Chapter 6

Detection Theory

In this chapter we start our investigation of statistical detection theory, also referred to as hypothesis testing or sometimes decision theory. The fundamental problem in statistical detection theory is summarized as follows: In a probability experiment, one and only one of several possible events has happened. After collecting observations with distributions that depend on which event happened, make a decision as to which one of the events actually happened. To illustrate this, consider the following example:

Example 6.1
A sonar system transmits pressure pulses into the water in a given direction, hoping to determine whether a submarine is present in that direction or not. The pulses propagate through the water, and interact with background as well as with a submarine if it is present. The sonar receiver listens for echoes, which may come from the submarine, as well as from background such as ocean floor features, large sea mammals, school of fish, etc. The receiver collects the echoes, and must decide whether there is a submarine present or not based on the received signal.

Note the key components of this problem. There are two possible events, corresponding to many different outcomes in the sample space: the event where a submarine is present in the direction of the sonar pulses, and the event where the submarine is absent. These events are disjoint, and in the terminology of probability events, collectively exhaustive: one of the two events must happen. We collect a measurement, which is a random variable that is a function of the outcome in the experiment. Based on the observed measurement, we must make a decision as to which one of the two possible events is "best to choose."

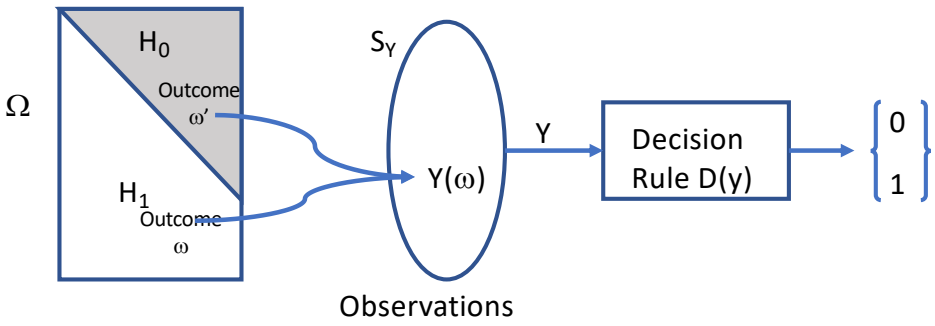


Figure 6.1: Detection problem components.

A general model of this process is shown in Figure 6.1. There are two possible events in the sample space Ω, each of which represents many outcomes. Each of these events is called a *hypothesis*. We use a measurement instrument that collects a random variable Y. Based on the measurement observation $Y = y$, we must design a rule to decide which is the correct hypothesis.

From Figure 6.1 we see that we will need three components in our model:

1. A model of generation processes that creates H_0, H_1.
2. A model of the observation process that generates the observation $Y = y$.
3. A decision rule $D(y)$ that maps each possible observation value y to an associated decision.

In general, the first two elements are set by the experiment or the restrictions of the physical data gathering situation, and we need to model them, but we don't control their design. For example, if we are trying to decide whether an area in a breast cancer mammogram is cancerous or not, the true state of that area (cancerous or not) is selected by processes outside of our control. The measurement instrument (the X-Ray imager) is a physical sensor that generates noisy images depending on whether the area is cancerous or not.

We want to avoid generating a complete description of the probability space $(\Omega, \mathcal{E}, \mathbb{P})$ to model the relationship of the observations Y and the event hypotheses H_0, H_1. Assume Y is a discrete random variable. Using the Law of Total Probability yields

$$\begin{aligned}\mathbb{P}[\{Y=y\}] &= \mathbb{P}[\{Y=y\} \cap H_0] + \mathbb{P}[\{Y=y\} \cap H_1] = \mathbb{P}[\{Y=y\}|H_0]\mathbb{P}[H_0] + \mathbb{P}[\{Y=y\}|H_1]\mathbb{P}[H_1] \\ &= P_{Y|H_0}(y)\mathbb{P}[H_0] + P_{Y|H_1}(y)\mathbb{P}[H_1]\end{aligned}$$

This indicates the components of how we model the detection problem:

1. A model of generation processes that creates H_0, H_1: $\mathbb{P}[H_1], \mathbb{P}[H_0]$.
2. A model of the observation process that generates the observation $Y = y$: $\mathbb{P}[\{Y = y\}|H_0], \mathbb{P}[\{Y = y\}|H_1]$.

This is a compact, probabilistic description that represents the detection problem. Based on this model, we design a decision rule that maps the possible measurement values into a decision. When there are only two possible hypotheses H_0, H_1, this decision rule corresponds to a partition of the space of possible observations into two regions: the region where the decision will be H_1, and the region where the decision will be H_0, as illustrated in Figure 6.2.

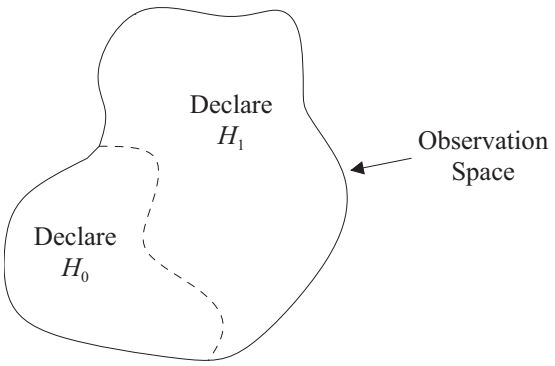


Figure 6.2: Illustration of a decision rule as a partition of the observation space into disjoint regions, illustrated here for the case of two possibilities.

We first discuss in detail the case that arises when there are only two possible hypotheses, termed binary hypothesis testing. Subsequently, we discuss the more general case of M hypotheses, for $M > 2$.

6.1 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. Let's examine a few motivating applications before proceeding to more detailed developments.

Example 6.2 (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.

Example 6.3 (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 6.4 (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.

6.1.1 Detection model

The detection problem is set in a probability space $(\Omega, \mathcal{E}, \mathbb{P})$, which we model in a very abbreviated way. Assume there are only two hypothesis, denoted as H_0 and H_1, which are events in $\mathcal{E}$ are events in the which are mutually disjoint, and collectively exhaustive ($H_0 \cup H_1 = \Omega$). We know $\mathbb{P}[H_0], \mathbb{P}[H_1]$. The figure on the right illustrates the events H_0, H_1 in the sample space, representing a partition of Ω.

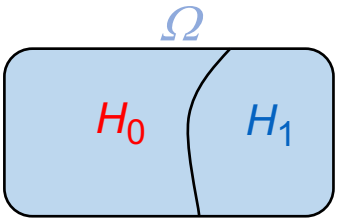


Figure 6.3: Events H_0, H_1.

Observation model: The measurement is a random variable Y defined on $(\Omega, \mathcal{E}, \mathbb{P})$. Y can be either discrete or continuous. For discrete Y, we model the measurement using a pair of conditional probability mass functions $P_{Y|H_1}(y), P_{Y|H_0}(y)$. For continuous Y, we model the measurement in terms of a pair of conditional probability density functions $f_{Y|H_1}(y), f_{Y|H_0}(y)$. These conditional probability functions are known, and are referred to as the **likelihoods** of the measurement $Y = y$ given the different hypotheses.

The figure on the right illustrates the observation model. Note that outcomes in H_0 and outcomes in H_1 can map to the same observation $Y = y$. However, it may be more likely to occur under one of those two hypotheses, as determined by the likelihoods $P_{Y|H_1}(y), P_{Y|H_0}(y)$ or $f_{Y|H_1}(y), f_{Y|H_0}(y)$. These likelihoods will influence which decisions to make.

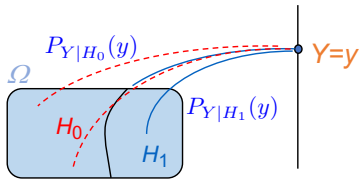


Figure 6.4: Likelihoods $P_{Y|H_1}(y), P_{Y|H_0}(y)$.

Decision rule: A decision rule is a function $U = D(Y)$ of the random variable Y, that maps Y into a decision $U \in \{0, 1\}$. The decision $D(y) = 0$ corresponds to deciding that H_0 is the selected hypothesis when the observation is $Y = y$, and $D(y) = 1$ indicates that H_1 is the selected hypothesis for $Y = y$. $U = D(Y)$ is a discrete random variable, mapping the range R_Y into two possible values. The sets $\{y \in R_Y : D(y) = 0\}$ and $\{y \in R_Y : D(y) = 1\}$ form a partition of R_Y, because $D(\cdot)$ is a function defined everywhere on R_Y. This is illustrated in Figure 6.2.

The decision rule is the solution we design for the detection problem. To do proper design, we select the decision rule on the basis of how good its performance will be.

One way to measure performance is in terms of the errors made by the decision rule. Specifically, when H_0 is true, and generates a measurement $Y = y$ such that $D(y) = 1$, the decision rule has made an error. The figure on the right illustrates the two types of error that the decision rule can make. When H_0 is the event that generates. measurement $Y = y$, and the decision rule selects $D(y) = 1$, we call this a **false alarm**. This terminology dates back to early detection problems such as detecting aircraft using radar, where H_0 was the hypothesis that no airplanes were present. Similarly, when the measurement y is generated by H_1, and $D(y)$ is such that $D(y) = 0$, we refer to this as a **missed detection.**

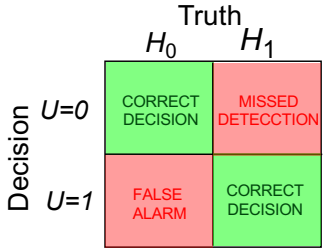


Figure 6.5: Types of Detection Errors.

Given a detection rule $U = D(Y)$, we can compute the probability of a missed detection using the likelihood $P_{Y|H_1}(y)$ if Y is discrete or $f_{Y|H_1}(h)$ if Y is continuous. Denote by A_0 the subset of the range of Y where $D(y) = 0$: $A_0 = \{y \in R_Y : D(y) = 0\}$. Then, the probability of a missed detection is

$$P_{MD} \equiv \mathbb{P}[y \in A_0 | H_1] = \begin{cases} \sum_{y \in A_0} P_{Y|H_1}(y) & Y \text{ is a discrete random variable,} \\ \int_{y \in A_0} f_{Y|H_1}(y)\, dy & Y \text{ is a continuous random variable.} \end{cases}$$

Thus, P_{MD} is the probability of making an erroneous decision when H_0 is true.

Similarly, let $A_1 = \{y \in R_Y : D(y) = 1\}$. Then, $A_0 \cup A_1 = R_Y$, the range of possible values of Y. The probability of a false alarm is computed using the likelihood $P_{Y|H_0}(y)$ if Y is discrete or $f_{Y|H_0}(h)$ if Y is continuous as follows:

$$P_{FA} \equiv \mathbb{P}[y \in A_1 | H_0] = \begin{cases} \sum_{y \in A_1} P_{Y|H_0}(y) & Y \text{ is a discrete random variable,} \\ \int_{y \in A_1} f_{Y|H_0}(y)\, dy & Y \text{ is a continuous random variable.} \end{cases}$$

P_{FA} is the probability of making an erroneous decision when H_1 is true.

