Jerónimo Arenas-García, Jesús Cid-Sueiro, Vanessa Gómez-Verdejo, Miguel Lázaro-Gredilla, and David Ramírez

Estimation and Detection Theory

Year 2021-22

February 17, 2025

Universidad Carlos III de Madrid

Chapter 1
Stochastic Processes

Chapter 2
Statistical Estimation Theory

Chapter 3
Linear Filtering

Chapter 4
Spectral Estimation

Chapter 5
Statistical Decision Theory

Contents

5.1 Introduction to Decision Theory

In this section we provide a formal presentation of decision theory. Appendix B provides some introductory examples.

Decision theory (also named **detection theory** or **hypothesis testing**) is a mathematical framework used to make optimal choices under conditions of uncertainty. It employs models and statistical analysis to evaluate and compare the outcomes of different decisions, aiming to identify the most advantageous option based on established criteria. This field utilizes concepts from statistics and economics to aid in strategic planning and risk management by calculating the probabilities and impacts of potential scenarios. Decision theory focuses on maximizing benefits and minimizing costs or risks, providing a structured approach to rational decision-making. Through precise quantitative methods, it guides individuals and organizations in policy formulation and decision implementation.

5.1.1 Hypotheses-based problems

In this course, we will only cover a particular class of detection or classification problems to which we will refer as *hypotheses-based problems*. The goal is to infer the correct hypothesis, which cannot be directly observed, from a set of measurements or observations. Thus, we consider a scenario with M hypotheses, and denote the random variable that identifies the hypothesis as H. This is depicted in Fig. 5.1, where $H \in \{0,1,\ldots,M-1\}$. We also assume that we have access to an observation vector $\mathbf{x}$, which can be considered as the realization of a random variable $\mathbf{X}$ lying in the observation space $\mathscr{X}$. We assume also that there is a certain statistical relationship between H and $\mathbf{X}$. Otherwise, i.e., if H and $\mathbf{X}$ were independent, it would make no sense to use $\mathbf{x}$ to make an informed inference about the value of H.

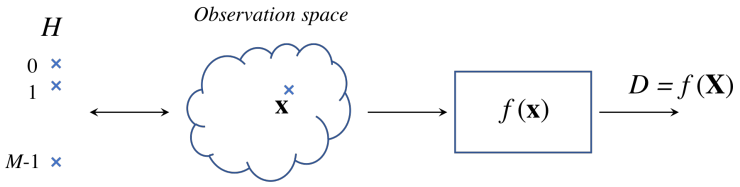


Fig. 5.1 Diagram block of hypothesis testing problems.

In this context, a detector is a function of $\mathbf{x}$ that outputs a value d in the range $\{0,1,\ldots M-1\}$, i.e., a guess on the value of the hypothesis that is unknown beforehand. Depending on the application scenario, the detector receives another names, like decision-maker or classifier. In this chapter we will take these terms as synonymous.

We should make a few considerations about the functions $f(\mathbf{x})$ that we admit as valid detectors in this course:

- We consider that $d = f(\mathbf{x})$ is a deterministic function. This implies that if the same vector is presented several times, the function will output the same value each time. Note that,

even though $f(\cdot)$ is deterministic, its output can be modeled as a random variable since the input is the random vector $\mathbf{X}$.

- The function is surjective, that is, every input $\mathbf{x}$ generates one and only one output, but several inputs could generate the same output. Hence, the function divides the observation space into M non-overlapping regions, $\mathscr{X}_d$, $d = 0, 1, \ldots M-1$, i.e., one region per hypotheses. The boundaries between regions are known as *decision boundaries*.

Example 5.1 The detector $f(x) = u(x^2 - 1)$, where $u(\cdot)$ is the step function, is defined for any x on the real line, and is characterized by the following decision regions:

$$\mathscr{X}_0 = \left\{x \in \Re | x^2 - 1 < 0\right\} = (-1, 1),$$
$$\mathscr{X}_1 = \left\{x \in \Re | x^2 - 1 \geq 0\right\} = (-\infty, -1] \cup [1, \infty).$$

where we have assumed $u(0) = 1$. In this example, the regions are connected and non-empty.

Example 5.2 The detector $f(\mathbf{x}) = \arg\min_i y_i(\mathbf{x})$ defined over $\mathscr{X} = [0,1]^2$, with

$$y_0(\mathbf{x}) = \|\mathbf{x}\|^2,$$
$$y_1(\mathbf{x}) = x_1 - x_0 + 1,$$
$$y_2(\mathbf{x}) = x_0 - x_1 + 1,$$

is characterized by the decision regions depicted in Fig. 5.2.

