random variable under either hypothesis. In particular, we have:

$$\begin{aligned} p_{L|H_1}(\ell \mid H_1) &= N\left(\ell; E, \sigma^2\right)) && (8.44)\\ p_{L|H_0}(\ell \mid H_0) &= N\left(\ell; 0, \sigma^2\right)) && (8.45) \end{aligned}$$

Now we can combine these densities with (8.40) and (8.43) to find P_D and P_F as we vary Γ from $(-\infty, \infty)$, which is the range of Γ which results from variations in η. Explicitly, we have

$$\begin{aligned} P_D &= \int_{\Gamma}^{\infty} p_{L|H_1}(\ell \mid H_1)\, d\ell && (8.46)\\ &= \int_{\Gamma}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\ell-E)^2}{2\sigma^2}}\, d\ell \end{aligned}$$

$$P_F = \int_\Gamma^\infty p_{L|H_0}(\ell \mid H_0)\, d\ell = \int_\Gamma^\infty \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\ell^2}{2\sigma^2}}\, d\ell \tag{8.47}$$

These calculations of P_D and P_F are illustrated in Figure 8.7. Since these probabilities depend on the integral of Gaussian densities, we can express them in terms of the standard Q function $Q(x) = \frac{1}{2\pi}\int_x^\infty e^{-z^2/2} dz$ as follows:

$$P_D = Q\left(\frac{\Gamma - E}{\sigma}\right) \tag{8.48}$$

$$P_F = Q\left(\frac{\Gamma}{\sigma}\right) \tag{8.49}$$

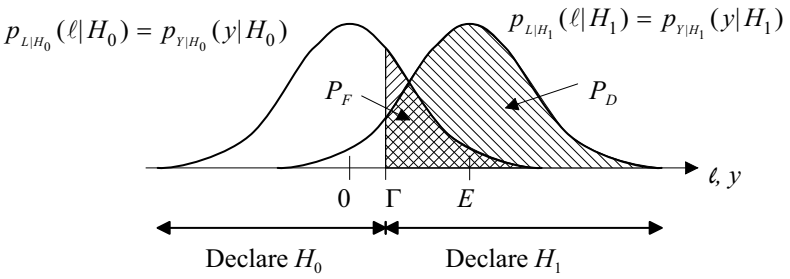

Figure 8.7: Illustration of P_D and P_F calculation.

Note that for this Gaussian detection example the performance of the detection rule really only depends on the separation of the means of the test statistic $\ell(y)$ under each hypothesis relative to the variance of the test statistic under each hypothesis – i.e. the normalized "distance" between the conditional densities. This relative or normalized distance is often an important indicator of the difficulty of a detection problem. As a result, this idea has been formalized in the definition of the so called "d^2 statistic":

$$d^2 \equiv \frac{\left(E\left[\ell \mid H_1\right] - E\left[\ell \mid H_0\right]\right)^2}{\sqrt{\mathrm{Var}\left(\ell \mid H_1\right)\mathrm{Var}\left(\ell \mid H_0\right)}} \tag{8.50}$$

The quantity d^2 can be seen to be a measure of the normalized distance between two hypotheses. In general, larger values of d^2 correspond to easier detection problems.

Example 8.9 (Scalar Gaussian Detection)
Let us continue Example 8.8. Note that:

$$d^2 = \frac{E^2}{\sigma^2} \tag{8.51}$$

which is a measure of the relative separation of the means under each hypothesis. Further we can express P_D and P_F in terms of d as follows:

$$P_F = Q\left(\frac{\Gamma}{\sigma}\right) \qquad P_D = Q\left(\frac{\Gamma}{\sigma} - d\right) \tag{8.52}$$

Larger values of d result in higher values of P_D for a given value of P_F.

8.2.1 Properties of the ROC

If we examine the expressions for P_D and P_F for Example 8.8 in more detail we can see that the corresponding ROC will possess a number of properties. First, $P_D \geq P_F$ for all thresholds Γ or η. In addition,

$\lim_{\Gamma\to-\infty} P_D = \lim_{\Gamma\to-\infty} = 1$. At the other extreme, $\lim_{\Gamma\to+\infty} P_D = \lim_{\Gamma\to+\infty} P_F = 0$. Finally, $P_D \leq 1$ and $P_F \leq 1$. Thus, the sketch in Figure 8.6 reasonably reflects this ROC. More interestingly, these properties (and others) are true for general ROC curves, and not just for the present example. We discuss these properties of the ROC next, starting with those we have just seen for our Gaussian example. We consider general likelihood ratio tests with threshold η as given in (8.11).

Property 1. The points $(P_F, P_D) = (0, 0)$ and $(P_F, P_D) = (1, 1)$ are always on the ROC. To see this, suppose we set the threshold $\eta = 0$. In this case since the densities are non-negative, the decision rule will always select H_1. In this case, $P_D = P_F = 1$. At the other extreme, assume the threshold $\eta = +\infty$. In this case the hypothesis H_0 is always selected[1]. Since H_0 is always selected $P_F = 0$ and $P_D = 0$.

Property 2. The ROC is the boundary between what is achievable by *any* decision rule and what is not. In particular, the (P_F, P_D) curve of any detection rule (including detection rules that are not LRTs) cannot lie in the shaded region shown in Figure 8.8.

Now, it is straightforward to see that we cannot get better P_D for a given P_F than that achieved by the LRT for the problem, since that would imply a detection rule resulting in lower Bayes risk (which would contradict our finding that the optimal Bayes risk decision rule is a LRT). What is perhaps less immediately obvious is that no decision rule can perform *worse* than the performance corresponding to the "reflection" of the ROC below the 45 degree line. The detector with this maximally bad performance is obtained by simply switching the decision regions for each value of η (and thus is doing the worst thing to do for every threshold). The reason is simple – if it were possible to design a decision rule with arbitrarily bad performance, than by just exchanging the decision regions we could obtain a decision rule with arbitrarily good performance. Note that the result of swapping the decision regions is that $P_D \Rightarrow 1 - P_D$ and $P_F \Rightarrow 1 - P_F$.

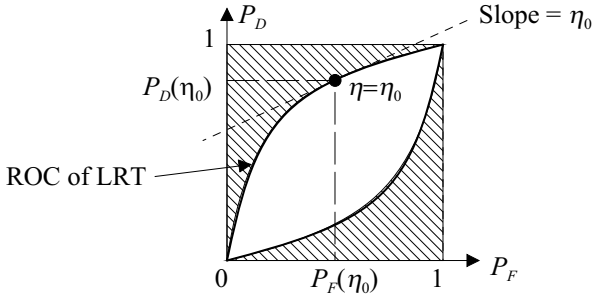

Figure 8.8: Illustration ROC properties.

Property 3. For a LRT with threshold η, the slope of the (continuous) ROC at the corresponding $(P_F(\eta), P_D(\eta))$ point is η.

To show this, first note that we may express P_D as follows:

$$P_D = \int_{\{y|\mathcal{L}(y)>\eta\}} p_{Y|H}(y \mid H_1)\, dy = \int_{\{y|\mathcal{L}(y)>\eta\}} \mathcal{L}(y)\, p_{Y|H}(y \mid H_0)\, dy \tag{8.53}$$

$$= \int_{\eta}^{\infty} Z\, p_{\mathcal{L}|H_0}(Z \mid H_0)\, dZ \tag{8.54}$$

Now, differentiating (8.54) with respect to η we obtain:

$$\frac{dP_D(\eta)}{d\eta} = -\eta\, p_{\mathcal{L}|H_0}(\eta \mid H_0) \tag{8.55}$$

[1]Note that the only way that H_1 would be selected is if we had an observation such that $p_{Y|H}(y \mid H_0) = 0$. However, for such observations there is no possibility of a false alarm, since those value cannot be generated under H_0!

Now we also know that

$$P_D = \int_\eta^\infty p_{\mathcal{L}|H_1}(\mathcal{L} \mid H_1)\, d\mathcal{L} \tag{8.56}$$

$$P_F = \int_\eta^\infty p_{\mathcal{L}|H_0}(\mathcal{L} \mid H_0)\, d\mathcal{L} \tag{8.57}$$

Differentiating these expressions with respect to η we also obtain:

$$\frac{dP_D}{d\eta} = -p_{\mathcal{L}|H_1}(\eta \mid H_1) \tag{8.58}$$

$$\frac{dP_F}{d\eta} = -p_{\mathcal{L}|H_0}(\eta \mid H_0) \tag{8.59}$$

Now equating (8.55) to (8.58) we obtain the result that:

$$\frac{p_{\mathcal{L}|H_1}(\eta \mid H_1)}{p_{\mathcal{L}|H_0}(\eta \mid H_0)} = \eta \tag{8.60}$$

Finally, the slope of the ROC is given by the derivative of P_D with respect to P_F:

$$\frac{dP_D}{dP_F} = \frac{\frac{dP_D}{d\eta}}{\frac{dP_F}{d\eta}} = \frac{-p_{\mathcal{L}|H_1}(\eta \mid H_1)}{-p_{\mathcal{L}|H_0}(\eta \mid H_0)} = \eta \tag{8.61}$$

which shows the result.

This property is illustrated in Figure 8.8. Note that a consequence of this property is that the ROC has zero slope at the point $(P_F, P_D) = (1, 1)$ $(\eta = 0)$ and infinite slope at the point $(P_F, P_D) = (0, 0)$ $(\eta = \infty)$.

