C H A P T E R 13

Hypothesis Testing

INTRODUCTION

The topic of hypothesis testing arises in many contexts in signal processing and communications, as well as in medicine, statistics and other settings in which a choice among multiple options or hypotheses is made on the basis of limited and noisy data. For example, from tests on such data, we may need to determine: whether a person does or doesn't have a particular disease; whether or not a particular radar return indicates the presence of an aircraft; which of four values was transmitted at a given time in a PAM system; and so on.

Hypothesis testing provides a framework for selecting among M possible choices or hypotheses in some principled or optimal way. In our discussion we will initially focus on $M = 2$, i.e., on binary hypothesis testing, to illustrate the key concepts. Though Section 13.1 introduces the discussion in the context of binary pulse amplitude modulation in noise, the presentation and results in Section 13.2 apply to the general problem of binary hypothesis testing. In Sections 13.3 and 13.4 we explicitly treat the case of more than two hypotheses.

13.1 BINARY PULSE AMPLITUDE MODULATION IN NOISE

In Chapter 12 we introduced the basic principles of pulse amplitude modulation, and considered the effects of pulse rate, pulse shape, and channel and receiver filtering in PAM systems. We also developed and discussed the condition for no inter-symbol interference (the no-ISI condition). Under the assumption of no ISI, we want to now examine the effect of noise in the channel. Toward this end, we again consider the overall PAM model in Figure 13.1, with the channel noise $v(t)$ represented as an additive term.

For now we will assume no post-filtering at the receiver, i.e., assume $f(t) = \delta(t)$. In Chapter 14 we will see how performance is improved with the use of filtering in the receiver. The basic pulse $p(t)$ going through the channel with impulse response $h(t)$ produces a signal at the channel output that we represent by $s(t) = p(t) * h(t)$. Figure 13.1 thus reduces to the overall system shown in Figure 13.2.

Since we are assuming no ISI, we can carry out our discussion for just a single pulse index n, which we will choose as $n = 0$ for convenience. We therefore focus, in the system of Figure 13.2, on

$$b[0] = r(0) = a[0]s(0) + v(0) \ . \qquad (13.1)$$

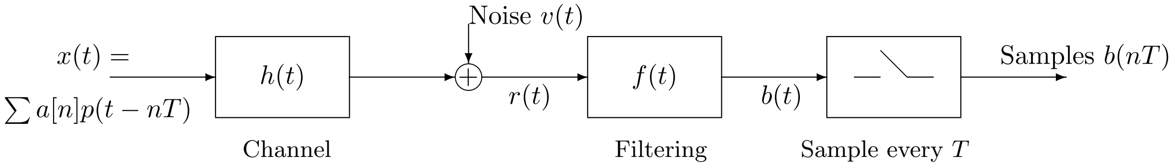

FIGURE 13.1 Overall model of a PAM system.

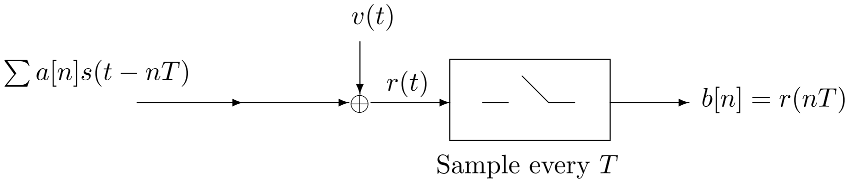

FIGURE 13.2 Simplified representation of a PAM system.

Writing $r(0)$, $a[0]$ and $v(0)$ simply as r, a and v respectively, and setting $s(0) = 1$ without loss of generality, the relation of interest to us is

$$r = a + v \; . \tag{13.2}$$

Our broad objective is to determine the value of a as well as possible, given the measured value r. There are several variations of this problem, depending on the nature of the transmitted sequence $a[n]$ and the characteristics of the noise. The amplitude $a[n]$ may span a continuous range or it may be discrete (e.g., binary). The amplitude may correspondingly be modeled as a random variable A with a known PDF or PMF; then a is the specific value that A takes in a particular outcome or instance of the probabilistic model. The contribution of the noise also is typically represented as a random variable V, usually continuous, with v being the specific value that it takes. We may thus model the quantity r at the receiver as the observation of a random variable R, with

$$R = A + V \; , \tag{13.3}$$

and we want to estimate the value that the random variable A takes, given that $R = r$. Consequently, we need to add a further processing step to our receiver, in which an estimate of A is obtained.

In the case where the pulse amplitude can be only one of two values, i.e., in the case of binary signaling, finding an estimate of A reduces to deciding, on the basis of the observed value r of R, which of the two possible amplitudes was transmitted. Two common forms of binary signaling in PAM systems are on/off signaling and antipodal signaling. Letting a_1 and a_0 denote the two possible amplitudes (representing for example a binary "one" or "zero"), in on/off signaling we have $a_0 = 0$, $a_1 \neq 0$, whereas in antipodal signaling $a_0 = -a_1 \neq 0$.

Thus, in binary signaling, the required post-processing corresponds to deciding between two alternatives or hypotheses, where the available information may include some prior information along with a measurement r of the single continuous random variable R. (The extension to multiple hypotheses and multiple measurements will be straightforward once the two-hypothesis case is understood.) The hypotheses are listed below:

Hypothesis H_0: the transmitted amplitude A takes the value a_0, so $R = a_0 + V$.

Hypothesis H_1: the transmitted amplitude A takes the value a_1, so $R = a_1 + V$.

Our task now is to decide, given the measurement $R = r$, whether H_0 or H_1 is responsible for the measurement. The next section develops a framework for this sort of hypothesis testing task.