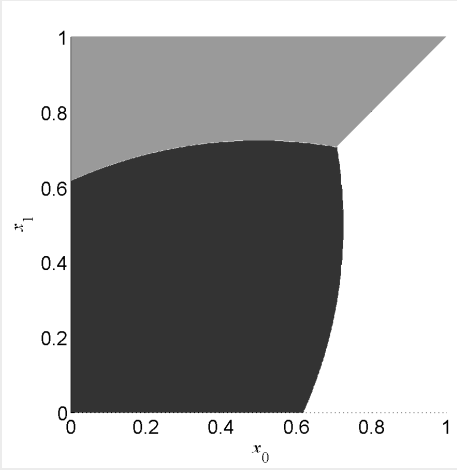


Fig. 5.2 Decision regions for the detector given in Example 5.2: $\mathscr{X}_0$ (black), $\mathscr{X}_1$ (grey), and $\mathscr{X}_2$ (white).

5.1.2 Modeling uncertainty

We review now the main distributions that will be employed in detection problems:

- *A priori* probability distribution of the hypotheses: This is a discrete distribution that quantifies the probability of each hypothesis independently of the observations. If we did not have access to any observations, our design would have to rely entirely on these probabilities, as it was the case in Section B.1.1,

$$P_H(h), \qquad \text{for} \quad h = 0, 1, \ldots, M-1.$$

- Likelihoods of the hypotheses: This represents the probability of the observations given the hypothesis. Note that, even though we refer to these distribution as the likelihoods of the hypotheses, what we actually have is a collection of distributions over the random variable X (unidimensional case) or $\mathbf{X}$ (multidimensional case), one for each hypothesis,

$$p_{\mathbf{X}|H}(\mathbf{x}|h) \qquad \text{for} \quad \mathbf{x} \in \mathscr{X} \text{ and } h = 0, 1, \ldots, M-1,$$

  where we have assumed a multidimensional case with continuous observations. Note that random variable $\mathbf{X}$ may lie in different regions depending on the hypothesis.
- *A posteriori* distribution of the hypotheses: This distribution provides information about the probabilities of the hypothesis, but conditioning them on each possible value of the observation vector

$$P_{H|\mathbf{X}}(h|\mathbf{x}), \qquad \text{for} \quad h = 0, 1, \ldots, M-1.$$

  Since designing a detector consists in deciding what should be the decision for each value of the observation vector, and this distribution expresses directly what are the probabilities of the hypothesis conditioned on every $\mathbf{x}$, *a posteriori* probabilities play a fundamental role for the statistical design of detectors.

  A priori and *a posteriori* probabilities are related by Bayes' Theorem, which states

$$P_{H|\mathbf{X}}(h|\mathbf{x}) = \frac{p_{\mathbf{X}|H}(\mathbf{x}|h) P_H(h)}{p_{\mathbf{X}}(\mathbf{x})}.$$

Bayes' Theorem shows how observing $\mathbf{x}$ modifies the information about the probabilities of the different hypotheses. Without them, we could only use $P_H(h)$ to make decisions. However, once the observation vector comes into play, a more accurate estimation of these probabilities can be achieved via $P_{H|\mathbf{X}}(h|\mathbf{x})$, and these probabilities can be used to obtain a more informed decision. Note also that if we know both the *a priori* probabilities of the hypothesis and their likelihoods, the joint distribution of $\mathbf{X}$ and H can be calculated. This joint distribution is the most complete characterization of the random variables, and from it any other probability function can be calculated as well.

In the following, we consider two different kinds of problems involving M-ary hypothesis testing problems:

- Analysis of detectors: Here, the detector is given, and the objective is to analyze its performance with respect to certain performance metrics.
- Detector design: The goal is to build a function $f(\mathbf{x})$ to optimize a desired performance metric.

5.2 Performance metrics

The first problem that we consider is the evaluation of the performance of a given detector. In this section, we review different metrics that can be used to assess performance. In all cases, we consider first the multiple hypothesis test scenario, and afterwards we specialize it to the binary case.