Property 4. The ROC of the LRT is convex downward. In particular, $P_D \geq P_F$.

To show this property we use the concept of randomized decision rules, discussed in the following section on detection from discrete-valued observations. Suppose we select the endpoints of a randomized decision rule to be on the optimal ROC itself, as illustrated in Figure 8.9. Note that such a randomized decision rule is not necessary optimal. As a result, the optimal test must have performance (i.e. P_D for a given P_F) that is better than any randomized test. In particular, if (P_F^*, P_D^*) are the points on the ROC for the optimal Bayes decision rule, then we must have:

$$P_D^* \geq P_D(p) \quad \text{when} \quad P_F^* = P_F(p) \tag{8.62}$$

This argument shows that points on the optimal ROC between our chosen endpoints must lie above the line connecting the endpoints, and thus that the optimal ROC is convex, as shown in Figure 8.9

To see how the ROC can be used to compare the performance of different problems and detection rules, consider the following example, where we examine how the ROC changes as a function of the amount of data.

Example 8.10

Suppose we observe N independent samples of a random variable: y_i, $i = 1, \cdots, N$. Under hypothesis H_0, $p_{Y_i|H_0}(y_i \mid H_0) \sim N(0, \sigma^2)$, and under $H_1, p_{Y_i|H_0}(y_i \mid H_1) \sim N(1, \sigma^2)$. Define the vector $\underline{y}$ to be the collection of samples. Our problem is to decide whether our vector of observations came from the H_0 distribution or the H_1 distribution. This problem is similar to the N-pulse radar detection problem of Example 8.5. Using our analysis there we find that the optimal test can be written as:

$$\ell(\underline{y}) = \frac{1}{N}\sum_{i=1}^{N} y_i \underset{H_0}{\overset{H_1}{\gtrless}} \frac{1}{2} + \frac{\sigma^2 \ln(\eta)}{N} \tag{8.63}$$

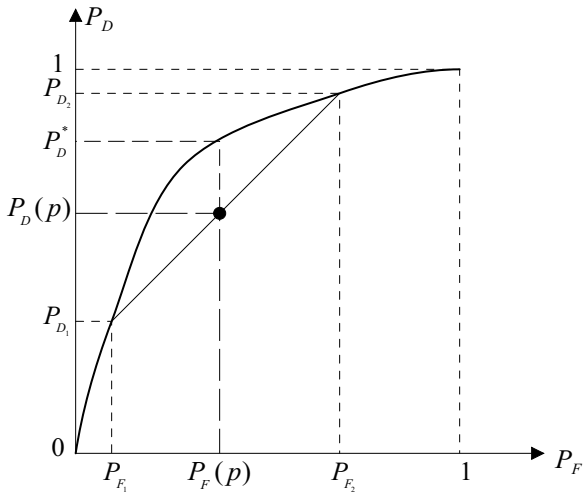

Figure 8.9: Illustration ROC convexity using randomized decision rules.

Now note that since the observations y_i are independent, the sufficient statistic for the test $\ell(\underline{y}) = \frac{1}{n}\sum_{i=1}^n y_i$ has the following probability density functions under each hypothesis:

$$H_0 \quad : \quad p_{L|H_0}(\ell \mid H_0) \sim N(0, \sigma^2/N) \tag{8.64}$$

$$H_1 \quad : \quad p_{L|H_1}(\ell \mid H_1) \sim N(1, \sigma^2/N) \tag{8.65}$$

Thus, the probability of false alarm for a given threshold $\Gamma = 1/2 + \frac{\sigma^2 \ln(\eta)}{N}$ is given by

$$P_F = \frac{1}{\sqrt{2\pi\sigma^2/n}} \int_\Gamma^\infty e^{-\frac{Nx^2}{2\sigma^2}}\, dx = Q\left(\frac{N^{1/2}\Gamma}{\sigma}\right) \tag{8.66}$$

where $Q(\Gamma) = \frac{1}{\sqrt{2\pi}} \int_\Gamma^\infty e^{-x^2/2} dx$. Similarly,

$$P_D = \frac{1}{\sqrt{2\pi\sigma^2/N}} \int_\Gamma^\infty e^{-\frac{N(x-1)^2}{2\sigma^2}}\, dx = Q\left(\frac{N^{1/2}(\Gamma-1)}{\sigma}\right) \tag{8.67}$$

Note that, as N increases the ROC curves are monotonically increasing in P_D for the same P_F, and thus nest. In particular, as we make more independent observations the curves move to the northwest and closer to their bounding box. In the limit, we have $\lim_{N\to\infty} P_D = 1, \lim_{N\to\infty} P_F = 0$, which indicates that, as $N \to \infty$. This effect is shown in Figure 8.10. Simply looking at the ROC curves for the different cases we can see the positive effect of using more observations.

Finally, note that the idea of using the ROC to evaluate the performance of decision rules is so powerful and pervasive that it is used to evaluate decision rules even when they are not, strictly speaking LRT rules for binary hypothesis testing problems.

8.2.2 Detection Based on Discrete-Valued Random Variables

The theory behind detection based on observations y that are discrete valued is essentially the same as when y is continuous valued. In particular, the LRT (8.11) is still the optimal decision rule, as considered in Example 8.7. There are some important unique characteristics of the discrete valued case that are worth discussing, however. When the observations y are discrete-valued the likelihood ratio $\mathcal{L}(y)$ will also be discrete-valued. In this case, varying the threshold η will have no effect on the values of P_F, P_D until the threshold crosses one of the discrete values of $\mathcal{L}(y)$. After crossing this discrete-value, the values of P_F, P_D will then change by a finite amount. As a result, the ROC "curve" in such a discrete observation case, obtained by varying the value of the threshold, will be a series of disconnected and isolated points. This is illustrated in the following examples.

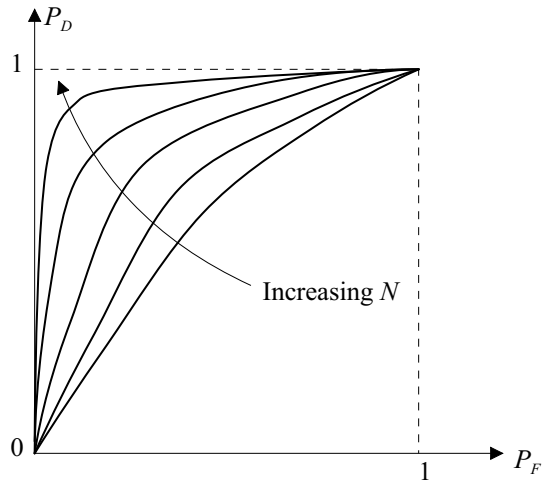

Figure 8.10: Illustration ROC behavior as we obtain more independent observations.

Example 8.11
Assume that y is a binomial random variable, resulting from the sum of two independent, identically distributed Bernoulli random variables:

$$y = x_1 + x_2 \tag{8.68}$$

The probabilities of the x_i under each hypothesis are given by:

$$\text{Under } H_0: \qquad \Pr(x_i = 1) = \frac{1}{4}; \Pr(x_i = 0) = \frac{3}{4}; \tag{8.69}$$

$$\text{Under } H_1: \qquad \Pr(x_i = 1) = \frac{1}{2}; \Pr(x_i = 0) = \frac{1}{2}; \tag{8.70}$$

Note that y can only take 3 values: 0, 1, or 2. Under these conditions, the likelihood ratio for the problem is given by:

$$\mathcal{L}(y) = \frac{p_{Y|H}(y \mid H_1)}{p_{Y|H}(y \mid H_0)} = \frac{\frac{2!}{y!(2-y)!}\left(\frac{1}{2}\right)^y\left(\frac{1}{2}\right)^{2-y}}{\frac{2!}{y!(2-y)!}\left(\frac{1}{4}\right)^y\left(\frac{3}{4}\right)^{2-y}} = \frac{1/4}{(1/4)^y(3/4)^{2-y}} \tag{8.71}$$

$$= \frac{4}{3^{2-y}} \tag{8.72}$$

The LRT for this problem is then given by:

$$\frac{4}{3^{2-y}} \underset{H_0}{\overset{H_1}{\gtrless}} \eta \tag{8.73}$$

Now note that the likelihood ratio can only take the values:

$$\mathcal{L}(y) = \begin{cases} 4/9 & \text{if } y = 0 \\ 4/3 & \text{if } y = 1 \\ 4 & \text{if } y = 2 \end{cases} \tag{8.74}$$

Now let us examine how P_D and P_F vary as we change η. If $\eta > 4$, hypothesis H_0 is always selected so that $P_D = 0$ and $P_F = 0$. Thus, these values of η correspond to the point $(P_F, P_D) = (0, 0)$ on the ROC. As η is reduced so that $4/3 < \eta < 4$, hypothesis H_1 is selected only when $y = 2$. The probability of detection is $P_D = P(y = 2 \mid H_1) = 1/4$, whereas the probability of false alarm is $P_F = P(y = 2 \mid H_0) = 1/16$. Note that P_D and P_F will have these values for *any* value of η in the range $4/3 < \eta < 4$. Thus, this entire range of η corresponds to the (isolated) point $(P_F, P_D) = (1/16, 1/4)$ on the ROC. Further reducing the threshold η so that $4/9 < \eta < 4/3$ implies that H_0 is selected only when $y = 0$. In this case, the probability of detection is $P_D = 1 - P(y = 0 \mid H_1) = 3/4$, and the probability of false alarm is $1 - P(y = 0 \mid H_0) = 7/16$. Again, note that P_D and P_F will have these values for *any* value of η in the range $4/9 < \eta < 4/3$. Again, this entire range of η thus corresponds to the (isolated) point $(P_F, P_D) = (7/16, 3/4)$ on the

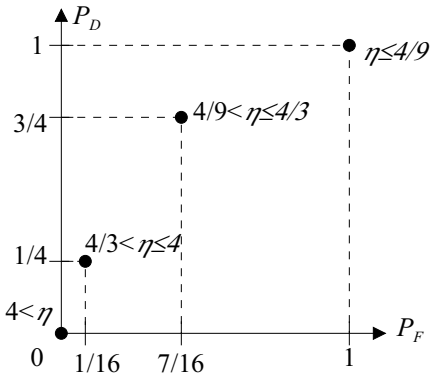

Figure 8.11: Illustration ROC for a discrete valued problem of Example 8.11.