Note that P_{FA}, P_{MD} are conditional statistics. If we know $\mathbb{P}[H_0], \mathbb{P}[H_1]$, we can compute unconditional statistics such as the average probability of error using the Law of Total Probability, as:

$$P_e \equiv \mathbb{P}[\text{Error}] = \mathbb{P}[\text{Error}|H_0]\mathbb{P}[H_0] + \mathbb{P}[\text{Error}|H_1]\mathbb{P}[H_1] = P_{FA}\mathbb{P}[H_0] + P_{MD}\mathbb{P}[H_1].$$

We can now use these performance measures to define criteria for selecting a decision rule. We describe different approaches for designing decision rules next.

6.2 Maximum Likelihood Detection

The most common approach for designing a decision rule is known as **maximum likelihood detection.** Assume that Y is a discrete random variable. Given a measurement y, we compute the likelihood of this measurement under each hypothesis, using $P_{Y|H_0}(y)$ and $P_{Y|H_1}(y)$. The maximum likelihood (ML) decision selects the hypothesis that has the largest likelihood for that measurement. That is,

$$D^{ML}(y) = \begin{cases} 1, & P_{Y|H_1}(y) \geq P_{Y|H_0}(y), \\ 0, & P_{Y|H_1}(y) < P_{Y|H_0}(y). \end{cases}$$

We break ties arbitrarily, so we assign a tie to 1.

The maximum likelihood method for detection and estimation was developed by the statistician R. A. Fisher in the early 20th century, although some limited results appeared earlier.

Example 6.5
Assume we have a coin, which may be biased so that the probability of obtaining heads is 0.6. Hypothesis H_1 is that the coin has probability of heads = 0.6. Hypothesis H_0 is that the coin is unbiased, so the probability of heads = 0.5. To detect whether the coin is biased or not, we conduct an experiment, where we flip the coin independently 5 times, and count the number of heads that appear in the experiment. Thus, the measurement in the experiment, Y, is the number of heads in five coin flips.

Y is a discrete random variable, with $R_Y = \{0, 1, 2, 3, 4, 5\}$. The above description lets us describe the likelihood functions: $P_{Y|H_0}(y)$ is the probability mass function of a Binomial(5,0.5) random variable, and $P_{Y|H_1}(y)$ is the probability mass function of a Binomial(5,0.6) random variable. In this case, the range R_Y is small, so we can enumerate the two probability mass functions, and compare their values for each $y \in R_Y$, as shown in the table below.

Y:	0	1	2	3	4	5
$P_{Y\|H_1}$	0.01024	0.0768	0.2304	**0.3456**	**0.2592**	**0.07776**
$P_{Y\|H_0}$	**0.03125**	**0.15625**	**0.3125**	0.3125	0.15625	0.03125

To compute the maximum likelihood decision, we compare the numbers in each column, and pick the larger of the two numbers. In the table above, we have highlighted the larger number in bold and magenta color. Thus we see that the maximum likelihood decision rule becomes:

$$D^{ML}(y) = \begin{cases} 1, & y = 3, 4, 5, \\ 0, & y = 0, 1, 2. \end{cases}$$

The decision agrees with intuition: a larger count of heads suggests the coin is more likely to be unbalanced, whereas a smaller count of heads indicates the coin is more likely to be balanced.

What is the performance of the maximum likelihood decision rule? Let's compute the probability of missed detection. As discussed above, this is the probability that, when H_1 is the correct hypothesis, we get a value y where the decision $D^{ML}(y)$ is 0. Therefore,

$$P_{MD} = \mathbb{P}[\{y = 0, 1, 2\}|H_1] = P_{Y|H_1}(0) + P_{Y|H_1}(1) + P_{Y|H_1}(2) = 0.31744.$$

Similarly, the probability of false alarm is the probability that, when H_0 is the correct hypothesis, we get a measurement $Y = y$ where $D^{ML}(y) = 1$. Then,

$$P_{FA} = \mathbb{P}[\{y = 3, 4, 5\}|H_0] = P_{Y|H_0}(3) + P_{Y|H_0}(4) + P_{Y|H_0}(5) = 0.5$$

Assuming that $\mathbb{P}[H_0] = \mathbb{P}[H_1] = 0.5$, we can compute the probability of error as

$$P_e = \mathbb{P}[H_0]P_{FA} + \mathbb{P}[H_1]P_{MD} = 0.40872.$$

We can rewrite the maximum likelihood decision rule in terms of a ratio. Define the **likelihood ratio** as a function of the measurement value $Y = y$, as

$$\mathcal{L}(y) = \frac{P_{Y|H_1}(y)}{P_{Y|H_0}(y)}.$$

The maximum likelihood decision rule can be written in terms of the likelihood ratio as

$$D^{ML}(y) = \begin{cases} 1, & \mathcal{L}(y) \geq 1, \\ 0, & \mathcal{L}(y) < 1. \end{cases}$$

We abbreviate this decision using this notation: $D^{ML}(y) = \{\mathcal{L}(y) = \frac{P_{Y|H_1}(y)}{P_{Y|H_0}(y)} \underset{H_0}{\overset{H_1}{\gtrless}} 1\}$. This indicates that, when the inequality is in the "greater than" direction, the decision selected is that of hypothesis H_1, and when the inequality is reversed, the decision selected is that of hypothesis H_0.

We can often compute the maximum likelihood decision rule analytically using the expressions for the probability mass functions and the likelihood ratio. For Example 6.5, the likelihood ratio is

$$\mathcal{L}(y) = \frac{\binom{5}{y}(0.4)^{5-y}(0.6)^y}{\binom{5}{y}(0.5)^{5-y}(0.5)^y} = \frac{(0.4)^{5-y}(0.6)^y}{(0.5)^5} = 2^5(0.4)^{5-y}(0.6)^y = (0.8)^5(1.5)^y$$

We want to compare $\mathcal{L}(y)$ to 1. Therefore, the maximum likelihood detection rule is $\mathcal{L}(y) = \frac{P_{Y|H_1}(y)}{P_{Y|H_0}(y)} \underset{H_0}{\overset{H_1}{\gtrless}} 1$.

To compute the performance of the maximum likelihood detector, we need to identify the values of $Y = y$ for which $D^{ML}(y) = 0$ and for which $D^{ML}(y) = 1$. When we enumerate the likelihoods for all values of $Y = y$ as in Example 6.5, this is straightforward. For larger R_Y, enumeration is impractical, so we need to further simplify the maximum likelihood decision rule to determine these regions.

To simplify this, we make the following observation: $\mathcal{L}(y) > 1 \iff \ln\big(\mathcal{L}(y)\big) > 0$. Computing the logarithm of the likelihood ratio $\mathcal{L}(y)$ yields $\ln\big(\mathcal{L}(y)\big) = 5\ln(0.8) + y\ln(1.5)$. Then,

$$\ln\big(\mathcal{L}(y)\big) > 0 \iff y > \frac{5\ln(1.25)}{\ln(1.5)} \approx 2.751.$$

Thus, for $y = 3, 4, 5$, the likelihood ratio $\mathcal{L}(y)$ is greater than 1, and for $y = 0, 1, 2$, the likelihood ratio is less than 1. This is the same maximum likelihood decision rule derived in Example 6.5.

Using logarithms often makes it easier to identify the decision rule in terms of a region of values of y, as we saw above. We can write the maximum likelihood decision rule in terms of the **log-likelihood ratio**, the logarithm of the likelihood ratio, as $D^{ML}(y) = \{\ln\left(\frac{P_{Y|H_1}(y)}{P_{Y|H_0}(y)}\right) \underset{H_0}{\overset{H_1}{\gtrless}} 0\}$.

Example 6.6

Radar systems usually send trains of pulses to detect the presence of aircraft in the direction the radar is aimed at. Each of these pulses potentially generates a reflection; for each pulse, a decision as to whether an aircraft is present or not can be made based on the received pulse signal strength, comparing it to a threshold. The final decision for detecting the presence of aircraft is based on the total number of pulses received that had sufficient signal strength. The detections on each pulse are assumed to be independent, conditioned on whether an aircraft is present or not.

Assume that the probability of detecting an aircraft in a single pulse, assuming the aircraft is present, is p_1. If the aircraft is not present, the probability of having enough background signal strength to generate a detection is p_0. Assume that n pulses get transmitted, and $p_1 > p_0$. What is the maximum likelihood detector?

The problem is stated in terms of two hypotheses: H_1 is where the aircraft is present, and H_0 is where there is no aircraft present. From the problem description, the observation Y consists of the number of pulses that generate a detection, which can take values in $\{0, 1, \ldots, n\}$. The likelihood $P_{Y|H_1}(y)$ is a Binomial(n, p_1) distribution, and the likelihood $P_{Y|H_0}(y)$ is a Binomial(n, p_0) distribution.

Since n, p_1, p_0 are left as variables, we cannot simply enumerate the possible values of Y in a table and find the best decision for each value of y. Nevertheless, we can analyze this using log-likelihood ratios, as:

$$\mathcal{L}(y) = \frac{P_{Y|H_1}(y)}{P_{Y|H_0}(y)} = \frac{\binom{n}{y}p_1^y(1-p_1)^{n-y}}{\binom{n}{y}p_0^y(1-p_0)^{n-y}} = \Big(\frac{1-p_1}{1-p_0}\Big)^n \Big(\frac{(p_1(1-p_0)}{p_0(1-p_1)}\Big)^y$$

$$\ln(\mathcal{L}(y)) = n\ln(\frac{1-p_1}{1-p_0}) + y\ln(\frac{(p_1(1-p_0)}{p_0(1-p_1)})$$

We see that the log-likelihood ratio is increasing in y (because $p_1 > p_0$, so $1 - p_1 < 1 - p_0$.) Furthermore for $y = 0$, the log-likelihood ratio is negative. Hence, there is a value of y for which the log-likelihood ratio equals 1. That value is

$$y^* = \frac{n\ln(1-p_0) - n\ln(1-p_1)}{\ln(p_1(1-p_0)) - \ln(p_0(1-p_1))}$$

For instance, if $p_1 = 0.7, p_2 = 0.2, n = 20$, we get $y^* \approx 8.78$, so the maximum likelihood detector declares a detection if 9 or more pulses are detected. Hence, $D^{ML}(y) = \{y \underset{H_0}{\overset{H_1}{\gtrless}} 8.78\}$, which is a simple detector to implement.

We can now compute the probabilities of missed detection and false alarm as sums, as

$$P_{MD} = \mathbb{P}[\{Y < y^*\}|H_1] = \sum_{y<y^*} \binom{n}{y} (p_1)^y (1-p_1)^{n-y}.$$

$$P_{FA} = \mathbb{P}[\{Y > y^*\}|H_0] = \sum_{y>y*} \binom{n}{y} (p_0)^y (1-p_0)^{n-y}.$$

For the values $p_1 = 0.7, p_2 = 0.2, n = 20$, we get $P_{MD}] \approx 0.005, P_{FA} \approx 0.010$, which shows that, even though single pulse detection is not very accurate, by sending 20 pulses we increase our performance to near-perfect detection.