5.2.1 Probability of error

The probability of error is the probability of a wrong decision, i.e., the output of the statistic is not equal to the actual hypothesis. Under a frequentist approach, this probability can be interpreted as the average number of experiments in which an incorrect decision is taken, when the number of experiments tends to infinity. However, since we are assuming that the statistical characterization of the problem is available through the different probability distributions that we just reviewed, the probability of error can be calculated in closed-form as:

$$\begin{aligned} P_e = P(D \neq H) &= 1 - P(D = H) \\ &= 1 - \sum_{h=0}^{M-1} P(D = h, H = h) \\ &= 1 - \sum_{h=0}^{M-1} P(D = h | H = h) P_H(h) \\ &= 1 - \sum_{h=0}^{M-1} P_H(h) \int_{\mathscr{X}_h} p_{\mathbf{X}|H}(\mathbf{x}|h) d\mathbf{x}, \end{aligned}$$

where we have exploited that the probability of error is one minus the probability of correct decision. This is, in most cases, more convenient since the number of combinations where D and H are equal is (much) smaller than the number of combinations where they differ. Moreover, the last line of the previous expression follows from

$$P(D = h | H = h) = P(\mathbf{x} \in \mathscr{X}_d | H = h) = \int_{\mathscr{X}_h} p_{\mathbf{X}|H}(\mathbf{x}|h) d\mathbf{x},$$

which states that, conditioned on $H = h$, the probability of $D = h$ is precisely the integral of the likelihood of that hypothesis in the region where the given detector decides in favor of hypothesis h, i.e., the region $\mathscr{X}_h$.

Finally, note that it is also possible to compute the probability of error for a particular observation vector $\mathbf{x}$. If $\mathbf{x}$ belongs to $\mathscr{X}_d$, the associated probability of error would be

$$P(H \neq d | \mathbf{x}) = 1 - P(H = d | \mathbf{x}) = 1 - P_{H|\mathbf{X}}(d|\mathbf{x}) = \sum_{\substack{l=0 \\ l \neq d}}^{M-1} P_{H|\mathbf{X}}(l|\mathbf{x}) \tag{5.1}$$

In other words, the probability of error at a particular $\mathbf{x} \in \mathscr{X}_d$ is the sum of the *a posteriori* probabilities of hypothesis different from d conditioned on this particular observation. For instance, imagine that in a three-hypothesis testing problem for a given $\mathbf{x}_o$ a detector selects hypothesis 0. Then, the probability of error for $\mathbf{x}_o$ is the sum of the probabilities of hypothesis 1 and 2 conditioned on $\mathbf{X} = \mathbf{x}_o$, i.e., the sum of *a posteriori* probabilities $P_{H|\mathbf{X}}(1|\mathbf{x}_o)$ and $P_{H|\mathbf{X}}(2|\mathbf{x}_o)$.

5.2.1.1 Binary case: P_e, P_{FA}, P_{M} and P_{D}

For the binary case, contrary to the multiple hypotheses test, computing the probability of error involves as many terms as the probability of a correct decision since

$$\begin{aligned} P_e &= P(D=0,H=1) + P(D=1,H=0) \\ &= P(D=0|H=1)P_H(1) + P(D=1|H=0)P_H(0). \end{aligned}$$

In the expression above we find two terms that are normally referred to as the *probability of false alarm* (also known as probability of Type I error or significance level) and the *probability of missing* (or probability of Type II error):

$$P_{\mathrm{FA}} = P(D=1|H=0) = \int_{\mathscr{X}_1} p_{\mathbf{X}|H}(\mathbf{x}|0)d\mathbf{x},$$
$$P_{\mathrm{M}} = P(D=0|H=1) = \int_{\mathscr{X}_0} p_{\mathbf{X}|H}(\mathbf{x}|1)d\mathbf{x}.$$

Similarly, the probability of detection (power or sensitivity) is defined as

$$P_{\mathrm{D}} = P(D=1|H=1) = 1 - P_{\mathrm{M}},$$

and

$$P(D=0|H=0) = 1 - P_{\mathrm{FA}},$$

is the specificity. Using these definitions, the probability of error can now be expressed more compactly as

$$P_e = P_{\mathrm{M}}P_H(1) + P_{\mathrm{FA}}P_H(0).$$

Interestingly, for the computation of P_{FA} and P_{M}, only likelihoods are required. However, in order to compute the overall probability of error, we also need to know the *a priori* probabilities of the hypothesis.

5.2.2 Receiver Operating Characteristic (ROC)

We also introduce here an important concept for the analysis of binary hypothesis tests: the receiver operating characteristic (ROC) curve. The ROC curve plots the probability of false alarm, P_{FA}, against the probability of detection, P_{D} for different values of some parameter. Figure 5.3 shows the ROC curves of two different detectors, Detector 1 and Detector 2. As can be seen in this figure, the performance of Detector 2 is clearly better than that of

Detector 1, since for each P_{FA}, the P_{D} of Detector 2 is equal or larger than that of Detector 1. Moreover, both detectors perform better than a random decision whose ROC curve is also shown in the figure. One final comment is in order. For almost all detectors it is not possible to increase the probability of detection without increasing the probability of false alarm.