13.2 BINARY HYPOTHESIS TESTING

Our general binary hypothesis testing task is to decide, on the basis of a measurement r of a random variable R, which of two hypotheses — H_0 or H_1 — is responsible for the measurement. We shall indicate these decisions by 'H_0' and 'H_1' respectively (where the quotation marks are intended to suggest the announcement of a decision). An alternative notation is $\widehat{H} = H_0$ and $\widehat{H} = H_1$ respectively, where $\widehat{H}$ denotes our estimate of, or decision on, the hypothesis H.

Suppose H is modeled as a random quantity, and assume we know the *a priori* (i.e., prior) probabilities

$$P(H_0 \text{ is true}) = P(H = H_0) = P(H_0) = p_0 \qquad (13.4)$$

and

$$P(H_1 \text{ is true}) = P(H = H_1) = P(H_1) = p_1 \qquad (13.5)$$

(where the last two equalities in each case simply define streamlined notation that we will be using). We shall also require the conditional densities $f_{R|H}(r|H_0)$ and $f_{R|H}(r|H_1)$ that tell us how the measured variable is distributed under the two respective hypotheses. These conditional densities in effect constitute the relevant specifications of how the measured data relates to the two hypotheses. For example, in the PAM setting, with R defined as in (13.3) and assuming V is independent of A under each hypothesis, these conditional densities are simply

$$f_{R|H}(r|H_0) = f_V(r - a_0) \quad \text{and} \quad f_{R|H}(r|H_1) = f_V(r - a_1) \ . \qquad (13.6)$$

It is natural in many settings, as in the case of digital communication by PAM, to want to minimize the probability of picking the wrong hypothesis, i.e., to choose with minimum probability of error between the hypotheses, given the measurement $R = r$. We will, for most of our discussion of hypothesis testing, focus on this criterion of minimum probability of error.

13.2.1 Deciding with Minimum Probability of Error: The MAP Rule

Consider first how one would choose between H_0 and H_1 with minimum probability of error in the absence of any measurement of R. If we make the choice 'H_0', then we make an error precisely when H_0 does not hold, so the probability of error with this choice is $1 - P(H_0) = 1 - p_0$. Similarly, if we chose 'H_1', then the probability of error is $1 - P(H_1) = 1 - p_1 = p_0$. Thus, for minimum probability of error, we should decide in favor of whichever hypothesis has maximum probability — an intuitively reasonable conclusion. (The preceding reasoning extends in the same way to choosing one from among many hypotheses, and leads to the same conclusion.)

What changes when we aim to choose between H_0 and H_1 with minimum probability of error, knowing that $R = r$? The same reasoning applies as in the preceding paragraph, except that all probabilities now need to be conditioned on the measurement $R = r$. We conclude that to minimize the conditional probability of error, $P(\text{error}|R = r)$, we need to decide in favor of whichever hypothesis has maximum conditional probability, conditioned on the measurement $R = r$. (If there were several random variables for which we had measurements, rather than just the single random variable R, we would simply condition on all the available measurements.) Thus, if $P(H_1|R = r) > P(H_0|R = r)$, we decide '$H_1$', and if $P(H_1|R = r) < P(H_0|R = r)$, we decide '$H_0$'. This may be compactly written as

$$P(H_1|R = r) \underset{\text{`}H_0\text{'}}{\overset{\text{`}H_1\text{'}}{\gtrless}} P(H_0|R = r) \; . \tag{13.7}$$

(If the two conditional probabilities happen to be equal, we get the same conditional probability of error whether we choose 'H_0' or 'H_1'.) The corresponding conditional probability of error is

$$P(\text{error}|R = r) = \min\{1 - P(H_0|R = r), 1 - P(H_1|R = r)\} \; . \tag{13.8}$$

The overall probability of error, P_e, associated with the use of the above decision rule (but before knowing what specific value of R is measured) is obtained by averaging the conditional probability of error in (13.8) over all possible values of r that might be measured, using the PDF $f_R(r)$ as a weighting function. We shall study P_e in more detail shortly.

The conditional probabilities $P(H_0|R = r)$ and $P(H_1|R = r)$ that appear in the expression (13.7) are referred to as the *a posteriori* or posterior probabilities of the hypotheses, to distinguish them from the *a priori* or prior probabilities, $P(H_0)$ and $P(H_1)$. The decision rule in (13.7) is accordingly referred to as the maximum *a posteriori* probability rule, usually abbreviated as the "MAP" rule.

To actually evaluate the posterior probabilities in (13.7), we use Bayes' rule to rewrite them in terms of known quantities, so the decision rule becomes

$$\frac{p_1 f_{R|H}(r|H_1)}{f_R(r)} \underset{\text{`}H_0\text{'}}{\overset{\text{`}H_1\text{'}}{\gtrless}} \frac{p_0 f_{R|H}(r|H_0)}{f_R(r)} , \tag{13.9}$$

under the reasonable assumption that $f_R(r) > 0$, i.e., that the PDF of R is positive at the value r that was actually measured. (In any case, we only need to specify our decision rule at values of r for which $f_R(r) > 0$, because the choices made at other values of r do not affect the overall probability of error, P_e.) Since the denominator is the same and positive on both sides of the above expression, we may further simplify it to

$$p_1 f_{R|H}(r|H_1) \underset{\text{`}H_0\text{'}}{\overset{\text{`}H_1\text{'}}{\gtrless}} p_0 f_{R|H}(r|H_0) . \tag{13.10}$$

This now provides us with an easily visualized and implemented decision rule. We first use the prior probabilities $p_i = P(H_i)$ to scale the PDFs $f_{R|H}(r|H_i)$ that describe how the measured quantity R is distributed under each of the hypotheses. We then decide in favor of the hypothesis associated with whichever scaled PDF is largest at the measured value r. (The preceding description also applies to choosing with minimum probability of error among multiple hypotheses, rather than just two, and given measurements of several associated random variables, rather than just one — the reasoning is identical.)