For continuous observations Y, the maximum likelihood rule is expressed in terms of the likelihood ratio using the conditional probability densities $f_{Y|H_1}(y), f_{Y|H_0}(y)$. In this case, $\mathcal{L}(y) = \frac{f_{Y|H_1}(y)}{f_{Y|H_0}(y)}$, and the maximum likelihood decision rule is given as $D^{ML}(y) = \{\frac{P_{Y|H_1}(y)}{P_{Y|H_0}(y)} \underset{H_0}{\overset{H_1}{\gtrless}} 1\}$. For continuous random variables, enumerating the likelihood values for each y is no longer possible; to find the regions $A_1 = \{y \in R_Y : D^{ML}(y) = 1\}$ and $A_0 = \{y \in R_Y : D^{ML}(y) = 0\}$, we use the log-likelihood ratio to solve for the region.

Example 6.7
You are interested in diagnosing whether a person has a fever associated with a particular disease based on measuring their temperature. If the person does not have a disease, the measured temperature is expected to be a Gaussian random variable with mean 98.1 degrees Fahrenheit and standard deviation 1 degree Fahrenheit. If the person has the disease, the average temperature is 101 degrees Fahrenheit and standard deviation 1 degree Fahrenheit. What is the maximum likelihood detector? For the maximum likelihood detector, what are the probabilities of missed detection and false alarm?

Let H_1 be the event where the person has the disease, and H_0 the event where the person does not have the disease. The maximum likelihood detector is readily written in terms of the likelihood ratio as:

$$D^{ML}(y) = \left\{ \frac{\frac{1}{\sqrt{2\pi}} e^{-\frac{(y-101)^2}{2}}}{\frac{1}{\sqrt{2\pi}} e^{-\frac{(y-98.1)^2}{2}}} \underset{H_0}{\overset{H_1}{\gtrless}} 1. \right\}$$

To evaluate the performance, we use the log-likelihood ratio, which is

$$\ln \mathcal{L}(y) = \ln \left(\frac{e^{-\frac{(y-101)^2}{2}}}{e^{-\frac{(y-98.1)^2}{2}}} \right) = -\frac{(y-101)^2}{2} + \frac{(y-98.1)^2}{2}$$
$$= (101 - 98.1)y - \frac{101^2}{2} + \frac{98.1^2}{2} = (101 - 98.1)y - \frac{(101 - 98.1)(101 + 98.1)}{2}.$$

Equating this to 0, we get that $y^* = \frac{(101+98.1)}{2} = 99.55$, the average of the two expected values. If $y > y^*$, then $D^{ML}(y) = 1$, and if $y < y^*, D^{ML} = 0$. This is illustrated in the figure on the right, where the vertical blue line shows the value of y^*. To the right of that blue line, we have $f_{Y|H_1}(y) > f_{Y|H_0}(y)$. To the left, the inequality is reversed. We can now compute the performance as follows:

$$P_{FA} = \mathbb{P}[\{y \geq 99.55\}|H_0] = 1 - \Phi(1.45) = Q(1.45).$$

where the threshold $y^* = 99.55$ is 1.45 standard deviations higher than the average 98.1. Similarly,

$$P_{MD} = \mathbb{P}[\{y \geq 99.55\}|H_1] = \Phi(-1.45) = Q(1.45).$$

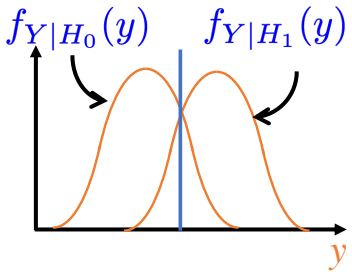


Figure 6.6: Example 6.7.

6.3 Maximum A Posteriori (MAP) Detection

In maximum likelihood detection, we designed the detection rule independent of the prior probabilities of each event hypothesis, $\mathbb{P}[H_0]$ and $\mathbb{P}[H_1]$. However, in many cases, the probabilities $\mathbb{P}[H_0]$ and $\mathbb{P}[H_1]$ can be very different. For instance, when testing for the presence of measles in a college-age student, the probability that the observed symptoms actually come from measles is small, as most college-age students have received an immunization vaccine. In this section, we show how to design detection algorithms that integrate this type of information.

Assume the measurements Y are discrete-valued, and we know $\mathbb{P}[H_0], \mathbb{P}[H_1]$. We refer to $\mathbb{P}[H_0], \mathbb{P}[H_1]$ as the prior probabilities, as they are known before measuring Y. After measuring Y, we compute the a posteriori or conditional probabilities of H_0 and H_1 given $Y = y$ using Bayes' Rule, as

$$\mathbb{P}[H_0|\{Y=y\}] = \frac{\mathbb{P}[H_0 \cap \{Y=y\}]}{\mathbb{P}[\{Y=y\}]} = \frac{\mathbb{P}[\{Y=y\}|H_0]\mathbb{P}[H_0]}{\mathbb{P}[\{Y=y\}]} = \frac{P_{Y|H_0}(y)\mathbb{P}[H_0]}{\mathbb{P}[\{Y=y\}]},$$

where the denominator is computed using the Law of Total Probability, as

$$\mathbb{P}[\{Y=y\}] = \mathbb{P}[\{Y=y\}|H_0]\mathbb{P}[H_0] + \mathbb{P}[\{Y=y\}|H_1]\mathbb{P}[H_1] = P_{Y|H_0}(y)\mathbb{P}[H_0] + P_{Y|H_1}(y)\mathbb{P}[H_1].$$

Similarly,

$$\mathbb{P}[H_1|\{Y=y\}] = \frac{\mathbb{P}[\{Y=y\}|H_1]\mathbb{P}[H_1]}{\mathbb{P}[\{Y=y\}]} = \frac{P_{Y|H_1}(y)\mathbb{P}[H_1]}{\mathbb{P}[\{Y=y\}]}.$$

The **maximum a posteriori (MAP)** decision rule is defined as follows:

$$D^{MAP}(y) = \begin{cases} 1, & \mathbb{P}[H_1|\{Y=y\}] \geq \mathbb{P}[H_0|\{Y=y\}], \\ 0, & \mathbb{P}[H_0|\{Y=y\}] > \mathbb{P}[H_1|\{Y=y\}]. \end{cases}$$

where we arbitrarily assign ties to 1. Since the denominator in Bayes' Rule is the same for $\mathbb{P}[H_1|\{Y=y\}]$ and $\mathbb{P}[H_0|\{Y=y\}]$, this rule is the same as

$$D^{MAP}(y) = \begin{cases} 1, & P_{Y|H_1}(y)\mathbb{P}[H_1] \geq P_{Y|H_0}(y)\mathbb{P}[H_0], \\ 0, & P_{Y|H_0}(y)\mathbb{P}[H_0]] > P_{Y|H_1}(y)\mathbb{P}[H_1]. \end{cases}$$

This allows us to rewrite the MAP decision rule in terms of the likelihood ratio, as

$$D^{MAP}(y) = \left\{ \mathcal{L}(y) = \frac{P_{Y|H_1}(y)}{P_{Y|H_0}(y)} \underset{H_0}{\overset{H_1}{\gtrless}} \frac{\mathbb{P}[H_0]}{\mathbb{P}[H_1]} \right\}.$$

Note that the data-dependent computation in the MAP decision rule is to compute the likelihood ratio, just as in the ML decision rule. What changes is the threshold that one compares the maximum likelihood to. In the ML case, the threshold is 1. This is also true in the MAP case if $\mathbb{P}[H_0] = \mathbb{P}[H_1]$. However, if $\mathbb{P}[H_1] > \mathbb{P}[H_0]$, the threshold is lower than 1, and the number of y for which the decision equals 1 is possibly increased. If $\mathbb{P}[H_0]$ is larger, then the threshold is larger than 1, and the number of y for which the decision equals 1 may be decreased.

Example 6.8

Assume we have the same problem as Example 6.5, but the prior probability that the coin is biased is only $\mathbb{P}[H_1] = 0.4$, so $\mathbb{P}[H_0] = 0.6$ because H_0, H_1 form a partition of Ω. From Example 6.5, we know the likelihoods of Y, the number of heads observed in 6 trials, are shown in the table below.

Y:	0	1	2	3	4	5
$P_{Y\|H_1}$	0.01024	0.0768	0.2304	0.3456	0.2592	**0.07776**
$P_{Y\|H_0}$	0.03125	0.15625	0.3125	0.3125	0.15625	0.03125
$\mathcal{L}(y)$	0.3277	0.4915	0.7373	1.1059	**1.6589**	**2.4883**

We have added to the table a row computing the likelihood ratio for each value of Y. The threshold in the MAP decision rule is $\frac{\mathbb{P}[H_0]}{\mathbb{P}[H_1]} = 1.5$. The values of $Y = y$ for which the likelihood ratio exceeds the threshold are highlighted in bold magenta above. We see that increasing the threshold has decreased the number of y for which the MAP decision is 1. The MAP decision rule and the ML decision rule from Example 6.5 are shown below:

To compute the maximum likelihood decision, we compare the numbers in each column, and pick the larger of the two numbers. In the table above, we have highlighted the larger number in bold and magenta color. Thus we see that the maximum likelihood decision rule becomes:

$$D^{MAP}(y) = \begin{cases} 1, & y = 4, 5, \\ 0, & y = 0, 1, 2, 3. \end{cases} \qquad D^{ML}(y) = \begin{cases} 1, & y = 3, 4, 5, \\ 0, & y = 0, 1, 2. \end{cases}$$

The decision agrees with intuition: a larger count of heads suggests the coin is more likely to be unbalanced, whereas a smaller count of heads indicates the coin is more likely to be balanced.

Since the ML and MAP decision rules are different, they have different performance. The probability of false alarm for the MAP decision rule is

$$P_{FA}^{MAP} = \mathbb{P}[\{Y = 4, 5\}|H_0] = P_{Y|H_0}(4) + P_{Y|H_0}(4) = 0.1875.$$

The probability of missed detection for the MAP decision rule is

$$P_{MD}^{MAP} = \mathbb{P}[\{Y = 0, 1, 2, 3\}|H_1] = P_{Y|H_1}(0) + P_{Y|H_1}(1) + P_{Y|H_1}(2) + P_{Y|H_1}(3) = 0.633.$$

In contrast, for the ML decision rule, $P_{FA}^{ML} = 0.5, P_{MD}^{ML} = 0.3174$. Thus, increasing the threshold reduced the probability of false alarm, and increased the probability of missed detection. The probability of error for each of the detectors is

$$P_e^{MAP} = \mathbb{P}[H_0]P_{FA}^{MAP} + \mathbb{P}[H_1]P_{MD}^{MAP} = 0.6 \cdot 0.1875 + 0.4 \cdot 0.633 \approx 0.3777.$$

$$P_e^{ML} = \mathbb{P}[H_0]P_{FA}^{ML} + \mathbb{P}[H_1]P_{MD}^{ML} = 0.6 \cdot 0.5 + 0.4 \cdot 0.3174 \approx 0.4270.$$

We will show later that the MAP decision rule achieves the minimum probability of error among all possible decision rules.