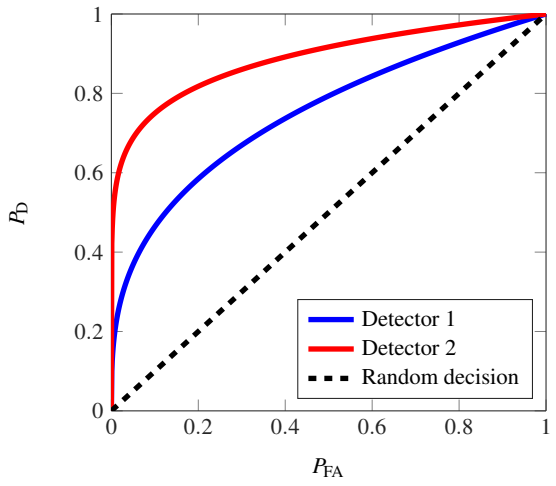


Fig. 5.3 ROC curves for two different detectors.

5.2.3 Risk

In scenarios where the consequences of each type of error are different, the probability of error is not an adequate performance measure. Imagine, for instance, a detector that discriminates whether there are or not suspicious tumor masses in a medical image. Such detector is used as a pre-diagnosis system, so that patients that can have a tumor are then explored with more accurate (but also invasive) techniques. In this case, there is a clear asymmetry between both kinds of errors: The incorrect decision that tumor masses are present would result in an unnecessary biopsy and inconvenience for the patient, but the opposite error could delay the diagnosis until a time when the process is irreversible.

To assign a penalty to different kinds of errors, we can define a cost function

$$c_{DH}, \qquad D,H = 0,\ldots,M-1.$$

Such function will take as many values as combinations of decisions and hypotheses, in such a way that each particular value c_{dh} is the cost of deciding $D = d$ when hypothesis $H = h$ is the true one. As already pointed out, we assume deterministic costs in this course, in the sense that the cost for each particular d and h is fixed. However, since the cost is a function of the random variables D and H, it is a random variable. Given a detector $D = \phi(\mathbf{X})$, we define the **risk** R_ϕ as the expected value of the cost,

$$R_\phi = \mathbb{E}\{c_{DH}\} = \sum_{h=0}^{M-1}\sum_{d=0}^{M-1} c_{dh} P_H(h) P_{D|H}(d|h). \tag{5.2}$$

Since $\phi(\mathbf{x}) = d$ when the observation belongs to the decision region of $D = d$, that is, $\mathbf{x} \in \mathscr{X}_d$, the conditional probabilities $P_{D|H}(d|h)$ can be calculated as

$$P_{D|H}(d|h) = P\{X \in \mathscr{X}_d | H = h\} = \int_{\mathscr{X}_d} p_{\mathbf{X}|H}(\mathbf{x}|h) d\mathbf{x} \tag{5.3}$$

therefore

$$R_\phi = \sum_{h=0}^{M-1} P_H(h) \sum_{d=0}^{M-1} c_{dh} \int_{\mathbf{x}\in\mathscr{X}_d} p_{\mathbf{X}|H}(\mathbf{x}|h) d\mathbf{x}. \tag{5.4}$$

Example 5.3 Consider a multiclass decision problem with three hypotheses whose likelihoods are:

$$\begin{array}{ll} p_{X|H}(x|0) = 1 & 0 < x < 1 \\ p_{X|H}(x|1) = 2(1-x) & 0 < x < 1 \\ p_{X|H}(x|2) = 2x & 0 < x < 1 \end{array}$$

knowing that the prior probabilities of the hypotheses are: $P_H(0) = 0.4$ and $P_H(1) = P_H(2) = 0.3$, and the cost policy is given by $c_{hh} = 0$, $h = 0,1,2$ and $c_{hd} = 1$, $h \neq d$. Obtain the risk of the decision-maker:

$$\phi(x) = \begin{cases} 1, & x < 0.5 \\ 2, & x > 0.5 \end{cases}$$

Applying the expression (5.2) to this problem we have:

$$\begin{aligned} R_\phi = {} & c_{10} P_H(0) P_{D|H}(1|0) + c_{20} P_H(0) P_{D|H}(2|0) + c_{01} P_H(1) P_{D|H}(0|1) \\ & + c_{21} P_H(1) P_{D|H}(2|1) + c_{02} P_H(2) P_{D|H}(0|2) + c_{12} P_H(2) P_{D|H}(1|2) \end{aligned}$$

where the terms $P_{D|H}(d|h)$ can be calculated using (5.3)

$$\begin{array}{l} P_{D|H}(0|1) = P_{D|H}(0|2) = 0 \\ P_{D|H}(1|0) = \int_{\mathscr{X}_1} p_{X|H}(x|0)\, dx = \int_0^{0.5} 1\, dx = 0.5 \\ P_{D|H}(2|0) = \int_{\mathscr{X}_2} p_{X|H}(x|0)\, dx = \int_{0.5}^{1} 1\, dx = 0.5 \\ P_{D|H}(1|2) = \int_{\mathscr{X}_1} p_{X|H}(x|2)\, dx = \int_0^{0.5} 2x\, dx = 0.25 \\ P_{D|H}(2|1) = \int_{\mathscr{X}_2} p_{X|H}(x|1)\, dx = \int_{0.5}^{1} 2(1-x)\, dx = 0.25 \end{array}$$

and substituting these, we arrive at

$$R_\phi = 0.4 \cdot 0.5 + 0.4 \cdot 0.5 + 0.3 \cdot 0.25 + 0.3 \cdot 0.25 = 0.55$$

Finally, we define the *conditional risk* as the expected cost conditioned on a given value of $\mathbf{x}$, $\mathbb{E}\{c_{dH}|\mathbf{x}\}$. Taking into account that, for a given $\mathbf{x}$ and a given detector, the decision value is fixed, it is only required to take expectations with respect to such hypothesis. The conditional risk for a given observation $\mathbf{x} \in \mathscr{X}_d$ is given by

$$\mathbb{E}\{c_{dH}|\mathbf{x}\} = \sum_{h=0}^{M-1} c_{dh} P_{H|X}(h|\mathbf{x}). \tag{5.5}$$

Example 5.4 Continuing with Example 5.3, the conditional risk of each decision can be calculated as:

$$\mathbb{E}\{c(d,H)|x\} = c_{d0}P_{H|X}(0|x) + c_{d1}P_{H|X}(1|x) + c_{d2}P_{H|X}(2|x)$$

where the posterior distributions can be obtained by applying Bayes' Theorem:

$$P_{H|X}(0|x) = \frac{p_{X|H}(x|0)P_H(0)}{\sum_{h=0}^{2} p_{X|H}(x|h)P_H(h)} = \frac{1\cdot 0.4}{1\cdot 0.4 + 2(1-x)\cdot 0.3 + 2x\cdot 0.3} = 0.4$$

$$P_{H|X}(1|x) = \frac{p_{X|H}(x|1)P_H(1)}{\sum_{h=0}^{2} p_{X|H}(x|h)P_H(h)} = \frac{2(1-x)\cdot 0.3}{1} = 0.6(1-x)$$

$$P_{H|X}(2|x) = \frac{p_{X|H}(x|2)P_H(2)}{\sum_{h=0}^{2} p_{X|H}(x|h)P_H(h)} = \frac{2x\cdot 0.3}{1} = 0.6x$$

This leads to:

- if $d = 0$:
$$\begin{aligned}\mathbb{E}\{c(0,H)|x\} &= c_{00}P_{H|X}(0|x) + c_{01}P_{H|X}(1|x) + c_{02}P_{H|X}(2|x)\\ &= 0\cdot 0.4 + 1\cdot 0.6(1-x) + 1\cdot 0.6x = 0.6\end{aligned}$$
- if $d = 1$:
$$\begin{aligned}\mathbb{E}\{c(1,H)|x\} &= c_{10}P_{H|X}(0|x) + c_{11}P_{H|X}(1|x) + c_{12}P_{H|X}(2|x)\\ &= 1\cdot 0.4 + 0\cdot 0.6(1-x) + 1\cdot 0.6x = 0.4 + 0.6x\end{aligned}$$
- if $d = 2$:
$$\begin{aligned}\mathbb{E}\{c(2,H)|x\} &= c_{20}P_{H|X}(0|x) + c_{21}P_{H|X}(1|x) + c_{22}P_{H|X}(2|x)\\ &= 1\cdot 0.4 + 1\cdot 0.6(1-x) + 0\cdot 0.6x = 1 - 0.6x\end{aligned}$$