13.2.2 Understanding P_e: False Alarm, Miss and Detection

The sample space that is relevant to evaluating a decision rule consists of the following four mutually exclusive and collectively exhaustive possibilities: H_i is true and we declare 'H_j', $i, j = 1, 2$. Of the four possible outcomes, the two that represent errors are $(H_0, \text{`}H_1\text{'})$ and $(H_1, \text{`}H_0\text{'})$. Therefore, the probability of error P_e — averaged over all possible values of the measured random variable — is given by

$$\begin{aligned} P_e &= P(H_0, \text{`}H_1\text{'}) + P(H_1, \text{`}H_0\text{'}) \\ &= p_0 P(\text{`}H_1\text{'}|H_0) + p_1 P(\text{`}H_0\text{'}|H_1) . \end{aligned} \tag{13.11}$$

The conditional probability $P(\text{`}H_1\text{'}|H_0)$ is referred to as the conditional probability of a false alarm, and denoted by $P_{FA}$. The conditional probability $P(\text{`}H_0\text{'}|H_1)$ is referred to as the conditional probability of a miss, and denoted by P_M. The word "conditional" is usually omitted from these terms in normal use, but it is important to keep in mind that the probability of a false alarm and the probability of a miss are defined as conditional probabilities, and are furthermore conditioned on different events.

The preceding terminology is historically motivated by the radar context, in which H_1 represents the presence of a target and H_0 the absence of a target. A false alarm then occurs if you declare that a target is present when it actually isn't, and a miss occurs if you declare that a target is absent when it actually isn't. We will also make reference to the conditional probability of detection,

$$P_D = P(\text{'}H_1\text{'}|H_1) \ . \tag{13.12}$$

In the radar context, this is the probability of declaring a target is present when it is actually present. As with P_{FA} and P_M, the word "conditional" is usually omitted in normal use, but it is important to keep in mind that the probability of detection is a conditional probability.

Expressing the probability of error in terms of P_{FA} and P_M, (13.11) becomes

$$P_e = p_0 P_{FA} + p_1 P_M \ . \tag{13.13}$$

Also note that

$$P(\text{'}H_0\text{'}|H_1) + P(\text{'}H_1\text{'}|H_1) = 1 \tag{13.14}$$

or

$$P_M = 1 - P_D \ . \tag{13.15}$$

To explicitly relate P_{FA} and P_M to whatever the corresponding decision rule is, it is helpful to introduce the notion of a decision region in measurement space. In the case of a decision rule based on measurement of a single random variable R, specifying the decision rule corresponds to choosing a range of values D_1 on the real line such that, when the measured value r of R falls in D_1, we declare 'H_1', and when r falls outside D_1 — a region that we shall denote by D_0 — then we declare 'H_0'. This is illustrated in Figure 13.3, for some arbitrary choice of D_1. (There is a direct generalization of this notion to the case where multiple random variables are measured.)

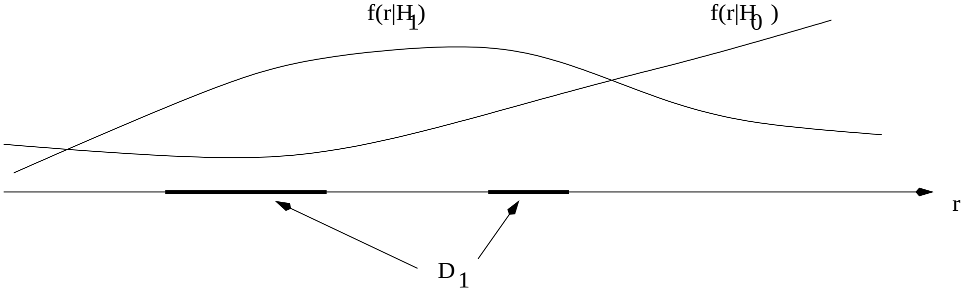

FIGURE 13.3 Decision regions. The choice of D_1 marked here is arbitrary, not the optimal choice for minimum probability of error.

With the preceding definitions, we can write

$$P_{FA} = \int_{D_1} f_{R|H}(r|H_0)dr \tag{13.16}$$

and

$$P_M = \int_{D_0} f_{R|H}(r|H_1)dr \; . \tag{13.17}$$

13.2.3 The Likelihood Ratio Test

Rewriting (13.10), we can state the minimum-P_e decision rule in the form

$$\Lambda(r) = \frac{f_{R|H}(r|H_1)}{f_{R|H}(r|H_0)} \begin{array}{c} \text{`}H_1\text{'} \\ > \\ < \\ \text{`}H_0\text{'} \end{array} \frac{p_0}{p_1} \tag{13.18}$$

or

$$\Lambda(r) \begin{array}{c} \text{`}H_1\text{'} \\ > \\ < \\ \text{`}H_0\text{'} \end{array} \eta \; , \tag{13.19}$$

where $\Lambda(r)$ is referred to as the likelihood ratio, and η is referred to as the threshold. This particular way of writing our decision rule is of interest because other formulations of the binary hypothesis testing problem — with criteria other than minimization of P_e — also often lead to a decision rule that involves comparing the likelihood ratio with a threshold. The only difference is that the threshold is picked differently in these other formulations. We describe two of these alternate formulations — the Neyman-Pearson approach, and minimum risk decisions — in later sections of this chapter.

13.2.4 Other Scenarios

While the above discussion of binary hypothesis testing was introduced in the context of binary PAM, it applies in many other scenarios. For example, in the medical literature, clinical tests are described using a hypothesis testing framework similar to that used here for communication and signal detection problems, with H_0 generally denoting the absence of a medical condition and H_1 its presence. The terminology in the medical context is slightly different, but still suggestive of the intent, as the following examples show:

- P_D is the sensitivity of the clinical test.
- $P(\text{`}H_1\text{'}|H_0)$ is the probability of a false positive (rather than of a false alarm).
- $1 - P_{FA}$ is the specificity of the test.
- $P(H_1)$ is the prevalence of the condition that the test is aimed at.
- $P(H_1\,|\text{`}H_1\text{'})$ is the positive predictive value of the test, and $P(H_0\,|\,\text{`}H_0\text{'})$ is the negative predictive value.