ROC. Finally, as the threshold is lowered so that $\eta < 4/9$, hypothesis H_0 is never selected, so that $P_D = 1$ and $P_F = 1$. These values of η therefore correspond to the point $(P_F, P_D) = (1, 1)$ on the ROC. In summary, varying the threshold η produces 4 isolated points for the ROC curve for this problem, as illustrated in Figure 8.11

Let us consider another discrete valued example, this time involving Poisson random variables.

Example 8.12
Consider observing a scalar value y, which is Poisson distributed under H_0 with mean m_0, and Poisson distributed under H_1 with mean m_1. This situation was considered in Example 8.7. The optimal decision rule for this problem was found in (8.34 to be given by:

$$y \underset{H_0}{\overset{H_1}{\gtrless}} \frac{(m_1 - m_0) + \ln(\eta)}{\ln\left(\frac{m_1}{m_0}\right)} = \Gamma \tag{8.75}$$

Since y is discrete-valued, fractional parts of the effective threshold Γ on the right hand side of (8.75) will have no effect, and the ROC will again have a countable number of points.

The probability of false alarm is thus a function of the integer part of the threshold Γ, and is given by:

$$P_F(\Gamma) = \sum_{y=\lceil\Gamma\rceil}^{\infty} \frac{m_0^y}{y!} e^{-m_0} \tag{8.76}$$

where $\lceil\Gamma\rceil$ denotes the smallest integer greater than Γ. Similarly, the probability of detection is given by:

$$P_D(\Gamma) = \sum_{y=\lceil\Gamma\rceil}^{\infty} \frac{m_1^y}{y!} e^{-m_1} \tag{8.77}$$

The ROC for this problem is illustrated in Figure 8.12

The discrete nature of the ROC when the observation is discrete-valued seems to suggest that we can only obtain detection performance at a finite number of (P_F, P_D) pairs. While this observation is true if we limit ourselves to deterministic decision rules, by introducing the concept of a *randomized decision rule* we can get a much wider set of detection performance points (i.e. (P_F, P_D) points).

To introduce the idea of a randomized decision rule, suppose we have a likelihood ratio $\mathcal{L}(y)$ for an arbitrary problem (i.e. not necessarily with discrete-valued observations) and two thresholds η_0 and η_1. We then essentially have two likelihood ratio decision rules. Assume the decision rule corresponding to η_0 has performance (P_{F_0}, P_{D_0}) and the decision rule corresponding to η_1 has performance (P_{F_1}, P_{D_1}). Suppose we now define a new (random) decision rule by deciding between H_0 and H_1 according to the following probabilistic scheme:

1. Select a Bernoulli random variable Z with $\Pr(Z = 1) = p$. This is equivalent to flipping a biased coin with $\Pr(\text{heads}) = p$.

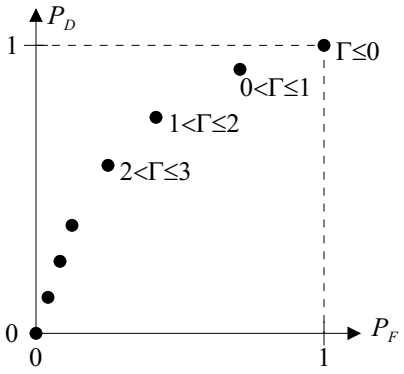

Figure 8.12: Illustration ROC for a discrete valued problem of Example 8.12.

2. If $Z = 1$ use a LRT with the threshold $\eta = \eta_1$ to make the decision.

$$\mathcal{L}(y) \underset{H_0}{\overset{H_1}{\gtrless}} \eta_1 \tag{8.78}$$

If $Z = 0$ use a LRT with the threshold $\eta = \eta_0$ to make the decision.

$$\mathcal{L}(y) \underset{H_0}{\overset{H_1}{\gtrless}} \eta_0 \tag{8.79}$$

Note that the resulting overall rule will result in a random decision. The $P_D(p)$, $P_F(p)$ performance of the overall new detection rule as a function of p can be found as:

$$\begin{aligned} P_D(p) &= \Pr(\text{Decide } H_1 \mid H_1) \\ &= \Pr(\text{Decide } H_1 \mid H_1, Z = 1)\Pr(Z = 1) + \Pr(\text{Decide } H_1 \mid H_1, Z = 0)\Pr(Z = 0) \\ &= pP_{D_1} + (1 - p)P_{D_0} \end{aligned} \tag{8.80}$$

$$\begin{aligned} P_F(p) &= \Pr(\text{Decide } H_1 \mid H_0) \\ &= \Pr(\text{Decide } H_1 \mid H_0, Z = 1)\Pr(Z = 1) + \Pr(\text{Decide } H_1 \mid H_0, Z = 0)\Pr(Z = 0) \\ &= pP_{F_1} + (1 - p)P_{F_0} \end{aligned} \tag{8.81}$$

Thus, the performance of the randomized decision rule is on the line connecting the points (P_{F_1}, P_{D_1}) and (P_{F_0}, P_{D_0}). These ideas are illustrated for a generic decision problem in Figure 8.13. By varying p we can obtain a decision rule with performance given by any (P_F, P_D) pair on the line connecting the points (P_{F_1}, P_{D_1}) and (P_{F_0}, P_{D_0}).

Now, let us return to the discrete-valued observation case. By using such randomized decision rules with the isolated points of the ROC of the deterministic decision rule as endpoints, we can obtain any (P_F, P_D) performance on the lines connecting these points. For example, the resulting ROC for Example 8.11 would be as shown in Figure 8.14. In general, we can obtain an ROC curve which is a piecewise-linear concave curve connecting the isolated points of the deterministic decision rule ROC. Further, c.f. ROC Property 2, it is impossible to get performance that is above this piecewise-linear curve (or below its mirror image).

Finally, note that ROC Property 3 can also be extended to discrete-valued random variables. Note that in this case the ROC curve is not differentiable at the discrete-valued points so the slope of the ROC curve is not defined at these points. At such points of non-differentiability, there is a range of possible slopes, defined by the slopes of the straight lines to the right and to the left of the isolated points. At these points, the value of η must be included in this range of possible slopes.

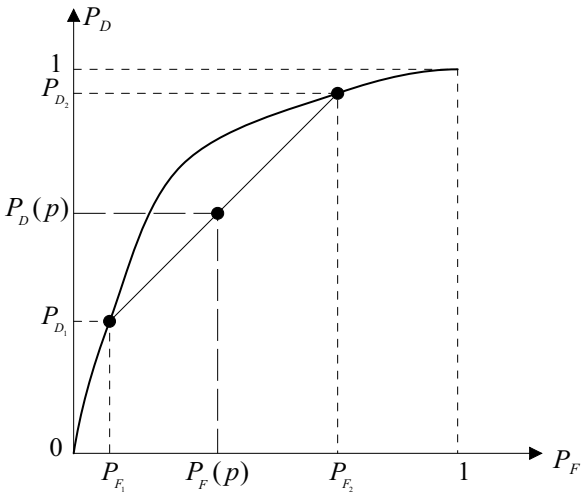

Figure 8.13: Illustration of the performance of a randomized decision rule.

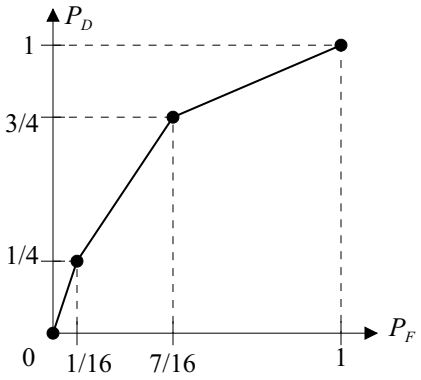

Figure 8.14: Illustration of the overall ROC obtained for a discrete valued observation problem using randomized rules.

8.3 Other Threshold Strategies

We have now determined that the form of the optimal Bayes risk test is the likelihood ratio test and have studied the performance of decision rules through use of the ROC. We have seen that the ROC compactly represents the performance of the LRT for all choices of the threshold η. In the general Bayes formulation the specific threshold η used for a given detection problem (and thus the specific operating point chosen on the ROC) is a function of the prior probabilities $P_i = \Pr(H_i)$ and the cost assignment C_{ij}:

$$\eta \equiv \frac{(C_{10} - C_{00})\, P_0}{(C_{01} - C_{11})\, P_1} \tag{8.82}$$

If we have knowledge of all these elements, then this is obviously the right (and easy) thing to do. Often, however, determining either the P_i or the C_{ij} is fraught with difficulties and an alternative strategy for picking the operating point is used. We discuss two such alternatives next.