The MAP decision rule for continuous-valued measurements Y is a straightforward extension of the MAP decision rule for discrete-valued measurements Y. We have to be a bit careful to define $\mathbb{P}[H_0|\{Y = y\}]$ and $\mathbb{P}[H_1|\{Y = y\}]$ using a limiting argument, as in Chapter 4.4.3, because $\mathbb{P}[\{Y = y\}] = 0$. Specifically,

$$\begin{aligned}\mathbb{P}[H_0|\{Y \in (y, y+\Delta]\}] &= \frac{\mathbb{P}[H_0 \cap \{Y \in (y, y+\Delta]\}]}{\mathbb{P}[\{Y \in (y, y+\Delta)\}]} = \frac{\mathbb{P}[\{Y \in (y, y+\Delta]\}|H_0]\mathbb{P}[H_0]}{\mathbb{P}[\{Y \in (y, y+\Delta)\}]} \\ &= \frac{(F_{Y|H_0}(y+\Delta) - F_{Y|H_0}(y))\mathbb{P}[H_0]}{F_Y(y+\Delta) - F_Y(y)}\end{aligned}$$

As $\Delta \to 0$, both numerator and denominator approach 0. We use L'Hopital's rule to evaluate the limit, as

$$\lim_{\Delta\to 0} \mathbb{P}[H_0|\{Y \in (y, y+\Delta]\}] = \lim_{\Delta\to 0} \frac{\frac{d}{d\Delta}(F_{Y|H_0}(y+\Delta) - F_{Y|H_0}(y))\mathbb{P}[H_0]}{\frac{d}{d\Delta}(F_Y(y+\Delta) - F_Y(y))} = \frac{f_{Y|H_0}(y)\mathbb{P}[H_0]}{f_Y(y)} = \mathbb{P}[H_0|Y = y].$$

Similarly, $\mathbb{P}[H_1|Y = y] = \frac{f_{Y|H_1}(y)\mathbb{P}[H_1]}{f_Y(y)}$, and the marginal density is obtained by the Law of Total Probability as

$$f_Y(y) = f_{Y|H_0}(y)\mathbb{P}[H_0] + f_{Y|H_1}(y)\mathbb{P}[H_1].$$

This leads to the MAP decision rule in terms of the likelihood ratio

$$D^{MAP}(y) = \left\{ \mathcal{L}(y) = \frac{f_{Y|H_1}(y)}{f_{Y|H_0}(y)} \underset{H_0}{\overset{H_1}{\gtrless}} \frac{\mathbb{P}[H_0]}{\mathbb{P}[H_1]} \right\}.$$

Example 6.9

The delay Y in arrival of an on-line order is modeled as an exponential random variable, but the rate of that random variable iis one of two possible rates. Under hypothesis H_1, the rate is 0.2/day, and under hypothesis H_0, the rate is 0.1/day. The prior probability that hypothesis H_0 is correct is $\mathbb{P}[H_0] = 0.6$. Assume we observe $Y = y$. What is the MAP decision rule, and what is its probability of error?

The threshold for the MAP decision rule for the probability of error is $T = \frac{\mathbb{P}[H_0]}{\mathbb{P}[H_1]} = \frac{3}{2}$. The likelihood ratio for the exponential random variables is

$$\mathcal{L}(y) = \frac{0.2e^{-0.2y}}{0.1e^{-0.1y}} = 2e^{-0.1y},$$

which is decreasing as y increases. Thus, longer observed delays y make hypothesis H_0 more likely, as its rate of arrival is smaller.

The boundary for the decision region in terms of y can be found by solving $\mathcal{L}(y) = 2e^{-0.1y} = \frac{3}{2}$. Taking logarithms,

$$-0.1y = \ln(3) - \ln(4) \Rightarrow y = 10(\ln(4) - \ln(3)) \approx 2.877.$$

Thus, if $y < 2.877$, select $D^{MAP}(y) = 1$; else, select $D^{MAP}(y) = 0$. With these regions, we have

$$P_{FA} = \int_0^{2.877} f_{Y|H_0}(y)\, dy = F_{Y|H_0}(2.877) = 1 - e^{-0.2877} = 0.25,$$

$$P_{MD} = \int_{2.877}^{\infty} f_{Y|H_1}(y)\,dy = e^{-0.2*2.877} = \frac{9}{16} = 0.5625.$$

The probability of error is

$$P_e = \mathbb{P}[H_0]P_{FA} + \mathbb{P}[H_1]P_{MD} = 0.6 * 0.25 + 0.4 * 0.5625 = 0.375.$$

We conclude this section by showing that the MAP decision rule minimizes the probability of error among all decision rules. For any decision rule $D(y)$, the probability of error conditioned on $Y = y$ is given as follows: Since H_0, H_1 are a partition of Ω,

$$\mathbb{P}[\text{Error}|Y = y] = \mathbb{P}[\text{Error} \cap H_1|Y = y] + \mathbb{P}[\text{Error} \cap H_0|Y = y].$$

$$\mathbb{P}[\text{Error} \cap H_1|Y = y] = \mathbb{P}[\text{Error}|Y = y, H_1]\mathbb{P}[H_1|Y = y];$$
$$\mathbb{P}[\text{Error} \cap H_0|Y = y] = \mathbb{P}[\text{Error}|Y = y, H_0]\mathbb{P}[H_0|Y = y],$$

which follows from the definition of conditional probability. Note that $\mathbb{P}[\text{Error}|Y = y, H_0] = I_{D(y)=1}$, where I_A is the indicator function that is 1 if A is true, and 0 elsewhere. Similarly, $\mathbb{P}[\text{Error}|Y = y, H_1] = I_{D(y)=0}$. Therefore,

$$\mathbb{P}[\text{Error}|Y = y] = I_{D(y)=1}\mathbb{P}[H_0|Y = y] + I_{D(y)=0}\mathbb{P}[H_1|Y = y].$$

Note that $D^{MAP}(y)$ selects the smallest of the two terms for each $Y = y$, and hence has the smallest probability of error for each $Y = y$. The unconditional probability of error is, assuming Y is discrete, as

$$P_e = \sum_{y \in R_Y} \Big(I_{D(y)=1}\mathbb{P}[H_0|Y = y] + I_{D(y)=0}\mathbb{P}[H_1|Y = y] \Big) P_Y(y),$$

which $D^{MAP}(y)$ will minimize because it minimizes each term in the sum.

For continuous Y, we get

$$P_e = \int_{y \in R_Y} \Big(I_{D(y)=1}\mathbb{P}[H_0|Y = y] + I_{D(y)=0}\mathbb{P}[H_1|Y = y] \Big) f_Y(y)\,dy,$$

which is minimized by $D^{MAP}(y)$ because $D^{MAP}(y)$ minimizes the integrand for every value of y, and hence it minimizes the integral.

6.4 Minimum Bayes Risk Detection

In many important situations, there is a different cost associated with the different types of errors. For instance, in luggage inspection, a false alarm can result in an unnecessary opening of a suitcase to check its contents. However, a missed detection can result in an explosive entering the airplane. In breast cancer diagnosis, a false alarm can lead to an unneeded biopsy, whereas a missed detection can be life-threatening.

To properly evaluate this tradeoff, we assign different costs to the different types of errors, and design a decision rule to minimize the total expected cost. Formally, let C_{ij} denote the cost of deciding U_i when H_j is true. We typically select $C_{11} = 0, C_{00} = 0$, so that correct decisions involve no cost; while this is not essential, it is wasted space to consider the full generality, as it is never used in practice. The key tradeoff is the relative cost of a missed detection C_{01} and a false alarm C_{10}. The Figure on the right illustrates the indexing as to what the costs mean for different values of decision and true hypothesis.

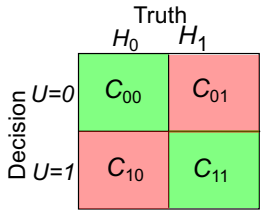


Figure 6.7: Bayes' Costs.

We follow closely the development in the previous section where we showed the MAP decision rule minimized P_e, the probability of making an error. For an arbitrary decision rule $D(y)$, let R denote the cost

of the decision rule. R is a random variable defined on the experiment, which depends on the outcome s and the observation y, as $R(s,y) = C_{01}I_{\{\omega\in H_1\}\cap\{D(y(\omega))=0\}} + C_{10}I_{\{\omega\in H_0\}\cap\{D(y(\omega))=1\}}$. Then, the conditional probability mass function of R given $Y=y$ and $D(y)=0$ is

$$P_{R|Y,\{D=0\}}(r|y) = \begin{cases}\mathbb{P}[H_1|Y=y], & r=C_{01}\\ \mathbb{P}[H_0|Y=y], & r=0.\end{cases} \quad P_{R|Y,\{D=1\}}(r|y) = \begin{cases}\mathbb{P}[H_0|Y=y], & r=C_{10}\\ \mathbb{P}[H_1|Y=y], & r=0.\end{cases}$$

Then,

$$\mathbb{E}[R|Y=y] = C_{01}\mathbb{P}[H_1|Y=y]I_{D(y)=0} + C_{10}\mathbb{P}[H_0|Y=y]I_{D(y)=1}.$$

The decision that minimizes this conditional expected risk given measurement $Y=y$ is

$$D^{MBR}(y) = \begin{cases}1, & C_{01}\mathbb{P}[H_1|Y=y] \geq C_{10}\mathbb{P}[H_0|Y=y]\\ 0, & C_{01}\mathbb{P}[H_1|Y=y] < C_{10}\mathbb{P}[H_0|Y=y].\end{cases}$$

For discrete random variables Y, the expected risk for any decision $D(y)$ is written as:

$$\mathbb{E}[R] = \sum_{y\in R_Y} \mathbb{E}[R|Y=y]P_Y(y).$$

Since the minimum Bayes risk (MBR) minimizes each term of the sum among all decision rules, it is the optimal decision rule for minimizing the expected Bayes risk. For continuous random variables Y, the expected Bayes risk of any decision rule is

$$\mathbb{E}[R] = \int_{y\in R_Y} \mathbb{E}[R|Y=y]f_Y(y)\,dy.$$

The MBR decision rule $D^{MBR}(y)$ minimizes the integrand for each y, and hence minimizes the expectation.

We can write D^{MBR} in terms of the likelihood ratio $\mathcal{L}(y)$ using Bayes' Rule: for discrete Y,

$$\mathbb{P}[H_1|Y=y] = \frac{P_{Y|H_1}(y)\mathbb{P}[H_1]}{P_Y(y)}; \quad \mathbb{P}[H_0|Y=y] = \frac{P_{Y|H_0}(y)\mathbb{P}[H_0]}{P_Y(y)}.$$

Recall that $\iff$ means "if and only if"; then,

$$\begin{aligned} C_{01}\mathbb{P}[H_1|Y=y] \geq C_{10}\mathbb{P}[H_0|Y=y] &\iff C_{01}P_{Y|H_1}(y)\mathbb{P}[H_1] \geq C_{10}P_{Y|H_0}(y)\mathbb{P}[H_0]\\ &\iff \frac{P_{Y|H_1}(y)}{P_{Y|H_0}(y)} \geq \frac{C_{10}\mathbb{P}[H_0]}{C_{01}\mathbb{P}[H_1]}\end{aligned}$$

Thus, the minimum Bayes risk decision rule is

$$D^{MBR}(y) = \left\{\frac{P_{Y|H_1}(y)}{P_{Y|H_0}(y)} \underset{H_0}{\overset{H_1}{\gtrless}} \frac{C_{10}\mathbb{P}[H_0]}{C_{01}\mathbb{P}[H_1]}\right\}.$$

For continuous measurements $Y=y$, the minimum Bayes risk decision rule is

$$D^{MBR}(y) = \left\{\frac{f_{Y|H_1}(y)}{f_{Y|H_0}(y)} \underset{H_0}{\overset{H_1}{\gtrless}} \frac{C_{10}\mathbb{P}[H_0]}{C_{01}\mathbb{P}[H_1]}\right\}.$$

Note the following: The MAP decision rule is a special case of the MBR decision rule when $C_{10}=C_{01}$. The ML decision rule is another case of the MBR decision rule when $C_{10}=C_{01}, \mathbb{P}[H_1]=\mathbb{P}[H_0]$. In general, all MBR decision rules are based on comparing the likelihood ratio value for $Y=y$ to a threshold, where the threshold is computed from the relative costs and the prior probabilities of H_0, H_1.