Some easy exploration using Bayes' rule and the above terminology will lead you to recognize how small the positive predictive value of a test can be if the prevalence of the targeted medical condition is low, even if the test is highly sensitive and specific.

Another important context for binary hypothesis testing is in target detection, such as aircraft detection and tracking, in which a radar pulse is transmitted and the decision on the presence or absence of an aircraft is based on the presence or absence of reflected energy.

13.2.5 Neyman-Pearson Detection and Receiver Operating Characteristics

A difficulty with using the minimization of P_e as the decision criterion in many of these other contexts is that it relies heavily on knowing the *a priori* probabilities p_0 and p_1, and in many situations there is little basis for coming up with these numbers. One alternative that often makes sense is to maximize the probability of detection P_D, while keeping P_{FA} below some specified tolerable level. These conditional probabilities are determined by the measurement models under the different hypotheses, and by the decision rule, but not by the probabilities governing the selection of hypotheses. Such a formulation of the hypothesis testing problem again leads to a decision rule that involves comparing the likelihood ratio with a threshold; the only difference now is that the threshold is picked differently in this formulation. This approach is referred to as Neyman-Pearson detection, and is elaborated on below.

Consider a context in which we want to maximize the probability of detection,

$$P_D = P(\text{`}H_1\text{'}|H_1) = \int_{D_1} f_{R|H}(r|H_1)dr \; , \tag{13.20}$$

while keeping the probability of false alarm,

$$P_{FA} = P(\text{`}H_1\text{'}|H_0) = \int_{D_1} f_{R|H}(r|H_0)dr \; , \tag{13.21}$$

below a pre-specified level. (Both integrals are over the decision region D_1, and augmenting D_1 by adding more of the real axis to it will not decrease either probability.) As we show shortly, we can achieve our objective by picking the decision region D_1 to comprise those values of r for which the likelihood ratio $\Lambda(r)$ exceeds a certain threshold η, so

$$\Lambda(r) = \frac{f_{R|H}(r|H_1)}{f_{R|H}(r|H_0)} \begin{array}{c} \text{`}H_1\text{'} \\ > \\ < \\ \text{`}H_0\text{'} \end{array} \eta \; . \tag{13.22}$$

The threshold η is picked to provide the largest possible P_D while ensuring that P_{FA} is not larger than the pre-specified level. The smaller the η, the larger the decision region D_1 and the value of P_D become, but the larger P_{FA} grows as well, so one would pick the smallest η that is consistent with the given bound on P_{FA}.

To understand why the decision rule in this setting takes the form of (13.22), note that our objective is to include in D_1 values of r that contribute as much as possible to the integral that defines P_D, and as little as possible to the integral that defines P_{FA}. If we start with a high value of the threshold η, we will be including in D_1 those r for which $\Lambda(r)$ is large, and therefore where the contribution to P_D is relatively large compared to the contribution to P_{FA}. Moving η lower, we increase both P_D and P_{FA}, but the rate of increase of P_D drops, while the rate of increase of P_{FA} rises. These increases in P_D and P_{FA} may not be continuous in η. (Reducing η from infinitesimally above some value $\overline{\eta}$ to infinitesimally below this value will give rise to a finite upward jump in both P_D and P_{FA} if $f_{R|H}(r|H_1) = \overline{\eta}\, f_{R|H}(r|H_0)$ throughout some interval of r where both these PDFs are positive.) Typically, though, the variation of P_D and P_{FA} with η is indeed continuous, so as η is lowered we reach a point where the specified bound on P_{FA} is attained, or $P_D = 1$ is reached. This is the value of η used in the Neyman-Pearson test. (In the rare situation where P_{FA} jumps discontinuously from a value below its tolerable level to one above its tolerable level as η is lowered through some value $\overline{\eta}$, it turns out that a randomized decision rule allows one to come right up to the tolerable P_{FA} level, and ! thereby maximize P_D. A case like this is explored in a problem at the end of this chapter.)

The following argument shows in a little more detail, though still informally, why the Neyman-Pearson criterion is equivalent to a likeliood ratio test. If the decision region D_1 is optimal for the Neyman-Pearson criterion, then any change in D_1 that keeps P_{FA} the same cannot lead to an improvement in P_D. So suppose we take a infinitesimal segment of width dr at a point r in the optimal D_1 region and convert it to be part of D_0. In order to keep P_{FA} unchanged, we must correspondingly take an infinitesimal segment of width dr' at an arbitrary point r' in the optimal D_0 region, and convert it to be a part of D_1.

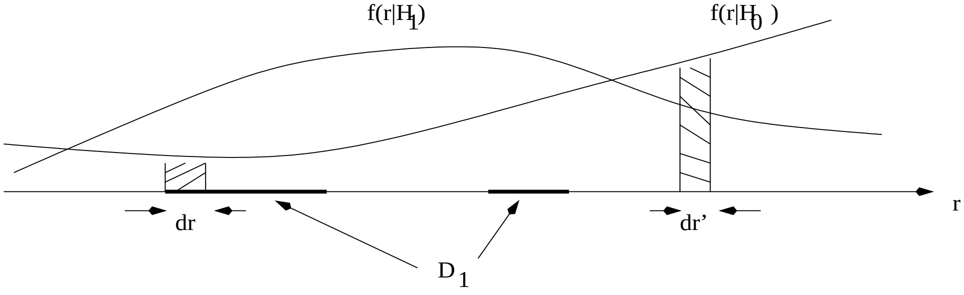

FIGURE 13.4 Illustrating the construction used in deriving the likelihood ratio test for the Neyman-Pearson criterion.