8.3.1 Minimax Hypothesis Testing

For a given detection problem suppose that we have a cost assignment C_{ij} we believe in, but are unsure of the true prior probabilities used by nature, which are P_1^* and P_0^*. Now suppose we design a decision rule (i.e., choose a threshold) based on the costs C_{ij} and a set of *assumed* (but possibility incorrect) prior probabilities P_1 and $P_0 = 1 - P_1$. Let the performance of the resulting decision rule be given by the operating point $(P_F(P_1), P_D(P_1))$, which, as we have indicated, will be a function of our choice of P_1. Since, in general, the P_i we use to design our decision rule will be different from the true underlying P_i^*, our test will not have the minimum cost or Bayes risk for this problem. One reasonable approach in such a situation is to assume that nature will do the worst thing possible and to choose our design values of P_i (i.e. choose our threshold η) to minimize the maximum value of the cost or Bayes risk as a function of the true values P_i^*. Such a strategy leads to the *minimax decision rule*.

Now, from (8.36), the resulting cost (i.e. the Bayes risk) of a decision rule using assumed values P_i when truth is P_i^* is given by:

$$\begin{aligned} E\left(C, P_1, P_1^*\right) &= C_{00}P_0^* + C_{01}P_1^* + \left(C_{10} - C_{00}\right)P_0^* P_F(P_1) - \left(C_{01} - C_{11}\right)P_1^* P_D(P_1) \\ &= \left[\left(C_{01} - C_{00}\right) - \left(C_{10} - C_{00}\right)P_F - \left(C_{01} - C_{11}\right)P_D\right]P_1^* + C_{00} + \left(C_{10} - C_{00}\right)P_F \end{aligned} \tag{8.83}$$

where we have used the fact that $P_0^* = (1 - P_1^*)$.

On the left in Figure 8.15 we illustrate how the expected cost changes as the true prior probability P_1^* is varied. When an arbitrary fixed value of P_1 is used, the threshold is fixed, so the corresponding values of P_F and P_D are fixed. In this case we see from (8.83) that $E(C)$ will be a linear function of the true prior probability P_1^*. This is plotted as the upper curve in Figure 8.15 (left). Now if we knew P_1^* we could design an optimal LRT using an optimal threshold. In this case the threshold would change as P_1^* varied and thus so would P_F and P_D and the resulting cost. The cost of this optimal decision rule is the lower curve in Figure 8.15 (left). The two curves touch when the design value of P_1 matches the true value of P_1^*. Thus, they will always be tangent at this matched point. For the example in the figure, the maximum value of the expected cost for the non-optimal rule is obtained at the left endpoint of the curve.

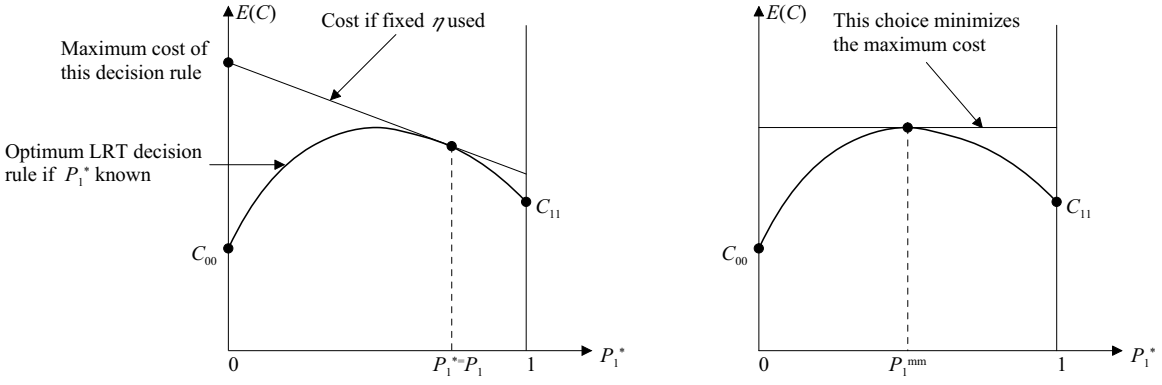

Figure 8.15: Left: Illustration of the expected cost of a decision rule using an arbitrary fixed threshold as a function of the true prior probability P_1^*. The maximum cost of this decision rule is at the left endpoint. The lower curve is the corresponding expected cost of the optimal LRT. Right: The expected cost of the minimax decision rule as a function of the true prior probability P_1^*.

In general, we would like to minimize the maximum value of (8.83). Examining Figure 8.15, we can see that this goal is accomplished if we choose our value of P_1 (or equivalently, our operating point on the ROC) so that the line (8.83) is tangent to the optimal Bayes risk curve at its maximum, as shown on the right in the figure. This happens when the slope of the curve is zero, i.e. when:

$$\left[\left(C_{01} - C_{00}\right) - \left(C_{10} - C_{00}\right)P_F - \left(C_{01} - C_{11}\right)P_D\right] = 0 \tag{8.84}$$

This result is valid as long as the maximum of the optimal Bayes cost curve is interior to the interval. When the maximum is at the boundary of the interval, then that is value of P_1 to choose.

Equation (8.84) is sometimes termed the *minimax equation* and defines the general minimax operating point. We can rewrite (8.84) in the following form:

$$P_D = \left(\frac{C_{01} - C_{00}}{C_{01} - C_{11}}\right) - \left(\frac{C_{10} - C_{00}}{C_{01} - C_{11}}\right) P_F \tag{8.85}$$

which is just a line in (P_F, P_D) space. Thus the minimax choice of operating point can be found as the intersection of the straight line (8.85) with the ROC for the optimal LRT, as shown in Figure 8.16. For example, if we use the MPE cost assignment, $C_{01} = C_{10} = 1$, $C_{00} = C_{11} = 0$, then (8.85) reduces to $P_D = 1 - P_F$ and the minimax line is just the -45 degree line.

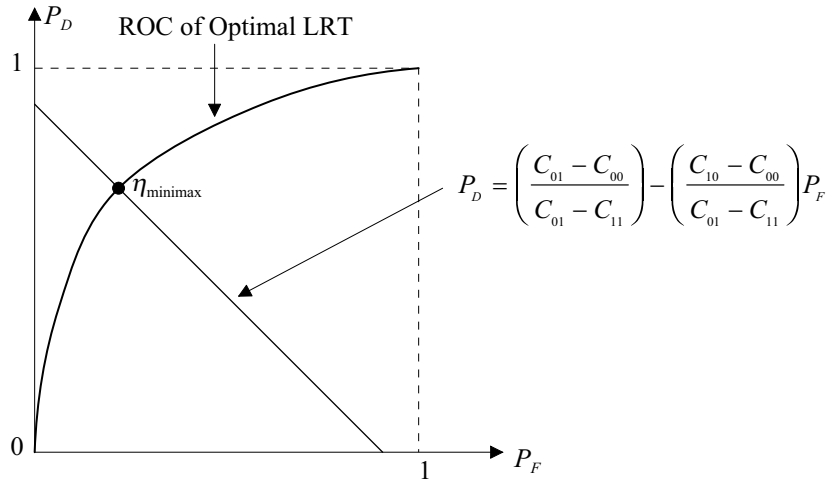

Figure 8.16: Finding the minimax operating point by intersecting (8.85) with the ROC for the optimal LRT.

8.3.2 Neyman-Pearson Hypothesis Testing

In the minimax case we assume that the costs C_{ij} can be meaningfully assigned, but that we do not know the prior probabilities. In many cases, finding such meaningful costs assignments can be difficult. This raises the question of how to choose an operating point when *neither* the prior probabilities P_i or the costs C_{ij} can be found. In general, we would like to make P_F as small as possible and P_D as large as possible. As the ROC shows, these two desires work in opposition to each other. What is often done is practice is to constrain P_F and then to maximize P_D subject to this constraint. Mathematically, one wants to solve:

$$\max P_D \quad \text{subject to } P_F \leq \alpha \tag{8.86}$$

The solution of this problem is called a *Neyman-Pearson detection rule* or "NP rule".

Note that the optimal Bayes LRT has the highest P_D for any P_F, and thus the solution of the Neyman-Pearson problem must be an optimal LRT for some choice of threshold η. So we are again in the position of needing to find an appropriate operating point on the optimal ROC. Since the ROC of the optimal LRT has P_D as a monotonically non-decreasing function of P_F, the solution of the NP problem must correspond to the point $(\alpha, P_D(\alpha))$. In the continuous-observation case, the corresponding optimal threshold η is then the slope of the ROC at this point. When the observations are discrete, we can use randomized decision rules to obtain the best P_D for any $P_F = \alpha$ and the corresponding threshold η can be found from the thresholds of the endpoint. Indeed, the desire to perform NP decision rules is one motivation for randomized decision rules in the discrete case!

Example 8.13 (Neyman-Pearson)
Suppose that the likelihoods under each hypothesis for a binary detection problem are as given in Figure 8.17. We want to find the decision rule that maximizes P_D subject to $P_F \leq 1/2$.