The threshold varies with the relative cost of false alarms and missed detections in an intuitive manner. If missed detection are more expensive than false alarms, then the threshold for the likelihood ratio is set lower, so that one is more likely to decide that H_1 is the correct hypothesis.

Example 6.10
Consider Example 6.7, which sought to diagnose the presence of a disease by measuring the temperature. Assume a priori that the probability of having the disease ($\mathbb{P}[H_1]$) is 0.4, and thus the probability of not having the disease ($\mathbb{P}[H_0]$) is 0.6. However, the cost of a missed detection is 10 (C_{01}), whereas the cost of a false alarm (C_{10}) is 1. What is the minimum Bayes risk decision rule, and what are the resulting probabilities of false alarm and missed detection?

The MBR decision rule is

$$D^{MBR}(y) = \left\{ \mathcal{L}(y) \underset{H_0}{\overset{H_1}{\gtrless}} \frac{C_{10}\mathbb{P}[H_0]}{C_{01}\mathbb{P}[H_1]} = \frac{3}{20} \right\}.$$

From the results of Example 6.7, the likelihood ratio is

$$\mathcal{L}(y) = \frac{e^{-\frac{(y-101)^2}{2}}}{e^{-\frac{(y-98.1)^2}{2}}}$$

Then,

$$\mathcal{L}(y) \leq \frac{3}{20} \iff \ln\Big(\mathcal{L}(y)\Big) \leq \ln(3) - \ln(20)$$

$$\ln\Big(\mathcal{L}(y)\Big) = (101 - 98.1)y - \frac{(101 - 98.1)(101 + 98.1)}{2} = 2.9y - (2.9) \cdot (99.55)$$

$$D^{MBR}(y) = \left\{ y \underset{H_0}{\overset{H_1}{\gtrless}} 99.55 + \frac{1}{2.9}(\ln(3) - \ln(20)) = 99.2159 \right\}.$$

Thus, we see that the threshold for the decision rule has been lowered as compared to the ML decision rule of Example 6.7. This means that the probability of missed detection decreases, and the probability of false alarm increases. Since the mean under H_0 is 98.1, the threshold is 1.1159 standard deviations higher than the mean, so $P_{FA} = Q(1.1159)$. Since the mean under H_0 is 101, the threshold is 1.7841 standard deviations lower than the mean, so $P_{MD} = \Phi(-1.7841) = Q(1.7841)$.

The minimum expected Bayes risk is given in terms of these measures, as

$$\mathbb{E}[R] = \mathbb{P}[H_0]C_{10}P_{FA} + \mathbb{P}[H_1]C_{01}P_{MD} = 0.6P_{FA} + 4P_{MD} = 0.6Q(1.1159) + 4Q(1.841)$$

for the MBR decision rule, and

$$\mathbb{E}[R] = 0.6Q(1.45) + 4Q(1.45)$$

for the ML decision rule, which is higher than the MBR expected Bayes risk.

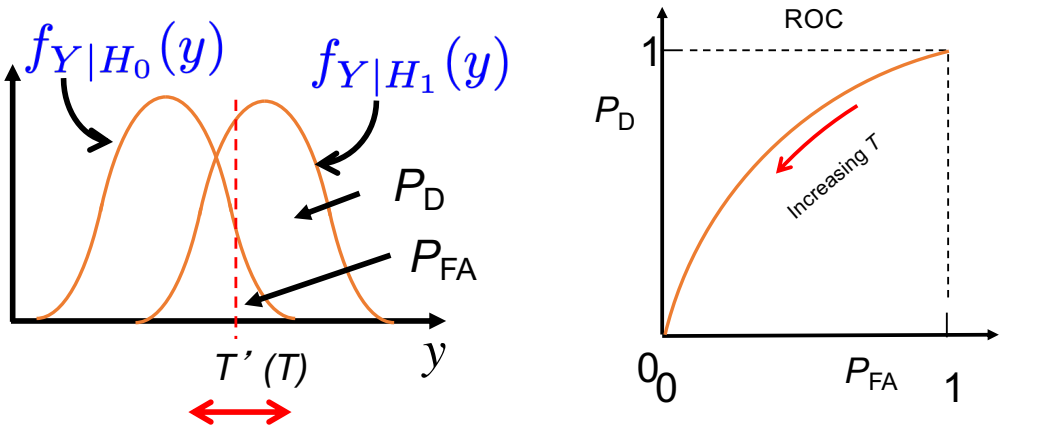
6.5 Performance and the Receiver Operating Characteristic

In the discussion so far, we have found that the optimal decision rule for binary hypotheses is a likelihood ratio test, where we compute a function of the measured data (the likelihood ratio) and compare it to a threshold. The choice of threshold depends on the prior probabilities of each hypotheses, plus the costs of making a missed detection. These four parameters are summarized in a single threshold T; to design an optimal decision rule, we simply select this threshold T, and the decision rule is

$$D(y) = \left\{ \mathcal{L}(y) \underset{H_0}{\overset{H_1}{\gtrless}} T \right\}.$$

The choice of threshold T controls the tradeoff between the conditional performance statistics P_{MD} and P_{FA}. As T increases, the decision rule selects H_1 less often, which increases P_{MD} and decreases P_{FA}.

Define the probability of detection $P_D = 1 - P_{MD}$. As the threshold T decreases to 0, the region of measurements $Y = y$ for which the decision is 1 increases, eventually becoming the entire range R_Y. When the threshold is 0, the performance statistics are $P_D = 1, P_{FA} = 1$, since the decision is always 1. Similarly, as the threshold increases to ∞, the region of measurements $Y = y$ for which the decision is 1 decreases, eventually becoming empty. For a threshold of ∞, the performance statistics are $P_D = 0, P_{FA} = 0$. As the threshold T is varied from 0 to ∞, we can trace a locus of performance of $P_D(T)$ versus $P_{FA}(T)$, which is called the **Receiver Operating Characteristic** or ROC for the detection problem. The design of an optimal decision rule based on likelihood ratios reduces to selecting a point on the ROC that trades off P_D versus P_{FA}. An illustration of a ROC is given in Figure 6.8.



Densities indicating decision regions Receiver Operating Characteristic (ROC)

Figure 6.8: Illustration of ROC for detection involving two Gaussian Distributions.

Let us emphasize some features of the ROC. First, note that the threshold T 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 $\mathbb{P}[H_i]$. These terms can be used to determine a particular threshold, and thus a particular operating point corresponding to the optimal Bayes risk detector. A couple of important properties of the ROC are:

- The ROC is monotone non-decreasing. Increasing P_{FA} results in increasing P_D.

- The ROC is a concave curve, with the graph above the $P_D = P_{FA}$ line. Performance on the line $P_D = P_{FA}$ correspond to detectors that that randomly guess $D(y) = 1$ with probability p, independent of the measured value y. The optimal detectors achieve better performance by using the information in y. This argument can be extended to show the ROC is a concave curve.

Determining the ROC requires computing the region $A_1(T) = \{y \in R_Y : \mathcal{L}(y) \geq T\}$ where the likelihood ratio decision rule results in decision 1 for threshold T. If we know that region, then $P_D(T) = \mathbb{P}[\{y \in A_1(T)\}|H_1], P_{FA}(T) = \mathbb{P}[\{y \in A_1(T)\}|H_0]$. By varying T, we obtain the points on the ROC. We discuss examples to show how this is done.

Example 6.11
We have a coin that may be biased so that the probability of Heads is 0.8 (Hypothesis H_1.) If the coin is unbiased, the probability of Heads is 0.5 (Hypothesis H_0.) We conduct three independent flips and count the number of heads as our measurement Y. The likelihoods and the likelihood ratio are shown in the table below:

Y:	0	1	2	3
$P_{Y\|H_1}$	0.0080	0.0960	0.3840	0.5120
$P_{Y\|H_0}$	0.1250	0.3750	0.3750	0.1250
$\mathcal{L}(y)$	0.0640	0.2560	1.0240	4.0960

We see that, for thresholds T above 4.1, the decision is always $D(y) = 0$, and so $P_D = 0, P_{FA} = 0$. For thresholds around $T = 1.1$, $D(y) = 1$ if $y = 3$, and 0 otherwise. Thus, $P_D = 0.5120, P_{FA} = 0.1250$. As we lower the threshold to between 0.26 and 1.02, $D(y) = 1$ for $y = 2, 3$ and 0 for $y = 0, 1$. Then, $P_D = 0.896, P_{FA} = 0.5$. We continue this and obtain the various points, plotted on the ROC figure on the right.

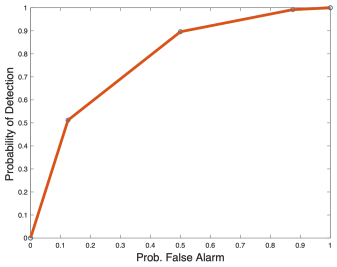


Figure 6.9: ROC for example.

Note that we have connected the discrete points in the ROC with straight lines. One can achieve performance on those straight lines by randomly switching between the thresholds corresponding to the two endpoints on the line. That type of random decision rule can be used to achieve a desired P_{FA} that is different from the finite ones obtained by the discrete breakpoints in the likelihood ratio table above.

Example 6.12
We have a light source that can either have an intensity of 100 photons/second, or 200 photons/second. We measure the number of photons emitted over a 1 second period, and have to decide which is the correct intensity for the light source. Let H_1 correspond to intensity of 120 photons/second, and H_0 correspond to intensity of 100 photons/second. If H_1 is correct, the number of photons measured is a Poisson(120) random variable; if H_0 is correct, the number is a Poisson(100) random variable.

The likelihood ratio for this problem is

$$\mathcal{L}(y) = \frac{P_{Y|H_1}(y)}{P_{Y|H_0}(y)} = \frac{\frac{120^y}{y!}e^{-120}}{\frac{100^y}{y!}e^{-100}} = (1.2)^y e^{-20}$$

An optimal likelihood ratio test is $\left\{\mathcal{L}(y) \underset{H_0}{\overset{H_1}{\gtrless}} T\right\}$ for a threshold T. Taking logarithms, of both sides, this reduces to

$$\ln\left(\mathcal{L}(y)\right) = y\ln(1.2) - 20 > \ln(T) \iff y > \frac{\ln(T) + 20}{\ln 1.2}.$$

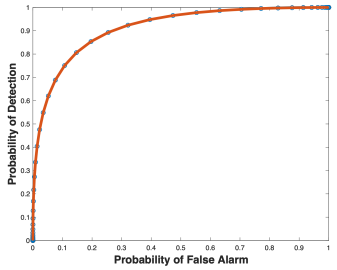


Figure 6.10: ROC for example.

For instance, for the ML decision rule, $T = 1$, and so the ML decision rule is $\{y \underset{H_0}{\overset{H_1}{\gtrless}} 109.7\}$. The ROC is shown in Figure 6.10 , where we have connected the discrete points in the ROC with straight lines.