The requirement that P_{FA} be unchanged then imposes the condition

$$f_{R|H}(r'|H_0)\, dr' = f_{R|H}(r|H_0)\, dr \;, \tag{13.23}$$

while the requirement that the new P_D not be larger than the old implies that

$$f_{R|H}(r'|H_1)\,dr' \leq f_{R|H}(r|H_1)\,dr\ . \tag{13.24}$$

Combining (13.23) and (13.24), we find

$$\Lambda(r') \leq \Lambda(r)\ . \tag{13.25}$$

What (13.25) shows is that the likelihood ratio cannot be less inside D_1 than it is in D_0. We can therefore conclude that the optimum solution to the Neyman-Pearson formulation is in fact based on a threshold test on the likelihood ratio:

$$\Lambda(r) = \frac{f_{R|H}(r|H_1)}{f_{R|H}(r|H_0)} \begin{array}{c} \text{‘}H_1\text{’} \\ > \\ < \\ \text{‘}H_0\text{’} \end{array} \eta\ , \tag{13.26}$$

where the threshold η is picked to obtain the largest possible P_D while ensuring that P_{FA} is not larger than the pre-specified bound.

The above derivation has made various implicit assumptions. However, our purpose is only to convey the essence of how one arrives at a likelihood ratio test in this case.

Receiver Operating Characteristic. In considering which value of P_{FA} to choose as a bound in the Neyman-Pearson test, it is often useful to look at a curve of P_D versus P_{FA} as the parameter η is varied. This is referred to as the Receiver Operating Characteristic (ROC). More generally, such an ROC can be defined for any decision rule that causes P_D to be uniquely fixed, once P_{FA} is specified. The ROC can be used to identify whether, for instance, modifying the variable parameters in a given test to permit a slightly higher P_{FA} results in a significantly higher P_D. The ROC can also be used to compare different tests.

EXAMPLE 13.1 Detection and ROC for Signal in Gaussian Noise

Consider a scenario in which a radar pulse is emitted from a ground station. If an aircraft is located in the propagation path, a reflected pulse will travel back towards the radar station. We assume that the received signal will then consist of noise alone if no aircraft is present, and noise plus the reflected pulse if an aircraft is present. The processing of the received signal results in a number that we model as the realization of a random variable R. If an aircraft is not present, then $R = W$, where W is a random variable denoting the result of processing just the noise. If an aircraft is present, then $R = s + W$, where the constant s is due to processing of the reflected pulse, and is assumed here to be a known value. We thus have the following two hypotheses:

$$H_0 :\quad R = W \tag{13.27}$$

$$H_1 :\quad R = s + W\ . \tag{13.28}$$

Assume that the additive noise term W is Gaussian with zero mean and unit variance, i.e.,

$$f_W(w) = \frac{1}{\sqrt{2\pi}} e^{-w^2/2}. \tag{13.29}$$

Consequently,

$$f_{R|H}(r|H_0) = \frac{1}{\sqrt{2\pi}} e^{-r^2/2} \tag{13.30}$$

$$f_{R|H}(r|H_1) = \frac{1}{\sqrt{2\pi}} e^{-(r-s)^2/2} \ . \tag{13.31}$$

The likelihood ratio as defined in (13.18) is then

$$\begin{aligned} \Lambda(r) &= \exp\Big[-\frac{(r-s)^2}{2} + \frac{r^2}{2}\Big] \\ &= \exp\Big[sr - \frac{s^2}{2}\Big] \ . \end{aligned} \tag{13.32}$$

For detection with minimum probability of error, the decision rule corresponds to evaluating this likelihood ratio at the received value r, and comparing the result against the threshold p_0/p_1, as stated in (13.18):

$$\exp\Big[sr - \frac{s^2}{2}\Big] \ \underset{\text{‘}H_0\text{’}}{\overset{\text{‘}H_1\text{’}}{\gtrless}} \ \eta = \frac{p_0}{p_1} \tag{13.33}$$

It is interesting and important to note that, for this case, the threshold test on the likelihood ratio can be rewritten as a threshold test on the received value r. Specifically, (13.33) can equivalently be expressed as

$$[sr - \frac{s^2}{2}\Big] \ \underset{\text{‘}H_0\text{’}}{\overset{\text{‘}H_1\text{’}}{\gtrless}} \ \ln\eta \ , \tag{13.34}$$

or, if $s > 0$,

$$r \ \underset{\text{‘}H_0\text{’}}{\overset{\text{‘}H_1\text{’}}{\gtrless}} \ \frac{1}{s}\Big[\frac{s^2}{2} + \ln\eta\Big] = \gamma \ , \tag{13.35}$$

where γ denotes the threshold on r. (If $s < 0$, the inequalities in (13.35) are simply reversed.) For example, if both hypotheses are equally likely *a priori*, so that $p_0 = p_1$, then $\ln\eta = 0$ and the decision rule for minimum probability of error when $s > 0$ is simply

$$r \ \underset{\text{‘}H_0\text{’}}{\overset{\text{‘}H_1\text{’}}{\gtrless}} \ \frac{s}{2} = \gamma \ . \tag{13.36}$$

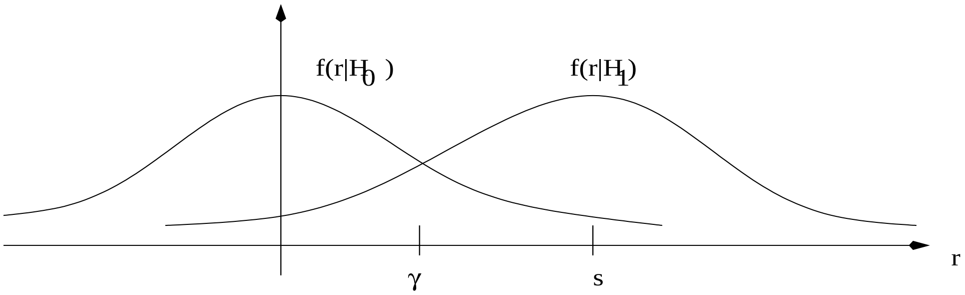

FIGURE 13.5 Threshold γ on measured value r.