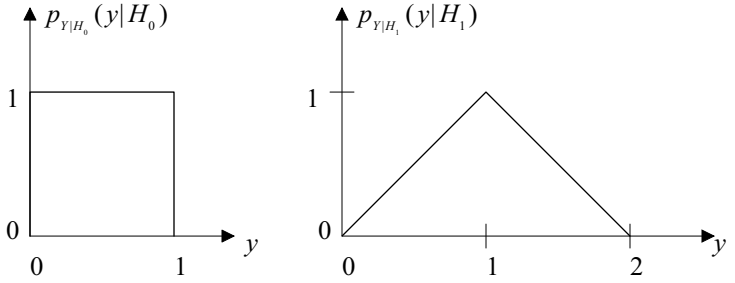

Figure 8.17: Likelihoods for a Neyman-Pearson problem.

This decision rule will be a Neyman-Pearson rule. The observation is continuous valued, so the ROC will be as well. Thus the optimal NP rule will be a LRT with threshold η chosen so that $P_F = 1/2$. We can write this rule as follows:

$$p_{Y|H}(y \mid H_1) \underset{H_0}{\overset{H_1}{\gtrless}} \eta\, p_{Y|H}(y \mid H_0) \tag{8.87}$$

Figure 8.18 shows $p_{Y|H}(y \mid H_1)$ and $\eta\, p_{Y|H}(y \mid H_0)$ on the same axes when $\eta < 1$. The corresponding decision regions are also shown. On the right of Figure 8.18 the corresponding value of $P_F = (1 - \eta)$ is shown. Now we want $P_F = 1/2$, thus we have:

$$\eta = 1 - 1/2 = 1/2 \tag{8.88}$$

The resulting decision rule is given by:

$$\frac{p_{Y|H}(y \mid H_1)}{p_{Y|H}(y \mid H_0)} \underset{H_0}{\overset{H_1}{\gtrless}} \frac{1}{2} \tag{8.89}$$

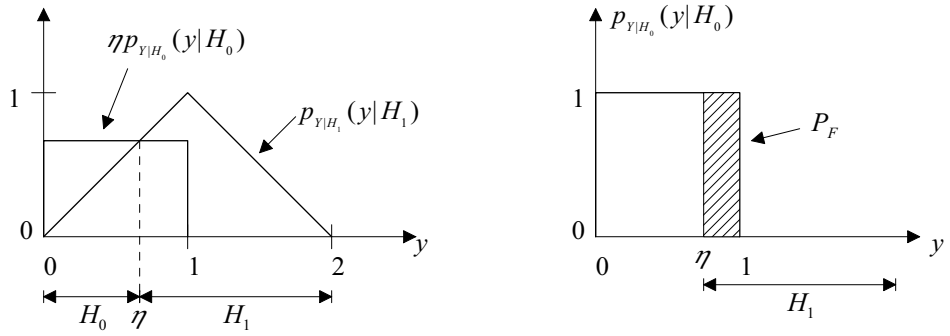

Figure 8.18: Scaled densities, decision regions and P_F for the problem of Example 8.13.

In practical problems, the bound α on P_F is determined by engineering considerations, and includes such constraints as the amount of computing power or other resources available to process false alarms. For example, a common situation we have all experienced relating to false alarm rate is in connection with car alarms. If the threshold of the car alarm is set too high, it will not trigger when the car is assaulted by thieves. On the other hand, if the threshold is set too low, the alarm will often go off even when no thief is present – creating a false alarm. If too many false alarms are generated people become exhausted and cease to check them out.

8.4 M-ary Hypothesis Testing

The exposition so far has focused on binary hypothesis testing problems. When there are M possibilities or hypotheses, we term the problem an *M-ary detection or hypothesis testing problem.* We can again take a minimum Bayes risk approach, with the same 3 three problem elements we had in the binary case:

1. **Model of Nature:** In the M-ary case there are M possibilities, denoted as H_i, $i = 0, \cdots, M-1$. Our knowledge of these possibilities is captured by the prior probabilities $P_i = \Pr(H = H_i)$, $i = 0, \cdots, M-1$. Note that $\sum_i P_i = 1$.

2. **Observation Model:** This relationship is given in the M-ary case by the M conditional densities $p_{Y|H}(y \mid h_i)$.

3. **Decision Rule:** Our decision rule $D(y)$ will again be obtained by minimizing the average cost or Bayes risk. Again, C_{ij} denotes the cost of deciding hypothesis $D(y) = H_i$ when hypothesis H_j is true and the Bayes risk is given by $E\left[C_{D(y),H}\right]$.

Note that in the M-ary case, the decision rule $D(y)$ is nothing more than a labeling of each point in the observation space with one of the corresponding possible decision outcomes H_i.

In an identical argument to the binary case, we have that the expected value of the cost is given by:

$$E\left[C_{D(y)}\right] = \int E\left[C_{D(y)} \mid y\right] p_Y(y)\, dy \tag{8.90}$$

and as before the expression is minimized by minimizing $E\left[C_{D(y)} \mid y\right]$. In particular, we should choose the decision resulting in the smallest value of this quantity. Now the expected cost of deciding H_k given y is:

$$E\left[C_{D(y)=H_k} \mid y\right] = \sum_{j=0}^{M-1} C_{kj} p_{H|Y}\left(H_j \mid y\right) \tag{8.91}$$

Thus the optimal decision rule is to choose hypothesis H_k given the observation y if:

$$\sum_{j=0}^{M-1} C_{kj} p_{H|Y}\left(H_j \mid y\right) \leq \sum_{j=0}^{M-1} C_{ij} p_{H|Y}\left(H_j \mid y\right) \qquad \forall\, i \tag{8.92}$$

The left hand side of (8.92) is the conditional cost of assigning y to the H_k decision region and the right hand side of (8.92) is the conditional cost of assigning y to the H_i decision region. Note that if the left hand side is the smallest, than assigning the given observation y to H_k is the best thing to do. Unlike the binary case, however, if the left hand side is not the smallest, we do not immediately know what the optimal hypothesis assignment is. All we know is that it is not H_k. Using this insight we can recast (8.92) in the following form, which is similar in spirit to (8.6):

$$\sum_{j=0}^{M-1} C_{kj} p_{H|Y}\left(H_j \mid y\right) \underset{\text{Not } H_i}{\overset{\text{Not } H_k}{\gtrless}} \sum_{j=0}^{M-1} C_{ij} p_{H|Y}\left(H_j \mid y\right) \qquad \forall \,\text{unique } i,k \text{ pairs} \tag{8.93}$$

where $\underset{\text{Not } H_i}{\overset{\text{Not } H_k}{\gtrless}}$ denotes eliminating hypothesis H_k if the inequality is $>$ and eliminating hypothesis H_i if the inequality is $<$. In the binary case there is only one comparison needed to define the optimal decision rule. In contrast, in the M-ary case, we need $\frac{M(M-1)}{2}$ comparisons to define the optimal decision rule. Each such comparison eliminates one of the hypotheses.

We can make (8.93) more similar to the binary case through some manipulations. In analogy with (8.11), let us define the following set of likelihood ratios:

$$\mathcal{L}_j(y) = \frac{p_{Y|H}(y \mid H_j)}{p_{Y|H}(y \mid H_0)} \qquad j = 0, \cdots, M-1 \tag{8.94}$$

where we take $\mathcal{L}_0(y) = 1$. Then, combining these likelihood ratios with Bayes rule (8.7) we have the following form for the optimal Bayes M-ary decision rule:

$$\sum_{j=0}^{M-1} C_{kj} P_j \mathcal{L}_j(y) \underset{\text{Not } H_i}{\overset{\text{Not } H_k}{\gtrless}} \sum_{j=0}^{M-1} C_{ij} P_j \mathcal{L}_j(y) \quad \forall \text{ unique } i,k \text{ pairs} \tag{8.95}$$

Note that quantities $\mathcal{L}_j(y)$ form a set of sufficient statistics for the M-ary detection problem. Further, this set of inequalities defines $M(M-1)/2$ linear decision boundaries in the space of the sufficient statistics $\mathcal{L}_i(y)$.

For example, consider the three-hypothesis case where $M = 3$. In this case, there are three comparisons that need to be performed:

$$k = 0, i = 1: \quad P_1 (C_{01} - C_{11}) \mathcal{L}_1(y) \underset{\text{Not } H_1}{\overset{\text{Not } H_0}{\gtrless}} P_0 (C_{10} - C_{00}) + P_2 (C_{12} - C_{02}) \mathcal{L}_2(y) \tag{8.96}$$

$$k = 1, i = 2: \quad P_1 (C_{11} - C_{21}) \mathcal{L}_1(y) \underset{\text{Not } H_2}{\overset{\text{Not } H_1}{\gtrless}} P_0 (C_{20} - C_{10}) + P_2 (C_{22} - C_{12}) \mathcal{L}_2(y) \tag{8.97}$$

$$k = 2, i = 0: \quad P_1 (C_{21} - C_{01}) \mathcal{L}_1(y) \underset{\text{Not } H_0}{\overset{\text{Not } H_2}{\gtrless}} P_0 (C_{00} - C_{20}) + P_2 (C_{02} - C_{22}) \mathcal{L}_2(y) \tag{8.98}$$

These comparisons are shown for a generic case in Figure 8.19. Each comparison eliminates one hypothesis. Taken together the set of comparisons labels the space of the test statistics. Note that in the space of the likelihood ratio test statistics the decision regions are always linear. In the space of the observations y this will not be true, in general. Further, the dimension of the "likelihood space" is dependent on number of hypotheses, not the dimension of the observation, which may be greater than or less than the likelihood dimension.