5.2.3.1 Binary case: risk

For the binary case, a simpler expression can be obtained in terms of P_{FA}, P_{M}, and P_{D} as follows

$$\begin{aligned}R_\phi &= c_{00}P(D=0,H=0) + c_{01}P(D=0,H=1)\\ &= c_{00}P(D=0|H=0)P_H(0) + c_{01}P_{\text{M}}P_H(1) + c_{10}P_{\text{FA}}P_H(0) + c_{11}P_{\text{D}}P_H(1).\\ &= c_{00}(1-P_{\text{FA}})P_H(0) + c_{01}P_{\text{M}}P_H(1) + c_{10}P_{\text{FA}}P_H(0) + c_{11}(1-P_{\text{M}})P_H(1).\\ &= (c_{00}P_H(0) + c_{11}P_H(1)) + (c_{01}-c_{11})P_{\text{M}}P_H(1) + (c_{10}-c_{00})P_{\text{FA}}P_H(0)\end{aligned} \tag{5.6}$$

The previous expression shows that the risk of a decision-maker is the sum of three components:

- $(c_{00}P_H(0) + c_{11}P_H(1))$ is the minimum risk of the ideal decision-maker, the one with $P_{\text{M}} = 0$ and $P_{\text{FA}} = 0$ who succeeds with probability 1.
- $(c_{01} - c_{11})P_H(1)P_{\text{M}}$ is the increase in risk caused by miss errors.
- $(c_{10} - c_{00})P_H(0)P_{\text{FA}}$ is the increase in risk caused by false alarms.

Note that the ideal decision-maker is, in general, unachievable, because if the likelihoods of the hypotheses overlap, it is not possible to avoid errors. The optimal decision-maker will be the one who finds a good compromise between miss errors and false alarms, such that the risk in (5.6) is minimized.

5.3 Detector design

Once we have studied different ways of analyzing the performance of a given detector, we turn our attention to the problem of designing detectors that maximize one of these performance metrics.

5.3.1 Maximum likelihood and maximum *a posteriori* detectors

A first possibility would be to rely directly on the maximization of the available probability density functions:

- The detector that maximizes the likelihood is known as the *maximum likelihood* (ML) detector:
$$d_{ML} = \arg\max_h p_{\mathbf{X}|H}(\mathbf{x}|h).$$
- The detector that selects the hypothesis with maximum *a posteriori* probability is known as the maximum *a posteriori* (MAP) detector:
$$d_{MAP} = \arg\max_h P_{H|\mathbf{X}}(h|\mathbf{x}).$$

These detectors proceed as follows. Designing a detector is equivalent to specifying a unique decision for each possible value of the observation vector $\mathbf{x}$. Then, the ML and MAP strategies are based on evaluating either the likelihoods or the *a posteriori* probabilities for each $\mathbf{x}$ in the observation space, and select, for each $\mathbf{x}$, the hypothesis that maximizes $p_{\mathbf{X}|H}(\mathbf{x}|h)$ (ML) or $P_{H|\mathbf{X}}(h|\mathbf{x})$ (MAP).

Finally, there are two properties that are worth considering with respect to these detectors:

1. When the *a priori* probabilities of the hypothesis are the same, i.e., $P_H(h) = 1/M$, the ML and MAP detectors are identical. This can be shown from the Bayes' Theorem, since in this case
$$d_{MAP} = \arg\max_h P_{H|\mathbf{X}}(h|\mathbf{x}) = \arg\max_h \frac{p_{\mathbf{X}|H}(\mathbf{x}|h)P_H(h)}{p_{\mathbf{X}}(\mathbf{x})} = \arg\max_h p_{\mathbf{X}|H}(\mathbf{x}|h) = d_{ML}.$$
2. The MAP detector minimizes the probability of error. Note that according to (5.1) the probability of error for a given $\mathbf{x}$ can be expressed as
$$P(D \neq H|\mathbf{x}) = 1 - P_{H|\mathbf{X}}(h|\mathbf{x}).$$

Since the MAP detector selects for every $\mathbf{x}$ the hypothesis that maximizes $P_{H|\mathbf{X}}(h|\mathbf{x})$, it therefore minimizes the probability of error for each vector of the observation space. Thus, as the probability of error is minimized for each point of the observation space, it is also minimized overall. That is,

$$P(D \neq H) = \int_{\mathscr{X}} P(D \neq H|\mathbf{x}) p_{\mathbf{X}}(\mathbf{x}) \mathbf{dx},$$

and we can check that the value of the integral (the probability of error) is minimized if, for each $\mathbf{x}$, the decisions minimize $P(D \neq H|\mathbf{x})$, i.e., the MAP detector.