The situation is represented in Figure 13.5.

The receiver operating characteristic displays P_D versus P_{FA} as η is varied, and is sketched in Figure 13.6.

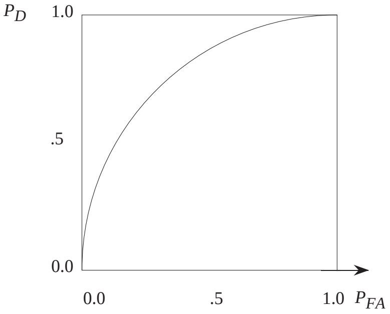

FIGURE 13.6 Receiver operating characteristic.

In a more general setting than the Gaussian case in Example 13.1, a threshold test on the likelihood ratio would not simply translate to a threshold test on the measurement r. Nevertheless, we could still decide to use a simple threshold test on r as our decision rule, and then generate and evaluate the associated receiver operating characteristic.

13.3 MINIMUM RISK DECISIONS

This section briefly describes a decision criterion, called minimum risk, that includes minimum probability of error as a special case, and that in the binary case again leads to a likelihood ratio test. We describe it for the general case of M hypotheses.

Let the available measurement be the value r of the random variable R (the same development holds if we have measurements of several random variables). Suppose we associate a cost c_{ij} with each combination of model H_j and decision 'H_i' for $0 \leq i, j \leq M-1$, reflecting the costs of actions and consequences that follow from this combination of model and decision. Our objective now is to pick whichever decision has minimum expected cost, or minimum "risk", given the measurement.

The expected cost of deciding 'H_i', conditioned on $R = r$, is given by

$$E[\text{Cost}|R = r, `H_i'] = \sum_{j=0}^{M-1} c_{ij} P(H_j|R = r, `H_i') = \sum_{j=0}^{M-1} c_{ij} P(H_j|R = r) \ , \quad (13.37)$$

where the last equality is a consequence of the fact that, given the received measurement $R = r$, the output of the decision rule conveys no additional information about which hypothesis actually holds. The next step is to compare these conditional expected costs for all i, and decide in favor of the hypothesis with minimum conditional expected cost. Specifying our decision for each possible r, we obtain the decision rule that minimizes the overall expected cost or risk.

[It is in this setting that hypothesis testing comes closest to the estimation problems for continuous random variables that we considered in our chapter on minimum mean-square-error estimation. We noted there that a variety of such estimation problems can be formulated in terms of minimizing an expected cost function. Establishing an estimate for a random variable is like carrying out a hypothesis test for a continuum of numerically specified hypotheses (rather than just M general hypotheses), with a cost function that penalizes some measure of the numerical distance between the actual hypothesis and the one we decide on.]

Note that if $c_{ii} = 0$ for all i and if $c_{ij} = 1$ for $j \neq i$, so we penalize all errors equally, then the conditional expected cost in (13.37) becomes

$$E[\text{Cost}|R = r, `H_i'] = \sum_{j \neq i} P(H_j|r) = 1 - P(H_i|r) \ . \quad (13.38)$$

This conditional expected cost is thus precisely the conditional probability of error associated with deciding 'H_i', conditioned on $R = r$. The right side of the equation then shows that to minimize this conditional probability of error we should decide in favor of the hypothesis with largest conditional probability. In other words, with this choice of costs, the risk (when the expectation is taken over all possible values of r) is exactly the probability of error P_e, and the optimum decision rule for minimizing this criterion is again seen to be the MAP rule.

Using Bayes' rule in (13.37) and noting that $f_R(r)$ — assumed positive — is common to all the quantities involved in our comparison, we see that an equivalent but more directly implementable procedure is to pick the hypothesis for which

$$\sum_{j=0}^{M-1} c_{ij} f(r|H_j) P(H_j) \quad (13.39)$$

is minimum. In the case of two hypotheses, and assuming $c_{01} > c_{11}$, it is easy to see that the decision rule based on (13.39) can be rewritten as

$$\Lambda(r) = \frac{f(r|H_1)}{f(r|H_0)} \begin{array}{c} \text{`}H_1\text{'} \\ > \\ < \\ \text{`}H_0\text{'} \end{array} \frac{P(H_0)(c_{10} - c_{00})}{P(H_1)(c_{01} - c_{11})} = \eta \ , \tag{13.40}$$

where $\Lambda(r)$ denotes the likelihood ratio, and η is the threshold. We have therefore again arrived at a decision rule that involves comparing a likelihood ratio with a threshold. If $c_{ii} = 0$ for $i = 0, 1$ and if $c_{ij} = 1$ for $j \neq i$, then we obtain the threshold associated with the MAP decision rule for minimum P_e, as expected.

The trouble with the above minimum risk approach to classification, and with the minimum error probability formulation that we have examined a few times already, is the requirement that the prior probabilities $P(H_i)$ be known.

It is often unrealistic to assume that prior probabilities are known, so we are led to consider alternative criteria. Most important among these alternatives is the Neyman-Pearson approach treated earlier, where the decision is based on the conditional probabilities P_D and P_{FA}, thereby avoiding the need for prior probabilities on the hypotheses.