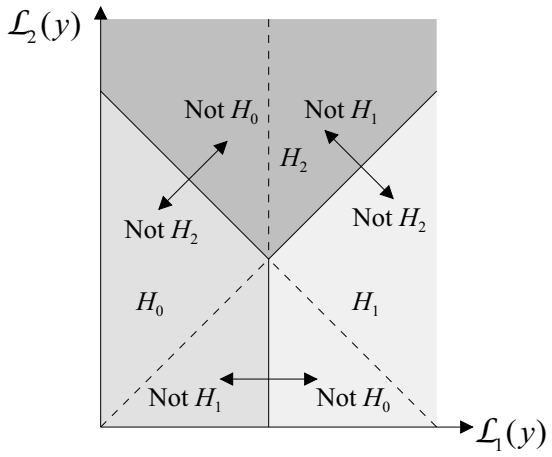

Figure 8.19: Decision boundaries in the space of the likelihoods for an M-ary problem.

8.4.1 Special Cases

Let us now consider some common special cases of the Bayes risk and the associated decision rules corresponding to them for the M-ary case.

MPE cost assignment and the MAP rule

Suppose we use the following "zero-one" cost assignment for an M-ary problem:

$$C_{ij} = 1 - \delta_{ij} \tag{8.99}$$

where $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ if $i \neq j$. Then the cost of all errors ($C_{10} = C_{01} = 1$) are the same and there is no cost for correct decisions ($C_{00} = C_{11} = 0$). As in the binary case, this cost assignment results in the Bayes risk also equaling the probability or error:

$$E\left[C_{D(y)}\right] = \sum_{j=0}^{M-1} \sum_{\substack{i=0 \\ i \neq j}}^{M-1} \Pr\left[\text{Decide } H_i,\, H_j \text{ true}\right] = \Pr\left[\text{Error}\right] \tag{8.100}$$

Thus the optimal decision rule for this cost assignment in the M-ary case also minimizes the probability of error. The corresponding decision rule (again termed the minimum probability of error (MPE) decision rule) is to choose hypothesis H_k given the observation y if:

$$p_{H|Y}\left(H_k \mid y\right) \geq p_{H|Y}\left(H_i \mid y\right) \qquad \forall\, i \tag{8.101}$$

This decision rule says that for minimum probability of error choose the hypothesis with the highest posterior probability. As in the binary case, this is termed the Maximum aposteriori probability or MAP rule. So again, the MPE cost assignment results in the MAP rule (independent of prior probabilities).

The MAP decision rule can also be expressed in terms of a series of comparisons of likelihood ratios, as in (8.95). By substituting the MPE cost structure into (8.95) and simplifying we obtain the following equivalent expression of the Bayes optimal M-ary MAP rule:

$$P_i \mathcal{L}_i(y) \underset{\text{Not } H_i}{\overset{\text{Not } H_k}{\gtrless}} P_k \mathcal{L}_k(y) \qquad \forall \text{ unique } i,k \text{ pairs} \tag{8.102}$$

Note that the details of the densities are hidden in the expressions for the likelihood ratios $\mathcal{L}_i(y)$.

The ML rule

Now suppose we again use the MPE cost criterion with $C_{ij} = 1 - \delta_{ij}$, but also have both hypotheses equally likely apriori so that $P_i = 1/M$. In this case we essentially have no prior preference for one hypothesis over the other. Applying these conditions together with Bayes rule to (8.101), this decision rule is to choose hypothesis H_k given the observation y if:

$$p_{Y|H}\left(y \mid H_k\right) \geq p_{Y|H}\left(y \mid H_i\right) \qquad \forall\, i \tag{8.103}$$

In this case the decision rule is to choose the hypothesis that gives the highest likelihood of the observation, which is again the maximum likelihood or ML rule.

As for the MAP rule, the ML decision rule can also be expressed in terms of a series of comparisons of likelihood ratios, as in (8.102). Note that the expression (8.102) already reflects the impact of the MPE cost structure. If we further incorporate the fact that $P_i = P_j$ into (8.102), we obtain the following equivalent expression of the Bayes optimal M-ary ML rule:

$$\mathcal{L}_i(y) \underset{\text{Not } H_i}{\overset{\text{Not } H_k}{\gtrless}} \mathcal{L}_k(y) \qquad \forall \text{ unique } i,k \text{ pairs} \tag{8.104}$$

8.4.2 Examples

Let us now consider some examples.

Example 8.14 (Known means in White Gaussian Noise)
Suppose we want to detect which of three possible N-dimensional signals is being received in the presence of noise. In particular, suppose that under hypothesis H_k the observation is given by:

$$\text{Under } H_k\text{:} \qquad \underline{y} = \underline{m}_k + \underline{w} \qquad k = 0, 1, 2 \tag{8.105}$$

where $\underline{w} \sim N(\underline{0}, I)$. Note that this implies that the observation densities under the different hypotheses are Gaussian, given by:

$$p_{\underline{Y}|H}\left(\underline{y} \mid H_k\right) = N\left(\underline{y}; \underline{m}_k, I\right) \qquad k = 0, 1, 2 \tag{8.106}$$

Assume that we want a minimum probability of error decision rule, which means we want the cost assignment $C_{ij} = 1 - \delta_{ij}$ and results in the MAP rule (8.101). We can also express this rule in the form (8.95). Substituting the densities given in (8.106) and simplifying, we obtain for the optimal decision rule for this example:

$$\ell_{ik}(\underline{y}) = \underline{y}^T \left(\frac{\underline{m}_k - \underline{m}_i}{\|\underline{m}_k - \underline{m}_i\|} \right) \underset{\text{Not } H_k}{\overset{\text{Not } H_i}{\gtrless}} \frac{1}{\|\underline{m}_k - \underline{m}_i\|} \left[\frac{\underline{m}_k^T \underline{m}_k - \underline{m}_i^T \underline{m}_i}{2} + \ln\left(\frac{P_i}{P_k}\right) \right] = \Gamma_{ik} \tag{8.107}$$

where we perform the comparisons over all unique i, k pairs.

Note a number of things. First, the set of $\ell_{ik}(\underline{y})$ are a set of sufficient statistics for the problem (as are the set of likelihood ratios $\mathcal{L}_i(\underline{y})$). In addition, the computation of these sufficient statistics (i.e. the processing of the data) consists of projecting the data vector onto the line between the different means and then comparing the result to a threshold. These ideas are illustrated in Figure 8.20 for a two-dimensional case. Note that the dimension of the space of the observation is independent of the number of hypotheses. Further, in this example of Gaussian densities with identical covariance matrices but different means, the decision boundaries of each comparison in (8.107) are lines (or hyperplanes, when the observations are higher dimensional). In general (i.e. when the likelihood densities are not Gaussian), these decision boundaries will not be simple linear/planar shapes.

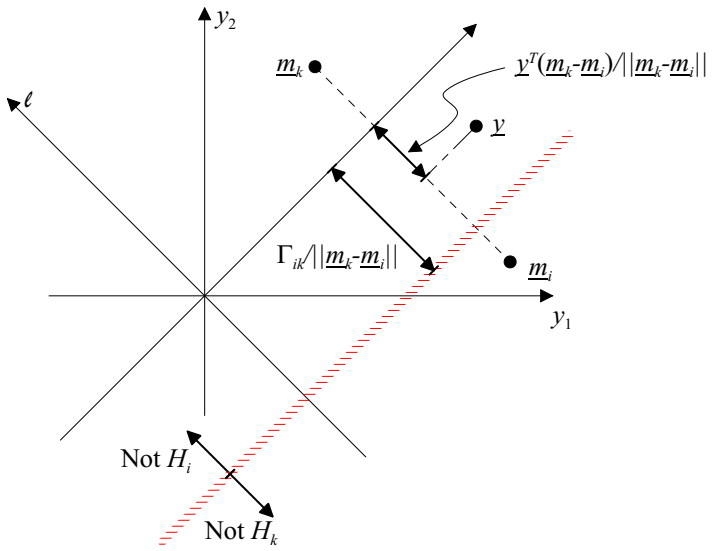

Figure 8.20: Illustration of the decision rule in the original data space.

Of course, we can also depict the decision rule for the MAP decision problem in the space of the likelihood ratios $\mathcal{L}_i$, as was done in Figure 8.19. In particular, if we express the MAP rule in the space of the original likelihood ratios for this 3 hypothesis case (i.e. by specializing (8.102) to the three hypotheses) we can express this rule as:

$$k = 0, i = 1: \qquad \mathcal{L}_1(y) \underset{\text{Not } H_1}{\overset{\text{Not } H_0}{\gtrless}} \frac{P_0}{P_1} \tag{8.108}$$

$$k = 1, i = 2: \qquad \mathcal{L}_2(y) \underset{\text{Not } H_2}{\overset{\text{Not } H_1}{\gtrless}} \left(\frac{P_1}{P_2}\right) \mathcal{L}_1(y) \tag{8.109}$$

$$k = 2, i = 0: \qquad \mathcal{L}_2(y) \underset{\text{Not } H_2}{\overset{\text{Not } H_0}{\gtrless}} \frac{P_0}{P_2} \tag{8.110}$$

In Figure 8.21 we show this decision rule in the likelihood space. When expressed in this way, the decision boundaries are independent of the specific likelihoods of the problem! That is, the decision regions for MAP rule for *any* 3-ary decision

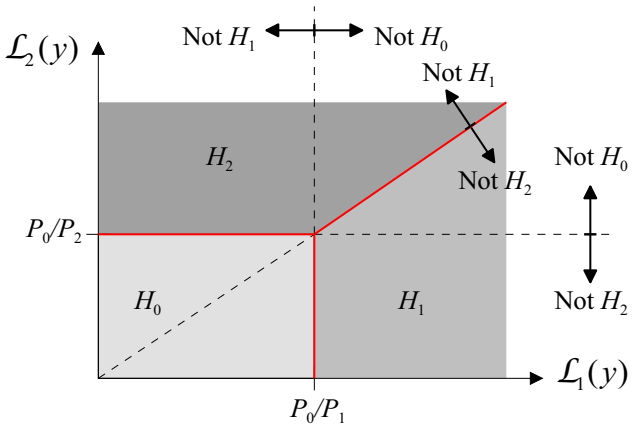

Figure 8.21: Illustration of the decision rule in the likelihood space.

problems is as given Figure 8.21. What has happened is that these likelihood details have been hidden in the likelihood ratios $\mathcal{L}(y)_i$.