Example 6.13 (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 $N(m, \sigma^2)$ under H_1. The likelihood ratio is

$$\mathcal{L}(y) = \frac{\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(y-m)^2}{2\sigma^2}}}{\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(y)^2}{2\sigma^2}}} = e^{-\frac{(y-m)^2}{2\sigma^2}+\frac{(y)^2}{2\sigma^2}}.$$

and the log-likelihood ratio is

$$\ln\left(\mathcal{L}(y)\right) = \frac{1}{2\sigma^2}(2my - m^2).$$

Hence, comparing the log-likelihood ratio to the log of a threshold T yields the decision rule

$$y \underset{H_0}{\overset{H_1}{\gtrless}} \frac{m}{2} + \frac{\sigma^2 \ln(T)}{m} = \Gamma.$$

From this, we can use the Gaussian likelihood formulas to obtain P_D and P_{FA} as:

$$P_D = 1 - \Phi(\frac{\Gamma - m}{\sigma}) = Q(\frac{\Gamma - m}{\sigma}); \quad P_{FA} = 1 - \Phi(\frac{\Gamma}{\sigma}) = Q(\frac{\Gamma}{\sigma}).$$

These calculations of P_D and P_F are illustrated in Figure 6.11.

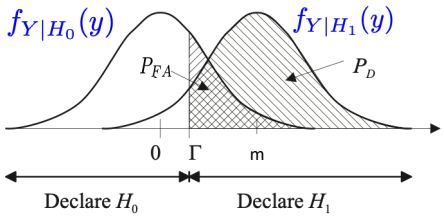


Figure 6.11: Illustration of P_D and P_{FA} calculation.

Example 6.14 (Gaussian detection with different variances)
Suppose y is scalar and distributed $N(0, \sigma_1^2)$ under H_0 and distributed $N(0, \sigma_0^2)$ under H_1. Assume $\sigma_1 < \sigma_0$. Thus, the Gaussians have the same mean, but different variances.

$$\mathcal{L}(y) = \frac{\frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(y)^2}{2\sigma_1^2}}}{\frac{1}{\sqrt{2\pi}\sigma_0} e^{-\frac{(y)^2}{2\sigma_0^2}}} = \frac{\sigma_1}{\sigma_0} e^{-\frac{y^2}{2\sigma_1^2} + \frac{y^2}{2\sigma_0^2}}.$$

The log-likelihood ratio is

$$\ln\Big(\mathcal{L}(y)\Big) = -\frac{y^2}{2}\left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_0^2}\right) + \ln(\sigma_1) - \ln(\sigma_0).$$

Hence, comparing the log-likelihood ratio to the log of a threshold T yields the decision rule

$$-y^2 \underset{H_0}{\overset{H_1}{\gtrless}} \frac{\sigma_1^2\sigma_0^2}{\sigma_0^2 - \sigma_1^2}\Big(\ln(\sigma_0) - \ln(\sigma_1) + \ln(T)\Big) = \Gamma.$$

Note we were careful in dividing by numbers that are positive, so the sign of the inequalities was preserved. Unlike the case where the means were different, the detector is quadratic in the measurement. Since the density of Y under H_0 has larger variance, higher magnitudes of the measured y provide more support for hypothesis H_0. We can simplify the decision rule: In terms of y, we select H_1 if $|y| \leq \sqrt{\Gamma}$, otherwise, we select H_0. From this, we can use the Gaussian likelihood formulas to obtain P_D and P_{FA} as:

$$P_D = \mathbb{P}[\{|Y| \leq \sqrt{\Gamma}\}|H_1] = \Phi\left(\frac{\sqrt{\Gamma}}{\sigma_1}\right) - \Phi\left(-\frac{\sqrt{\Gamma}}{\sigma_1}\right).$$

$$P_{FA} = \mathbb{P}[\{|Y| \leq \sqrt{\Gamma}\}|H_1] = \Phi\left(\frac{\sqrt{\Gamma}}{\sigma_0}\right) - \Phi\left(-\frac{\sqrt{\Gamma}}{\sigma_0}\right)$$

The ROC can now be obtained by varying Γ from 0 to ∞. The ROC is shown in Figure 6.12.

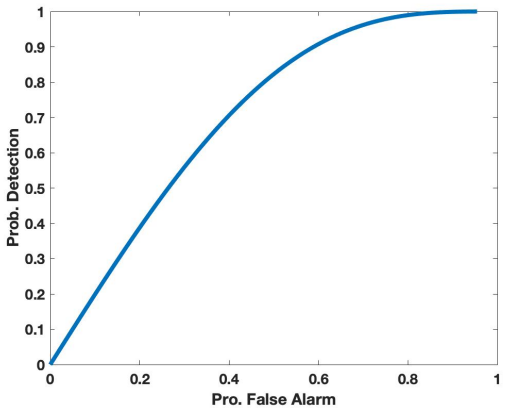


Figure 6.12: ROC for Gaussian hypotheses with different varlances.

6.6 Binary Hypothesis Testing with Vector Observations

The previous sections have assumed that the measurement Y is a scalar measurement, either discrete-valued or continuous valued. The previous theories developed extend completely to the case of pairs of measurements X, Y, or vector-valued measurements $\underline{Y}$. We briefly overview these extensions for pairs of measurements X, Y.

As before, we assume that the two hypotheses H_0, H_1 are events in $\mathcal{E}$ that form a partition of the sample space Ω. We assume that we observe a pair of random variables X, Y. If X, Y are discrete random variables, we assume that we are given the conditional joint probability mass functions $P_{X,Y|H_0}(x,y)$ and $P_{X,Y|H_1}(x,y)$. With this information, the likelihood ratio can be defined as a function of values $(x,y) \in R_{X,Y}$ by

$$\mathcal{L}(x,y) = \frac{P_{X,Y|H_1}(x,y)}{P_{X,Y|H_0}(x,y)}.$$

For jointly continuous measurements, the likelihoods are given by the conditional joint probability density functions $f_{X,Y|H_0}(x,y)$ and $f_{X,Y|H_1}(x,y)$. The likelihood ratio is defined as

$$\mathcal{L}(x,y) = \frac{f_{X,Y|H_1}(x,y)}{f_{X,Y|H_0}(x,y)}.$$

Once we have the likelihood ratios, the ML, MAP and MBR detectors are defined in identical manner to the scalar case:

$$D^{ML}(x,y) = \left\{ \mathcal{L}(x,y) \underset{H_0}{\overset{H_1}{\gtrless}} 1 \right\}.$$

$$D^{MAP}(x,y) = \left\{ \mathcal{L}(x,y) \underset{H_0}{\overset{H_1}{\gtrless}} \frac{\mathbb{P}[H_0]}{\mathbb{P}[H_1]} \right\}.$$

$$D^{ML}(x,y) = \left\{ \mathcal{L}(x,y) \underset{H_0}{\overset{H_1}{\gtrless}} \frac{C_{10}\mathbb{P}[H_0]}{C_{01}\mathbb{P}[H_1]} \right\}.$$

What is unique about the vector case is that the optimal decision rule depends only on a scalar function of the vector of observations X, Y. This holds true for higher-dimensional vectors: there is always a scalar function of the measurement vector $\underline{Y}$ that serves as a **sufficient statistic** to make an optimal decision.

The hard part of detection with vector observations is finding the decision regions so that we can compute performance metrics such as the probability of false alarm or the probability of missed detection. For pairs of random variables, we need to find the regions $\{(x,y) \in R_{X,Y} : D(x,y) = 0\}$ and $\{(x,y) \in R_{X,Y} : D(x,y) = 1\}$. For scalar measurements, we did this by analyzing the likelihood ratio test, and simplifying the equations to identify the regions. This is significantly harder for vector measurements, but there are special cases where we can do this.

We illustrate these extensions to vector observations with examples below.

Example 6.15

We are going to extend the diagnosis problem discussed in Example 6.7. The patient believes he has the flu. The hypothesis H_1 is the patient has the flu versus H_0 that the patient only has a cold. Let X be the measured temperature, and let Y be the results of a rapid influenza diagnostic test (RIDT) done on a mucus sample. We model the likelihood of X as a conditional Gaussian random variable with mean 98 degrees and standard deviation 2 degrees under H_0, and mean 102 degrees with standard deviation 2 degrees under H_1. The RIDT test is a color test, so we model the likelihood of Y in a very simple manner as a conditional Gaussian random variable (in the visible color spectrum) with mean wavelength 500 nm and standard deviation 100 nm under H_0, and mean wavelength 650 nm and standard deviation 100 nm under H_1. We assume that X, Y are conditionally independent given H_0, and also conditionally independent under H_1.

With the above information, we can now write the conditional joint probability density of (X,Y) given H_0 and H_1 as

$$f_{X,Y|H_0}(x,y) = \frac{1}{2\sqrt{2\pi}} e^{-\frac{(x-98)^2}{8}} \frac{1}{100\sqrt{2\pi}} e^{-\frac{(y-500)^2}{20000}}.$$

$$f_{X,Y|H_1}(x,y) = \frac{1}{2\sqrt{2\pi}} e^{-\frac{(x-102)^2}{8}} \frac{1}{100\sqrt{2\pi}} e^{-\frac{(y-650)^2}{20000}}.$$

The likelihood ratio is:

$$\begin{aligned}\mathcal{L}(x,y) &= \frac{f_{X,Y|H_1}(x,y)}{f_{X,Y|H_0}(x,y)} \\ &= \frac{\frac{1}{2\sqrt{2\pi}} e^{-\frac{(x-102)^2}{8}} \frac{1}{100\sqrt{2\pi}} e^{-\frac{(y-650)^2}{20000}}}{\frac{1}{2\sqrt{2\pi}} e^{-\frac{(x-98)^2}{8}} \frac{1}{100\sqrt{2\pi}} e^{-\frac{(y-500)^2}{20000}}} \\ &= e^{-\frac{(x-102)^2}{8}+\frac{(x-98)^2}{8}} e^{-\frac{(y-650)^2}{20000}+\frac{(y-500)^2}{20000}}\end{aligned}$$

Taking logarithms yields the log-likelihood ratio:

$$\begin{aligned}\ln\Big(\mathcal{L}(x,y)\Big) &= -\frac{(x-102)^2}{8} + \frac{(x-98)^2}{8} - \frac{(y-650)^2}{20000} + \frac{(y-500)^2}{20000} \\ &= \frac{(102-98)(2x-200)}{8} + \frac{150(2y-1150)}{20000} \\ &= x - \frac{4\cdot 200}{8} + \frac{3}{200}y - \frac{3\cdot 23}{8} \\ &= x + \frac{3}{200}y - \frac{869}{8}.\end{aligned}$$

The maximum likelihood detector compares the log-likelihood ratio to the threshold 0. This test becomes:

$$D^{ML}(x,y) = \Big\{x + \frac{3}{200}y \underset{H_0}{\overset{H_1}{\gtrless}} \frac{869}{8}\Big\}.$$

The decision rule reduces to comparing a scalar statistic $x + \frac{3}{20}y$ to a threshold. This defines a region in x-y space where the decision is 0, and another region where the decision is 1, separated by the line $x + \frac{3}{200}y = \frac{869}{8}$. With this definition of decision regions, we can now do compute P_{FA} and P_{MD} as two-dimensionl integrals.