13.4 HYPOTHESIS TESTING IN CODED DIGITAL COMMUNICATION

In our discussion of PAM earlier in this chapter, we considered binary hypothesis testing on a single received pulse. In modern communication systems, an alphabet of symbols may be transmitted, with each symbol encoded into a binary sequence of “ones” and “zeroes”. Consequently, in addition to making a binary decision on each received pulse, we may need to further decode a string of bits to make our best judgement of the transmitted symbol, and perhaps yet further processing to decide on the sequence of symbols that constitutes the entire message. It would in principle be better to take all the raw measurements and then make optimal decisions about the entire sequence of symbols that was transmitted, but this would be a hugely more complex task. In practice, therefore, the task is commonly broken down into three stages, as here, with locally optimal decisions made at the single-pulse level to decode sequences of “ones” and “zeros”, then further decisions made to decode at the symbol level, and still further decisions made at the symbol sequence level. In this section we illustrate the second of these decoding stages.

For concreteness, we center our discussion on the system in Figure 13.7. Suppose the transmitter randomly selects for transmission one of four possible symbols, which we label A, B, C and D. The probabilities with which these are selected will be denoted by $P(A)$, $P(B)$, $P(C)$ and $P(D)$ respectively. Whatever symbol the transmitter selects is now coded appropriately for transmission over the binary channel. The coding adds some redundancy to provide a basis for error correction at the receiver, in order to combat errors introduced by channel noise that may corrupt the individual bits. The resulting signal is then sent to the receiver. After

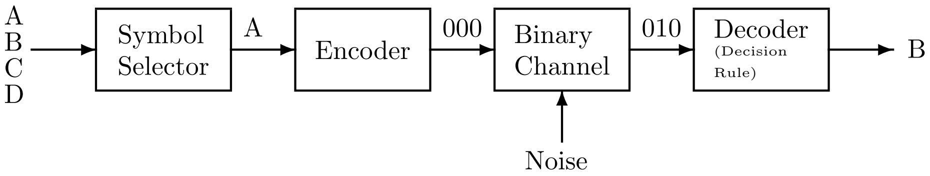

FIGURE 13.7 Communication over a binary channel.

the receiver decodes the received pulses, attempting to correct for channel noise in the process, it has to arrive at a decision as to which symbol was transmitted.

A natural criterion for measuring the performance of the receiver, with whatever decision process or decision rule it applies, is again the probability of error, P_e. It is natural, in a communications setting, to want minimum probability of error, and this is the criterion we adopt.

In the development below, rather than simply invoking the MAP rule we derived earlier, we repeat in this higher-level setting the line of reasoning that led to the MAP rule. We do this partly because there are some differences from what we considered earlier: we now have multiple hypotheses (four in our example), not just a pair of hypotheses; and the measured quantity is a discrete random symbol (more exactly, the received and possibly noise corrupted binary code for a transmitted symbol), rather than a continuous random variable. However, it will be clear that the problem here is not fundamentally different or harder.

13.4.1 Optimal a priori Decision

Consider, first of all, what the minimum-probability-of-error decision rule would be for the receiver if the channel was down, i.e., if the receiver had to decide on the transmitted signal without the benefit of any received signal, using only on *a priori* information. If the receiver guesses that the transmitter selected the symbol A, then the receiver is correct if A was indeed the transmitted symbol, and the receiver has made an error if A was not the transmitted symbol. Hence the receiver's probability of error with this choice is $1-P(A)$. Similar reasoning applies for the other symbols. So the minimum-probability-of-error decision rule for the receiver is to decide in favor of whichever symbol has maximum probability. This seems quite obvious for this simple case, and the general case (i.e., with the channel functioning) is not really any harder. We turn now to this general case, where the receiver actually receives the result of sending the transmitted signal through the noisy channel.

13.4.2 The Transmission Model

Let us model the channel as a binary channel, which accepts 1's and 0's from the transmitter, and delivers 1's and 0's to the receiver. Suppose that because of the noise in the channel there is a probability $p > 0$ that a transmitted 1 is received as a 0, and that a transmitted 0 is received as a 1. Because the probability is the same for both types of errors, this binary channel is called symmetric (we could treat the non-symmetric case as easily, apart from some increased notational burden). Implicit in our definition of this channel is the assumption that it is memoryless, i.e., its characteristics during any particular transmission slot are independent of what has been transmitted in other time slots. The channel is also assumed time-invariant, i.e., its characteristics do not vary with time.

Given such a channel, the transmitter needs to code the selected symbol into binary form. Suppose the transmitter uses 3 bits to code each symbol, as follows:

$$A : 000\ , \quad B : 011\ , \quad C : 101\ , \quad D : 110\ . \tag{13.41}$$

Because of the finite probability of bit-errors introduced by the channel, the received sequence for any of these transmissions could be any 3-bit binary number:

$$\begin{aligned} &R_0 = 000\ , \quad R_1 = 001\ , \quad R_2 = 010\ , \quad R_3 = 011\ , \\ &R_4 = 100\ , \quad R_5 = 101\ , \quad R_6 = 110\ , \quad R_7 = 111\ . \end{aligned} \tag{13.42}$$

The redundancy introduced by using 3 bits — rather than the 2 bits that would suffice to communicate our set of four symbols — is intended to provide some protection against channel noise. Notice that with our particular 3-bits/symbol code, a single bit-error would be recognized at the receiver as an error, because it would result in an invalid codeword. It takes two bit-errors (which are rarer than single bit-errors) to convert any valid codeword into another valid one, and thereby elude recognition of the error by the receiver.

There are now various probabilities that it might potentially be of interest to evaluate, such as:

- $P(R_1 \mid D)$, the probability that R_1 is received, given that D was sent;
- $P(D \mid R_1)$, the probability that D was sent, given that R_1 was received — this is the *a posteriori* probability of D, in contrast to $P(D)$, which is the *a priori* probability of D;
- $P(D, R_1)$, the probability that D is sent and R_1 is received;
- $P(R_1)$, the probability that R_1 is received.