Continuing with this example, suppose we additionally believe that each hypothesis is equally likely, so that $P_i = P_j = 1/3$. In this case, the decision rule will be the ML rule (8.103). Examining (8.107) and Figure 8.20, we can see that for our Gaussian example the ML rule but decision boundaries in the observation space halfway between each pair of means. Overall, the ML decision rule for this example becomes: Choose H_k if, for all i:

$$\|\underline{y} - \underline{m}_k\| \leq \|\underline{y} - \underline{m}_i\| \tag{8.111}$$

In particular, the decision rule chooses the hypothesis whose mean is closest to the given observation, resulting in the decision regions in the observation space shown in Figure 8.22 for a two-dimensional case. The decision boundaries are the bisectors of the lines connecting the means under the different hypotheses. In general, this type of decision strategy is called a *nearest neighbor classifier* or a *minimum distance receiver* in the literature. It is a strategy that is used rather widely in practice, even when it is not the optimum detector, due to its ease of implementation and understanding.

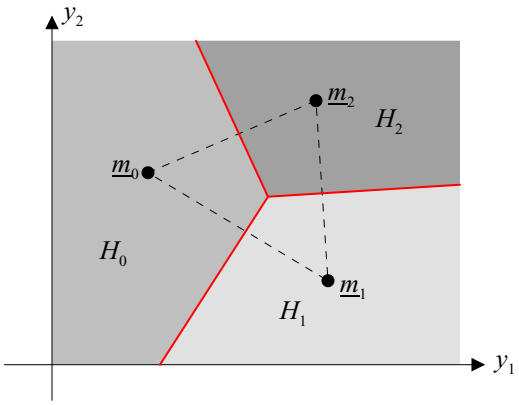

Figure 8.22: Illustration of the ML decision rule in the observation space.

Example 8.15 (Gaussians with different variances)
In this example, suppose we observe a one-dimensional random variable y and wish to determine which one of three-possible densities it could have come from. Under each of the three hypotheses the likelihoods are given by:

$$p_{Y|H}(y \mid H_i) \quad = \quad N(y; 0, \sigma_i^2) \qquad i = 0, 1, 2 \tag{8.112}$$

where $\sigma_0 < \sigma_1 < \sigma_2$. Further, suppose the hypotheses are equally likely and we wish to minimize the probability of error. In this case the decision rule will be the ML rule. Applying (8.104) and simplifying we obtain the following decision rule for this case:

$$y^2 \underset{\text{Not } H_i}{\overset{\text{Not } H_k}{\gtrless}} 2\left(\frac{\sigma_i^2\sigma_k^2}{\sigma_i^2-\sigma_k^2}\right)\ln\left(\frac{\sigma_i}{\sigma_k}\right) = \Gamma_{ik} \qquad \forall \text{ unique } i,k \text{ pairs} \tag{8.113}$$

This decision rule is shown in Figure 8.23. The decision rule in the space of the likelihoods is essentially the same as that in Figure 8.21 with $P_i/P_j = 1$.

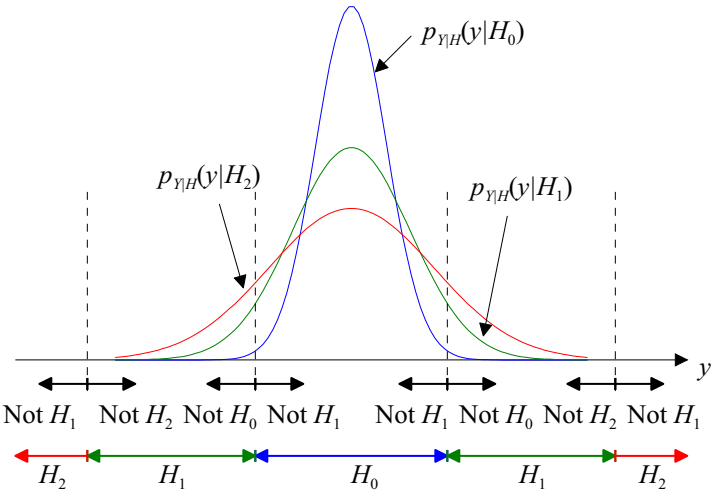

Figure 8.23: Illustration of decision rule in the observation space.

8.4.3 M-Ary Performance Calculations

The two performance metrics of the binary hypothesis testing problem were the expected value of the cost $E(C_{D(y)})$ and the probability of error Pr(Error). Both these criteria still make sense in the M-ary case, though the expressions are a bit different. In particular, whereas in the binary case we could express both the metrics in terms of only two conditional densities (P_D and P_F), in the M-ary case we need $M(M-1)$ conditional densities to express them.

First let us consider the expected value of the cost:

$$E\left[C_{D(y)}\right] = \sum_{i=1}^{M-1}\sum_{j=1}^{M-1} C_{ij}\Pr\left(\text{Decide } H_i \mid H_j\right) P_j \tag{8.114}$$

Thus, we now need $M(M-1)$ conditional densities to express the expected cost or Bayes risk versus the two needed in the binary case (i.e. P_D and P_F). So the situation is more complicated, but the idea is the same. To find the expected value of the cost (that is, the Bayes risk), we have to find a set of conditional probabilities, as before.

Consider the problem of Example 8.14 with the ML decision rule, shown in Figure 8.22. To find $\Pr\left(\text{Decide } H_0 \mid H_1\right)$ in the observation space we need to integrate the conditional density $p_{Y|H}(y \mid H_1)$ over the region of the space where we would choose hypothesis H_0. The density $p_{Y|H}(y \mid H_1)$ is a circularly symmetric Gaussian centered at the mean $\underline{m}_1$. Referring to Figure 8.22, the H_0 region of the space is the shaded region on the left. The term $\Pr\left(\text{Decide } H_0 \mid H_1\right)$ is thus the area of the Gaussian in the H_0 part of the space, as shown in Figure 8.24. The calculation of the other conditional densities is similar, where, in general, both the region of integration changes and the density being integrated changes.

Of course, if it is more convenient, we can also find these conditional densities in the space of a sufficient statistic. The basic idea is the same. Consider again the Example 8.14 with the ML decision rule, shown

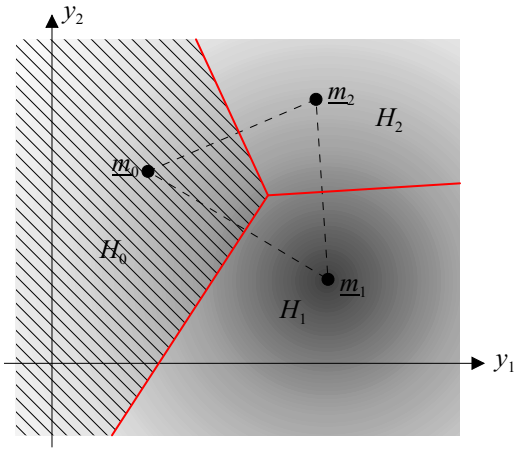

Figure 8.24: Illustration of the calculation of $\Pr(\text{Decide } H_0 \mid H_1)$ in the observation space.

this time in the space of the sufficient statistic provided by the likelihood ratios $\mathcal{L}_i(\underline{y})$ in Figure 8.21. To find $\Pr(\text{Decide } H_0 \mid H_1)$ we need to integrate the joint conditional density for the likelihood ratio sufficient statistics $p_{\mathcal{L}_1(y),\mathcal{L}_2(y)|H}(\mathcal{L}_1(y), \mathcal{L}_1(y) \mid H_0)$ over that part of the space of the likelihood ratios where we decide H_1. While the region of the likelihood space space is simply determined in this case, the required density may not be. In Example 8.14, even though the observations are Gaussian under any hypothesis, the likelihood ratios, being of the form $e^{\underline{y}^T \Sigma \underline{y}}$, will not be Gaussian random variables! All sufficient statistics are not equal, however, and a different choice of sufficient statistic may make the problem easier. Note for this example that the sufficient statistics $\ell_{ik}(y)$ defined in (8.107) are simply linear functions of the observations, and thus are themselves Gaussian random variables under any hypothesis. The decision regions are also relatively simple for these particular sufficient statistics. This discussion illustrates the issues we face in general when performing such calculations. The challenge is to find a sufficient statistic whose combination of decision regions and densities lead to a tractable set of calculations.