In this case, there is a simpler method for computing performance. Define the statistic $Z = X + \frac{3}{200}Y$ as a linear combination of X, Y. Z is a **sufficient statistic** for this problem, because the max-likelihood detector depends only on Z: $D^{ML}(x,y) = \Big\{z \underset{H_0}{\overset{H_1}{\gtrless}} \frac{869}{8}\Big\}$.

Since X, Y are jointly Gaussian conditioned on H_0, Z is a Gaussian random variable conditioned on H_0. Its conditional mean is $\mathbb{E}[Z|H_0] = \mathbb{E}[X|H_0] + \frac{3}{200}\mathbb{E}[Y|H_0] = 98 + 7.5 = 105.5$. Since X and Y are conditionally independent given H_0, we get

$$\mathsf{Var}[Z|H_0] = \mathsf{Var}[X|H_0] + \left(\frac{3}{200}\right)^2 \mathsf{Var}[Y|H_0] = 4 + \frac{9}{40000}10000 = 6.25.$$

Similarly, Z is Gaussian conditioned on H_1 with $\mathbb{E}[Z|H_1] = 102 + \frac{3}{200}650 = 111.75$, and $\mathsf{Var}[Z|H_1] = 6.25$. We write the ML detector in terms of Z as

$$D^{ML}(z) = \Big\{z \underset{H_0}{\overset{H_1}{\gtrless}} 108.625\Big\}.$$

and now we can analyze its performance the same way we did for a scalar Gaussian random variable decision rule. Thus,

$$P_{FA} = Q\Big(\frac{108.625 - 105.5}{\sqrt{6.25}}\Big); \qquad P_{MD} = Q\Big(\frac{111.75 - 105.5}{\sqrt{6.25}}\Big).$$

Example 6.16
Consider the radar detection example, where N independent pulses are sent out. However, instead of making a detection on each pulse return and counting the number of detections, we measure the signal strength of each return, so that a vector of signal strength measurements is collected. We assume that each pulse provides a measurement Y_i, where

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

where E is a known constant, $W_i, i = 1, \ldots, N$ are independent, zero-mean Gaussian random variables with variance σ^2.

The above model results in a vector of observations $\underline{Y}$, where the components Y_i are jointly Gaussian and independent. Under hypothesis H_1, each Y_i has mean E and variance σ^2, whereas under hypothesis H_0, each Y_i has mean 0 and variance σ^2. In this case, the likelihood ratio is given by

$$\mathcal{L}(\underline{y}) = \frac{f_{\underline{Y}|H_1}(\underline{y})}{f_{\underline{Y}|H_0}(\underline{y})} = \prod_{i=1}^{N} \frac{e^{\frac{-(y_i-E)^2}{2\sigma^2}}}{e^{\frac{-(y_i)^2}{2\sigma^2}}} = \prod_{i=1}^{N} e^{\frac{2Ey_i-E^2}{2\sigma^2}} = e^{\frac{2E\left(\sum_{i=1}^{N} y_i\right) - NE^2}{2\sigma^2}}$$

Taking logs of both sides the decision rule can be reduced to:

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

where T is the threshold used in the likelihood ratio test (e.g. 1 for the maximum likelihood detector.) In this case, note that a scalar sufficient statistic is $Z = \frac{1}{N}\sum_{i=1}^{N} Y_i$, which is a linear combination of $\underline{Y}$ and hence Gaussian conditioned on H_0 and on H_1. The mean of Z under H_1 is

$$\mathbb{E}[Z|H_1] = \mathbb{E}[\sum_{i=1}^{N} \frac{Y_i}{N}|H_1] = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}[Y_i|H_1 = \frac{1}{N}\sum_{i=1}^{N} E = E.$$

Similarly, the mean of Z under H_0 is 0. The variance of Z under both H_1 and H_0 is

$$\mathsf{Var}[Z|H_1] = \mathsf{Var}[\sum_{i=1}^{N} \frac{Y_i}{N}|H_1] = \sum_{i=1}^{N} \mathsf{Var}[\frac{Y_i}{N}|H_1] = \sum_{i=1}^{N} \frac{\sigma^2}{N^2} = \frac{\sigma^2}{N},$$

because the Y_i components are independent (and thus uncorrelated), so the variance of the sum is the sum of the variances of the individual components.

Thus, the effect of using N measurements is equivalent to using one measurement with variance reduced by a factor of $1/N$, thereby increasing the effective signal-to-noise ratio in the detector. Denote by $\Gamma = \frac{E}{2} + \frac{\sigma^2 \ln(T)}{NE}$ as the threshold used in the log-likelihood ratio test for Z. We can now compute the performance statistics as a function of this threshold using the Gaussian properties of Z, as

$$P_{FA} = \mathbb{P}[Z > \Gamma|H_0] = Q(\frac{\Gamma N^{\frac{1}{2}}}{\sigma}).$$

$$P_{MD} = \mathbb{P}[Z < \Gamma|H_1] = Q(\frac{(E-\Gamma)N^{\frac{1}{2}}}{\sigma}).$$

The effect of increasing N is to get a more accurate measurement. This means the performance of the detector, as captured in the ROC curve, improves. As $N \to \infty$, both P_{FA} and P_{MD} decrease to zero. The ROC for different values of N is illustrated in Figure 6.13.

6.7 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 have M events in $(\Omega, \mathcal{E}, \mathbb{P})$, denoted as $H_i, i = 0, \ldots, M-1$, which are mutually exclusive and collectively exhaustive, so they form a partition of Ω. We assume there are measurements $\underline{Y}$ which are random vectors that provide the information for detection. If $\underline{Y}$ is discrete-valued, we are provided the conditional probability mass functions $P_{\underline{Y}|H_i}(\underline{y})$ for $i = 0, 1, \ldots, M-1$. If $\underline{Y}$ is a jointly continuous random vector, we are provided the conditional probability density functions $f_{\underline{Y}|H_i}(\underline{y})$ for $i = 0, 1, \ldots, M-1$.

A decision rule $D(\underline{y})$ is. function that maps each observed value $\underline{y}$ into $\{0, 1, \ldots, M-1\}$ where decision k means that hypothesis H_k is the selected hypothesis. The concepts for designing decision rules that we presented previously for binary hypothesis testing extend naturally to this case. For the maximum likelihood decision rule, we want to select $D(\underline{y}) = k$ whenever

$$P_{\underline{Y}|H_k}(\underline{y}) \geq P_{\underline{Y}|H_j}(\underline{y}), \text{ for all } j \neq i(\ \underline{y} \text{ discrete}).$$

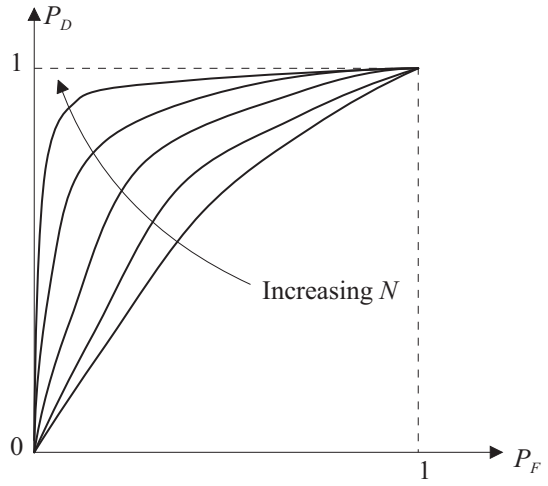


Figure 6.13: Illustration ROC behavior as we obtain more independent observations.

$$f_{\underline{Y}|H_k}(\underline{y}) \geq f_{\underline{Y}|H_j}(\underline{y}), \text{ for all } j \neq i(\ \underline{y} \text{ continuous}).$$

For the maximum a posteriori decision rule, we want to select $D(\underline{y}) = k$ whenever

$$\mathbb{P}[H_k|\underline{Y} = \underline{y}] \geq \mathbb{P}[H_j|\underline{Y} = \underline{y}], \text{ for all } j \neq i.$$

Equivalently, select $D(\underline{y}) = k$ whenever

$$P_{\underline{Y}|H_k}(\underline{y})\mathbb{P}[H_k] \geq P_{\underline{Y}|H_j}(\underline{y})\mathbb{P}[H_j], \text{ for all } j \neq i(\ \underline{y} \text{ discrete}).$$

$$f_{\underline{Y}|H_k}(\underline{y})\mathbb{P}[H_k] \geq f_{\underline{Y}|H_j}(\underline{y})\mathbb{P}[H_j], \text{ for all } j \neq i(\ \underline{y} \text{ continuous}).$$

As before, the MAP decision rule will minimize the average probability of error. If we defined costs C_{ij} associated with the cost of selecting decision U_i when hypothesis H_j is true, we can also define an equivalent theory for the minimum Bayes risk decision rule as in the binary hypothesis testing problems.

The biggest difference in the m-ary detection case is that there is no longer a sufficient scalar statistic like the likelihood ratio that we can compare to a threshold for optimal decision rules. Instead, the optimal decision rules must compute the M likelihoods, scale them appropriately, and pick the best decision on the basis of the resulting scaled values.

We illustrate m-ary detection problems with a couple of examples.

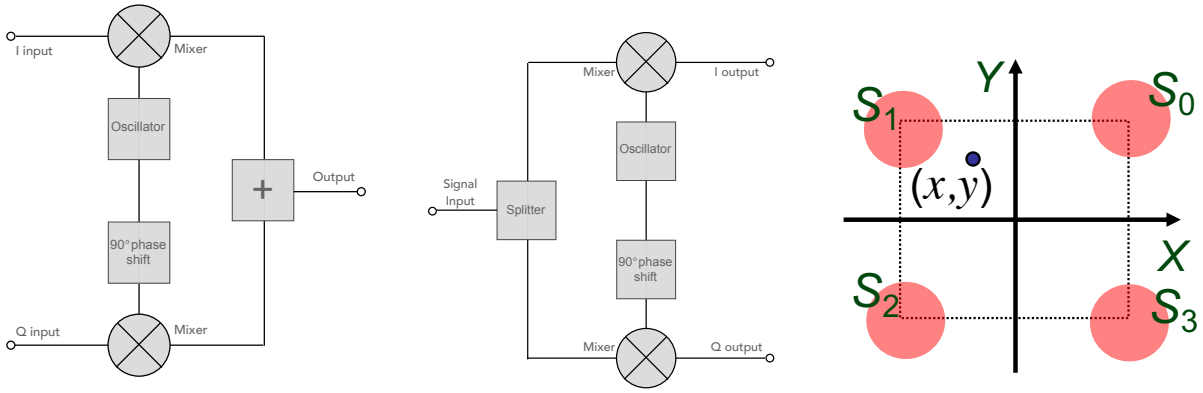
Example 6.17

Consider a communications problem where we try to communicate two bits at a time. We denote our two bits as pairs $A, B \in \{-1, 1\}$. We have four basic signals we are sending (1,1), (-1,1),(-1,-1),(1,-1), corresponding to hypotheses H_0, H_1, H_2, H_3 correspond to the transmitted symbols in this order.