The sample space of our probabilistic experiment can be described by Table 13.1, which contains an entry corresponding to every possible combination of transmitted symbol and received sequence. In the jth row of column A, we enter the probability $P(A, R_j)$ that A was transmitted and R_j received, and similarly for columns B, C, and D. The simplest way to actually compute this probability is by recognizing that $P(A, R_j) = P(R_j|A)P(A)$; the characterization of the channel permits computation of $P(R_j|A)$, while the characterization of the information source at the transmitter yields the prior probability $P(A)$. Note that we can also write $P(A, R_j) = P(A|R_j)P(R_j)$. Examples of these three ways of writing the probabilities of the outcomes of our experiment are shown in the table.

13.4.3 Optimal a posteriori Decision

We now want to design the decision rule for the receiver, i.e., the rule by which it decides or hypothesizes what symbol was transmitted, after the reception of a particular sequence. We would like to do this in such a way that the probability of error, P_e, is minimized.

Since a decision rule in our example selects one of the four possible symbols (or hypotheses), namely A, B, C, or D, for each possible R_j, it can be represented in Table 13.1 by selecting one (and only one) entry in each row; we shall mark the selected entry by a box. For instance, a particular decision rule may declare D to be the transmitted signal whenever it receives R_4; this is indicated on the table by putting a box around the entry in row R_4, column D, as shown. Each possible decision rule is therefore associated with a table of the preceding form, with precisely one entry boxed in each row.

Now, for a given decision rule, the probability of being correct is the sum of the probabilities in all the boxed entries, because this sum gives the total probability that the decision rule declares in favor of the same symbol that was transmitted. The probability of error, P_e, is therefore 1 minus the probability of being correct.

It follows that to specify the decision rule for minimum probability of error or maximum probability of being correct, we must pick in each row the box that has the maximum entry. (If more than one entry has the maximum value, we are free to pick one of these arbitrarily — P_e is not affected by which of these we pick.) For row R_j in Table 13.1, we should pick for the optimum decision rule the symbol for which we maximize

$$\begin{aligned} P(\text{symbol}, R_j) &= P(R_j|\,\text{symbol})P(\text{symbol}) \\ &= P(\text{symbol}\,|R_j)P(R_j)\ . \end{aligned} \tag{13.43}$$

Table 13.2 displays some examples of the required computation in a particular numerical case. The computation in this example is carried out according to the prescription on the right side in the first of the above pair of equations. As noted earlier, this is generally the form that yields the most direct computation in practice, because the characterization of the channel usually permits direct computation of $P(R_j|\,\text{symbol})$, while the characterization of the information source at the transmitter yields the prior probabilities $P(\text{symbol})$.

The right side of the second equation in (13.43) permits a nice, intuitive interpretation of what the optimum decision rule does. Since our comparison is being done across the row, for a given R_j the term $P(R_j)$ in the second equation stays the

	$A : 000$	$B : 011$	$C : 101$	$D : 110$
$R_0 = 000$	$P(A, R_0)$	$P(B, R_0)$ $= P(R_0 \vert B)P(B)$ $= p^2(1-p)P(B)$	$P(C, R_0)$ $= P(C \vert R_0)P(R_0)$	$P(D, R_0)$
$R_1 = 001$				
$R_2 = 010$				
$R_3 = 011$				
$R_4 = 100$	$P(A, R_4)$	$P(B, R_4)$	$P(C, R_4)$	$\boxed{P(D, R_4)}$
$R_5 = 101$				
$R_6 = 110$				
$R_7 = 111$				

TABLE 13.1 Each entry corresponds to a transmitted symbol and a received sequence.

same, so actually all that we need to compare are the *a posteriori* probabilities, $P(\text{symbol}\,|R_j)$, i.e. the probabilities of the various symbols, given the data. The optimum decision rule therefore picks the symbol with the maximum *a posteriori* probability. This is again the MAP decision rule that we derived previously in the binary hypothesis case. To summarize the important result we have arrived at here, and which we shall encounter again in more elaborate hypothesis testing contexts:

For minimum error probability P_e, decide in favor of the choice that has maximum *a posteriori* probability, i.e., the choice whose probability, conditioned on the available data, is maximum.

Note that the only difference from the minimum-P_e *a priori* decision rule we arrived at earlier, for the case where the channel was down, is the computation now has to involve conditional or *a posteriori* probabilities — conditioned on the received information — rather than the *a priori* probabilities. The receiver still decides in favor of the most probable choice, but now incorporating (i.e., conditioning on) the received information.

	000 A	011 B	101 C	110 D	Decision
R_0 000					
R_1 001					
R_2 010	$\left(\frac{3}{4}\right)^2 \frac{1}{4}\frac{1}{2}$	$\left(\frac{3}{4}\right)^2 \frac{1}{4}\frac{1}{4}$	$\left(\frac{1}{4}\right)^3 \frac{1}{8}$	$\left(\frac{3}{4}\right)^2 \frac{1}{4}\frac{1}{8}$	'A'
R_3 011					
R_4 100					
R_5 101					
R_6 110	$\left(\frac{1}{4}\right)^2 \frac{3}{4}\frac{1}{2}$	$\left(\frac{1}{4}\right)^2 \frac{3}{4}\frac{1}{4}$	$\left(\frac{1}{4}\right)^2 \frac{3}{4}\frac{1}{8}$	$\left(\frac{3}{4}\right)^3 \frac{1}{8}$	'D'
R_7 111					

TABLE 13.2 Designing the optimal decision rule, with $P(A) = \frac{1}{2}$, $P(B) = \frac{1}{4}$, $P(C) = \frac{1}{8}$, $P(D) = \frac{1}{8}$, $p = \frac{1}{4}$. The MAP rule chooses the symbol that maximizes the *a posteriori* probability, $P(\text{symbol} \mid \text{data})$.

6.011 Introduction to Communication, Control, and Signal Processing
Spring 2010