5.3.1.1 Binary case: ML and MAP detectors

The expressions of the ML and MAP detectors become fairly simple for the binary case:

- Maximum likelihood detector:

$$p_{\mathbf{X}|H}(\mathbf{x}|1) \underset{D=0}{\overset{D=1}{\gtrless}} p_{\mathbf{X}|H}(\mathbf{x}|0),$$

which can be expressed as a *likelihood ratio test* (LRT)

$$\frac{p_{\mathbf{X}|H}(\mathbf{x}|1)}{p_{\mathbf{X}|H}(\mathbf{x}|0)} \underset{D=0}{\overset{D=1}{\gtrless}} 1,$$

where we have taken into account that the likelihoods are non-negative. Sometimes, it will be more convenient to work with the *log-likelihood ratio test* (LLRT)

$$\log\left[\frac{p_{\mathbf{X}|H}(\mathbf{x}|1)}{p_{\mathbf{X}|H}(\mathbf{x}|0)}\right] = \log p_{\mathbf{X}|H}(\mathbf{x}|1) - \log p_{\mathbf{X}|H}(\mathbf{x}|0) \underset{D=0}{\overset{D=1}{\gtrless}} 0, \tag{5.7}$$

which can be done because the logarithm is a monotonically increasing function.

- Maximum *a posteriori* detector:

$$p_{H|\mathbf{X}}(1|\mathbf{x}) \underset{D=0}{\overset{D=1}{\gtrless}} p_{H|\mathbf{X}}(0|\mathbf{x}),$$

which can also be expressed as a LRT as

$$\frac{p_{\mathbf{X}|H}(\mathbf{x}|1)}{p_{\mathbf{X}|H}(\mathbf{x}|0)} \underset{D=0}{\overset{D=1}{\gtrless}} \frac{P_H(0)}{P_H(1)}. \tag{5.8}$$

As in the general case with M hypothesis, the MAP detector minimizes the probability of error and the ML and MAP detectors are the same if $P_H(0) = P_H(1) = 0.5$. Moreover, we can see that both detectors can be expressed as a LRT

$$\frac{p_{\mathbf{X}|H}(\mathbf{x}|1)}{p_{\mathbf{X}|H}(\mathbf{x}|0)} \underset{D=0}{\overset{D=1}{\gtrless}} \eta, \tag{5.9}$$

where η is a threshold. When this threshold is 1, the LRT is the ML detector and for $\eta = P_H(0)/P_H(1)$, the LRT becomes the MAP detector, that is, minimum P_e detector. Hence, we get two different points in the ROC curve. Actually, sweeping the value of the threshold generates the complete ROC curves in Figure 5.3.[1]

5.3.2 Bayesian decision-making: the minimum risk detector

As we have already studied, sometimes it makes more sense to measure the performance of a detector in terms of the expected cost. Therefore, it is important to tackle the problem of designing a detector that is optimum with respect to the expected cost.

Remember that the expected cost of a detector deciding d for an observation $\mathbf{x}$ is given by Equation (5.5), which we reproduce here for convenience:

$$\mathbb{E}\{c_{dH}|\mathbf{x}\} = \sum_{h=0}^{M-1} c_{dh} P_{H|\mathbf{X}}(h|\mathbf{x}). \tag{5.10}$$

Minimizing the expected cost over the whole observation space requires that decisions for each observation minimize the conditional expected cost. That is, for each $\mathbf{x}$ the above expression should be minimized, and the expression of the minimum mean cost detector can be stated as follows:

$$d^\star = \arg\min_d \sum_{h=0}^{M-1} c_{dh} P_{H|\mathbf{X}}(h|\mathbf{x}).$$

Hence, when designing the detector, we need to evaluate the cost of the different decisions for each observation vector, and select the decision for which the expected cost is minimized.