To send the symbols, we use a variation of quadrature amplitude modulation, using short pulses of the form $s(t) = A\cos(\omega t) + B\sin(\omega t), t \in [0, T]$. A typical QAM modulation scheme is show in Figure 6.14, where the input I is the in-phase component, corresponding to the symbol A, and the input Q is the quadrature component, corresponding to the symbol B. The resulting transmitted pulse is $s(t) = A\cos(\omega t) + B\sin(\omega t), t \in [0, T]$

The signals propagate through the environment to a receiver, that demodulates the signal using a quadrature demodulation scheme, as shown in Figure 6.14. In the demodulator, the received signal is split, and multiplied each by $\cos(\omega t)$ and $\sin(\omega t)$. The in-phase output of the demodulator, $I(t)$, corresponds to the signal $s(t)\cos(\omega t)$, and the quadrature output $Q(t)$ corresponds to the signal $s(t)\sin(\omega t)$.

Note that $I(t) = A\cos^2(\omega t) + B\cos(\omega t)\sin(\omega t)$. Thus, averaging $I(t)$ over an interval of a few periods yields the output $A/2$, as the second term averages to 0. Similarly, $Q(t) = A\cos(\omega t)\sin(\omega t) + B\sin^2(\omega t)$, which averages to $B/2$. This



QAM modulator QAM demodulator Signal Constellation

Figure 6.14: Figures for Example **??**.

averaging is a low-pass filter and generates measurements of the transmitted A, B bits. Of course, these measurements are corrupted by background noise in the detector, signal corruption in the transmission channel, and small phase errors in the oscillators between the modulator and the demodulator.

The mathematical model of this detection problem is as follows: we measure two continuous random variables (X, Y), corresponding to the averages of the I and Q outputs of the demodulator. We model the statistics of these random variables as follows: We assume that X, Y are conditionally independent and Gaussian given any of the four hypotheses $H_i, i = 0, 1, 2, 3$. Furthermore, under each hypothesis, the variance of X is σ^2, and the variance of Y is also σ^2. However, the means change between hypotheses:

- Under H_0, $\mathbb{E}[X|H_0] \equiv m_x^0 = 1, \mathbb{E}[Y|H_0] \equiv m_y^0 = 1$.
- Under H_1, $\mathbb{E}[X|H_1] \equiv m_x^1 = -1, \mathbb{E}[Y|H_1] \equiv m_y^1 = 1$.
- Under H_2, $\mathbb{E}[X|H_2] \equiv m_x^2 = -1, \mathbb{E}[Y|H_2] \equiv m_y^2 = -1$.
- Under H_3, $\mathbb{E}[X|H_3] \equiv m_x^3 = 1, \mathbb{E}[Y|H_3] \equiv m_y^3 = -1$.

The signals are illustrated in Figure 6.14.

The likelihood under H_i is thus $f_{X,Y|H_i}(x, y) = \frac{1}{2\pi\sigma^2} e^{-\frac{(x-m_x^i)^2+(y-m_y)^2}{2\sigma^2}}$. To pick the largest one, we can compare the logarithms of the likelihoods, and subtract a common constant from all of them, to get a different comparison function $c^i(x, y)$ as

$$c^i(x, y) = \ln\left(f_{X,Y|H_i}(x, y)\right) - \ln\left(\frac{1}{2\pi\sigma^2}\right) = -\frac{(x - m_x^i)^2 + (y - m_y)^2}{2\sigma^2}.$$

We can scale $c^i(x, y)$ and subtract the same term to all i, to get

$$d^i(x, y) = 2\sigma^2 c^i(x, y) + \frac{x^2}{2} + \frac{y^2}{2} = m_x^i x + m_y^i y - \frac{(m_i^x)^2 + (m_i^y)^2}{2} = m_x^i x + m_y^i y - 1.$$

Every transformation we did above preserved the order of the likelihoods $f_{X,Y|H_i}(x, y)$. Hence, the maximum likelihood decision is

$$D^{ML}(x, y) = U_{i^*}, \text{ where } i^* \in \arg\max_{i=0,1,2,3} m_x^i x + m_y^i y - 1.$$

Note the -1 is not important. Then, we decide 0 when: $x + y > x - y; x + y > -x - y; x + y > -x + y$. Combine these inequalities, we get the region $x > 0, y > 0$. Thus, we decide 0 if we measure x, y in the first quadrant. Similarly, we decide 1 if the measurement (x, y) is in $x < 0, y > 0$, U_2 if the measurement is in $x, y < 0$, and U_3 if $x > 0, y < 0$.

We've simplified the decision rule so we could identify the decision regions in terms of the regions of the measurement range $R_{X,Y}$. We can use this to analyze the performance. Note the following: we can compute $\mathbb{P}[D^{ML}(X, Y) = 0|H_0] = \mathbb{P}[X \geq 0, Y \geq 0|H_0] = \mathbb{P}[X \geq 0|H_0]\mathbb{P}[Y \geq 0|H_0]$ because of the conditional independence of X, Y. Thus,

$$\mathbb{P}[D^{ML}(X, Y) = 0|H_0] = \Phi(\frac{1}{\sigma})\Phi(\frac{1}{\sigma}) = \Phi(\frac{1}{\sigma})^2.$$

This is the probability that we don't make an error when H_0 is the correct hypothesis. By symmetry, this is also $\mathbb{P}[D^{ML}(X,Y) = U_i|H_i], i = 1,2,3$. If all the hypotheses had equal prior probability $\mathbb{P}[H_i]$, the expected probability of correct decoding is $\Phi(\frac{1}{\sigma})^2$.

Example 6.18
Suppose we want to detect which of three possible N-dimensional signals $\underline{m}^k, k = 0,1,2$ is being received in the presence of noise. Under hypothesis H_k the observation $\underline{Y}$ is given by the vector Gaussian density:

$$f_{\underline{Y}|H_k}(\underline{y}) = \prod_{j=1}^{N} \frac{1}{\sqrt{2\pi}} e^{-\frac{(y_j - m_j^k)^2}{2}}.$$

This means each of the components of the observation $\underline{Y}$ is conditionally independent, Gaussian, with variance 1 and conditional expectation given by the components of $\underline{m}^k$.

Assume that we want a minimum probability of error decision rule, namely the MAP decision rule. Let $P_k = \mathbb{P}[H_k]$, the prior probabilities. The MAP decision rule picks

$$D^{MAP}(\underline{y}) = U_{i^*}, \text{ where } i^* \in \arg\min_{k=0,1,2} P_k \prod_{j=1}^{N} \frac{1}{\sqrt{2\pi}} e^{-\frac{(y_j - m_j^k)^2}{2}}.$$

We now have a valid decision rule, but the decision regions are hard to figure out, and this requires a lot of computation. We simplify the decision rules by taking transformations that preserve the order: we first compute log-likelihoods, and subtract common constants, to define

$$c^k(\underline{y}) = \ln\left(f_{\underline{Y}|H_k}(\underline{y})\right) + \ln(P_k) - N\ln(\frac{1}{\sqrt{2\pi}}) = \ln(P_k) - \sum_{j=1}^{n} \frac{(y_j - m_j^k)^2}{2}.$$

We can further simplify this by adding the same term to all the $c^k(\underline{y})$, as

$$d^k(\underline{y}) = c^k(\underline{y}) + \sum_{j=1}^{N} \frac{y_j^2}{2} = \ln(P_k) + (\underline{m}^k)^T\underline{y} - \frac{1}{2}(\underline{m}^k)^T\underline{m}^k,$$

where we have used vector notation for transposes. The terms $d^k(\underline{y})$ are referred to as discriminant functions; in this case, they are linear functions of $\underline{y}$, which help establish the regions.

Thus, the decision 0 is made whenever

$$\ln(P_0) + (\underline{m}^0)^T\underline{y} - \frac{1}{2}(\underline{m}^0)^T\underline{m}^0 > \ln(P_1) + (\underline{m}^1)^T\underline{y} - \frac{1}{2}(\underline{m}^1)^T\underline{m}^1,$$

$$\ln(P_0) + (\underline{m}^0)^T\underline{y} - \frac{1}{2}(\underline{m}^0)^T\underline{m}^0 > \ln(P_2) + (\underline{m}^2)^T\underline{y} - \frac{1}{2}(\underline{m}^2)^T\underline{m}^2.$$

Combining the $\underline{y}$ terms on the left side of the first equation, we get:

$$(\underline{m}^0 - \underline{m}^1)^T\underline{y} > \ln(P_1) - \ln(P_0) + \frac{1}{2}\left((\underline{m}^0)^T\underline{m}^0 - (\underline{m}^1)^T\underline{m}^1\right)$$

which defines a half-plane perpendicular to the line connecting $\underline{m}^0$ and $\underline{m}^1$. Working with the second equation yields

$$(\underline{m}^0 - \underline{m}^2)^T\underline{y} > \ln(P_2) - \ln(P_0) + \frac{1}{2}\left((\underline{m}^0)^T\underline{m}^0 - (\underline{m}^2)^T\underline{m}^2\right)$$

which is another half plane perpendicular to the line connecting $\underline{m}^0$ and $\underline{m}^2$. The intersection of the two half-planes is the region of $\underline{y}$ where we decide 0. A similar analysis can be done to determine the regions for 1 and U_2.

It is worth noting that, if the prior probabilities are all equal to 1/3, then the half-plane separating $\underline{m}^0$ and $\underline{m}^1$ goes through the midpoint of the line connecting $\underline{m}^0$ and $\underline{m}^1$. This is because, setting $\underline{y} = \frac{\underline{m}^0 + \underline{m}^1}{2}$, we get

$$(\underline{m}^0)^T\underline{y} - \frac{1}{2}(\underline{m}^0)^T\underline{m}^0 = (\underline{m}^0)^T\underline{m}^1.$$

$$(\underline{m}^1)^T\underline{y} - \frac{1}{2}(\underline{m}^1)^T\underline{m}^1 = (\underline{m}^0)^T\underline{m}^1.$$

Thus, this value of $\underline{y}$ is on the boundary of the decision regions between $0,1$. The resulting decision regions are illustrated in Figure 6.15 for a two-dimensional case. The decision boundaries are the bisectors of the lines connecting the means

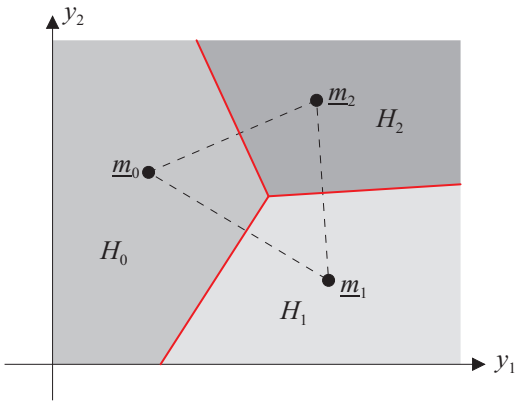


Figure 6.15: Illustration of the ML decision rule in the observation space.

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. Given the decision regions, we can now calculate performances, albeit with complicated integrals even in the case where we have conditionally independent measurements, because the decision regions are not parallel to the y_1, y_2 axes.

As a final comment in this Chapter, techniques such as nearest neighbor classifiers and linear discriminants are used extensively in data science and machine learning without much theoretical justification. In this Chapter, we have learned classes of statistical models for which nearest average classifiers and linear discriminants lead to optimal decision rules. We will use this to understand the hidden assumptions behind many classification methods in data science.