Our other performance metric was the probability of error $\Pr[\text{Error}]$. In the M-ary case this is given as:

$$\Pr[\text{Error}] = \sum_{j=0}^{M-1} \sum_{\substack{i=0 \\ i \neq j}}^{M-1} \Pr[\text{Decide } H_i \mid H_j \text{ true} P_j] \tag{8.115}$$

As in the calculation of the expected cost, the key is again the calculation of the conditional densities $\Pr[\text{Decide } H_i \mid H_j \text{ true} P_j]$. These probabilities can be calculated as illustrated in Figure 8.24 for Example 8.14. In the case of the $\Pr[\text{Error}]$ calculation there is an alternative form to the expression that is sometimes useful. It is based on the fact that the sum in (8.115) includes all the conditional densities except the "self term" $\Pr[\text{Decide } H_i \mid H_i \text{ true} P_j]$. As a result we may rewrite (8.115) as follows:

$$\Pr[\text{Error}] = \sum_{j=0}^{M-1} (1 - \Pr[\text{Decide } H_j \mid H_j \text{ true} P_j]) \tag{8.116}$$

Consider again the problem of Example 8.14 with the ML decision rule, shown in Figure 8.22. To find a self term, for example $\Pr(\text{Decide } H_1 \mid H_1)$, in the observation space we need to integrate the conditional density $p_{Y|H}(y \mid H_1)$ over the region of the space where we would choose hypothesis H_1. The term $\Pr(\text{Decide } H_1 \mid H_1)$ is thus the area of the Gaussian in the H_1 part of the space, as shown in Figure 8.25. The calculation of the other terms is similar. As in the case of the expected cost calculation, we may also perform such calculations in the space of a sufficient statistic if that is more convenient.

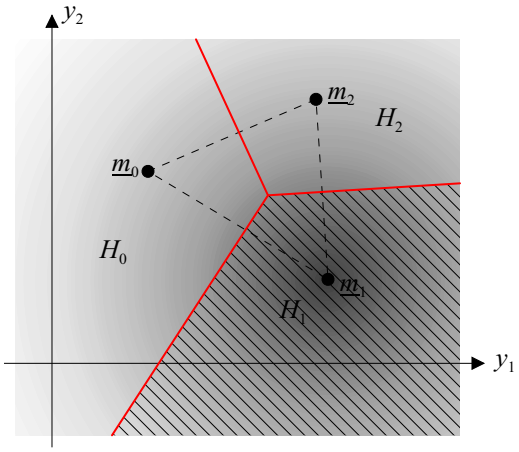

Figure 8.25: Illustration of the calculation of $\Pr(\text{Decide } H_1 \mid H_1)$ in the observation space.

8.5 Gaussian Examples

Gaussian detection problems are of general interest in many applications. In this section, several additional examples are discussed.

The general Gaussian likelihood ratio test is straightforward to compute. Let $\underline{y}$ be the n-dimensional observation vector, with hypothesized density $p_{\underline{Y}|H}(\underline{y} \mid H_0) \sim N(\underline{m}_0, \Sigma_0)$ under H_0 and density $p_{\underline{Y}|H}(\underline{y} \mid H_1) \sim N(\underline{m}_1, \Sigma_1)$ under H_1. Then the likelihood ratio test for the general Gaussian case is given by:

$$\mathcal{L}(\underline{y}) = \frac{p_{\underline{Y}|H}(\underline{y} \mid H_1)}{p_{\underline{Y}|H}(\underline{y} \mid H_0)} = \frac{\frac{1}{\sqrt{(2\pi)^N |\Sigma_1|}} e^{-\frac{1}{2}(y-\underline{m}_1)^T \Sigma_1^{-1}(y-\underline{m}_1)}}{\frac{1}{\sqrt{(2\pi)^N |\Sigma_0|}} e^{-\frac{1}{2}(y-\underline{m}_0)^T \Sigma_0^{-1}(y-\underline{m}_0)}} \underset{H_0}{\overset{H_1}{\gtrless}} \eta \tag{8.117}$$

where $|\Sigma_i|$ is the determinant of Σ_i. Taking logarithms of both sides and clearing out factors of $1/2$, one obtains the following form of the LRT:

$$\ell(\underline{y}) = -(\underline{y} - \underline{m}_1)^T \Sigma_1^{-1} (\underline{y} - \underline{m}_1) + (y - \underline{m}_0)^T \Sigma_0^{-1} (\underline{y} - \underline{m}_0) \underset{H_0}{\overset{H_1}{\gtrless}} 2\ln(\eta) + \ln(|\Sigma_1|) - \ln(|\Sigma_0|) \tag{8.118}$$

The above expression indicates that a sufficient statistic is $\ell(\underline{y}) = (\underline{y} - \underline{m}_0)^T \Sigma_0^{-1} (\underline{y} - \underline{m}_0) - (\underline{y} - \underline{m}_1)^T \Sigma_1^{-1} (\underline{y} - \underline{m}_1)$.

Example 8.16

Consider now the detection of known signals in additive Gaussian noise, where the elements y_j of the observation vector $\underline{y}$ are given by the following expression under each hypothesis:

$$H_i: \qquad y_j = m_{ij} + w_j \tag{8.119}$$

where the values of m_{ij} are known, and w_j is an independent, identically distributed sequence of Gaussian random variables with distribution $N(0, \sigma^2)$. Note that, in this case, $\Sigma_1 = \Sigma_0 = \sigma^2 I$, so that the sufficient statistic becomes:

$$\ell(\underline{y}) = \frac{2}{\sigma^2} (\underline{m}_1 - \underline{m}_0)^T \underline{y} + \frac{\underline{m}_0^T \underline{m}_0 - \underline{m}_1^T \underline{m}_1}{\sigma^2} \tag{8.120}$$

The optimal detector can be written as

$$\underline{y}^T (\underline{m}_1 - \underline{m}_0) \underset{H_0}{\overset{H_1}{\gtrless}} \frac{\underline{m}_1^T \underline{m}_1 - \underline{m}_0^T \underline{m}_0}{2} + \sigma^2 \ln(\eta) \tag{8.121}$$

Example 8.17 (Uniformly Most Powerful Test)
One can also detect unknown signals in Gaussian noise, as follows: Assume that the observations are distributed as

$$y_j = \begin{cases} w_j & \text{if } H_0 \text{ is true} \\ x_j + w_j & \text{otherwise} \end{cases} \tag{8.122}$$

where x_j is the j-th coefficient of a Gaussian vector $\underline{x}$ which is independent of $\underline{w}$, with distribution $N(\underline{m}_x, \Sigma_x)$. Again, this is a Gaussian detection problem, with $\underline{m_1} = \underline{m}_x$, $\Sigma_1 = \Sigma_x + \sigma^2 I$, $\Sigma_0 = \sigma^2 I$, $\underline{m}_0 = 0$. In this case, the sufficient statistic becomes

$$\ell(\underline{y}) = \sigma^{-2}(\underline{y}^T\underline{y}) - (\underline{y} - \underline{m}_x)^T\Sigma_1^{-1}(\underline{y} - \underline{m}_x) = \underline{y}^T[\sigma^{-2}\underline{y} - \Sigma_1^{-1}(y - \underline{m}_x) + \Sigma_1^{-1}\underline{m}_x] - \underline{m}_x^T\Sigma_1^{-1}\underline{m}_x \tag{8.123}$$

Thus, the optimal detector is to declare H_1 whenever

$$\underline{y}^T\left[\sigma^{-2}\underline{y} - \Sigma_1^{-1}\left(\underline{y} - \underline{m}_x\right) + \Sigma_1^{-1}\underline{m}_x\right] \underset{H_0}{\overset{H_1}{\gtrless}} 2\ln(\eta) + \underline{m}_x^T\Sigma_1^{-1}\underline{m}_x + \ln\left(|\mathsf{det}(\Sigma_1)|\right) - \ln\left(|\mathsf{det}(\Sigma_0)|\right) \tag{8.124}$$

It is interesting to examine the term on the right-hand side. In particular, note the following relationships which hold true under H_1:

$$\Sigma_{yy} = E\left[\left(\underline{y} - \underline{m}_x\right)\left(\underline{y} - \underline{m}_x\right)^T \middle| H_1\right] = \Sigma_1 = \Sigma_x + \sigma^2 I \tag{8.125}$$

$$\Sigma_{xy} = E[(\underline{x} - \underline{m}_x)(\underline{y} - \underline{m}_x)^T \mid H_1] = \Sigma_x \tag{8.126}$$

Thus,

$$\sigma^{-2}\underline{y} - \Sigma_1^{-1}\underline{y} = \Sigma_1^{-1}(\sigma^{-2}(\Sigma_x + \sigma^2 I) - I) = \sigma^{-2}\Sigma_1^{-1}\Sigma_x$$

The above expression can be given an interesting interpretation. Consider the case where $\underline{m}_x = 0$. Then, using the expression for Gaussian estimation,

$$E[\underline{x} \mid \underline{y}, H_1] = \Sigma_1^{-1}\Sigma_x\underline{y} \tag{8.127}$$

and the optimal detection rule selects H_1 whenever

$$\underline{y}^T E[\underline{x} \mid \underline{y}, H_1] > 2\sigma^2[\ln(T) + \ln(|\mathsf{det}(\Sigma_1)|) - \ln(|\mathsf{det}(\sigma^2 I)|)] \tag{8.128}$$

In particular, this decision rule is similar to the known signal case, except that the known difference in the means is replaced by $E[\underline{x} \mid y, H_1]$.