Example 5.5 Continuing with Example 5.3, the conditional risk of each decision can be calculated as:

$$\mathbb{E}\{c(d,H)|x\} = c_{d0}P_{H|X}(0|x) + c_{d1}P_{H|X}(1|x) + c_{d2}P_{H|X}(2|x)$$

where the posterior distributions can be obtained by applying Bayes' Theorem:

$$P_{H|X}(0|x) = \frac{p_{X|H}(x|0)P_H(0)}{\sum_{h=0}^{2} p_{X|H}(x|h)P_H(h)} = \frac{1\cdot 0.4}{1\cdot 0.4 + 2(1-x)\cdot 0.3 + 2x\cdot 0.3} = 0.4$$

$$P_{H|X}(1|x) = \frac{p_{X|H}(x|1)P_H(1)}{\sum_{h=0}^{2} p_{X|H}(x|h)P_H(h)} = \frac{2(1-x)\cdot 0.3}{1} = 0.6(1-x)$$

$$P_{H|X}(2|x) = \frac{p_{X|H}(x|2)P_H(2)}{\sum_{h=0}^{2} p_{X|H}(x|h)P_H(h)} = \frac{2x\cdot 0.3}{1} = 0.6x$$

This leads to:

[1] This actually applies to all detectors that can be written as $\phi(\mathbf{x}) \underset{D=0}{\overset{D=1}{\gtrless}} \eta$. That is, comparing a function of the observations with a threshold achieves a given $(P_{\mathrm{FA}}, P_{\mathrm{D}})$ point in the ROC curve. These detectors are known as threshold detectors.

- if $d=0$:
$$\begin{aligned}\mathbb{E}\{c(0,H)|x\} &= c_{00}P_{H|X}(0|x)+c_{01}P_{H|X}(1|x)+c_{02}P_{H|X}(2|x)\\ &= 0\cdot 0.4+1\cdot 0.6(1-x)+1\cdot 0.6x = 0.6\end{aligned}$$
- if $d=1$:
$$\begin{aligned}\mathbb{E}\{c(1,H)|x\} &= c_{10}P_{H|X}(0|x)+c_{11}P_{H|X}(1|x)+c_{12}P_{H|X}(2|x)\\ &= 1\cdot 0.4+0\cdot 0.6(1-x)+1\cdot 0.6x = 0.4+0.6x\end{aligned}$$
- if $d=2$:
$$\begin{aligned}\mathbb{E}\{c(2,H)|x\} &= c_{20}P_{H|X}(0|x)+c_{21}P_{H|X}(1|x)+c_{22}P_{H|X}(2|x)\\ &= 1\cdot 0.4+1\cdot 0.6(1-x)+0\cdot 0.6x = 1-0.6x\end{aligned}$$

It is interesting to point out that when the cost function penalizes equally all kinds of errors, i.e.,

$$c_{dh} = \begin{cases} 0, & d=h \\ c, & d\neq h \end{cases}$$

the detector with minimum expected cost becomes the MAP one. This is easily proved by replacing these costs into the expression for the minimum expected cost detector

$$\begin{aligned} d^\star &= \arg\min_d \sum_{h=0}^{M-1} c_{dh}P_{H|\mathbf{X}}(h|\mathbf{x}) \\ &= \arg\min_d \; c\sum_{h\neq d} P_{H|\mathbf{X}}(h|\mathbf{x}) \\ &= \arg\min_d 1-P_{H|\mathbf{X}}(d|\mathbf{x}) \\ &= \arg\max_d P_{H|\mathbf{X}}(d|\mathbf{x}) \\ &= d_{MAP}. \end{aligned} \tag{5.11}$$

5.3.2.1 Binary case: Minimum risk detector

In the binary case, we can also express the optimum detector with respect to a cost function as a LRT. Let us start by particularizing (5.10) for $d=0$ and $d=1$, and then follow the criterion of deciding in favor of the minimum cost, i.e.,

$$\mathbb{E}\{c_{0H}|\mathbf{x}\} \underset{D=0}{\overset{D=1}{\gtrless}} \mathbb{E}\{c_{1H}|\mathbf{x}\}.$$

Now, using the definition of expectation, the criterion becomes

$$c_{00}P_{H|\mathbf{X}}(0|\mathbf{x})+c_{01}P_{H|\mathbf{X}}(1|\mathbf{x}) \underset{D=0}{\overset{D=1}{\gtrless}} c_{10}P_{H|\mathbf{X}}(0|\mathbf{x})+c_{11}P_{H|\mathbf{X}}(1|\mathbf{x}),$$

which after some algebra can be rewritten as

$$\frac{P_{H|\mathbf{X}}(1|\mathbf{x})}{P_{H|\mathbf{X}}(0|\mathbf{x})} \underset{D=0}{\overset{D=1}{\gtrless}} \frac{c_{10}-c_{00}}{c_{01}-c_{11}}.$$

Finally, using Bayes' Theorem, we may rewrite the *a posteriori* probabilities in terms of the likelihoods and the *a priori* probabilities, which finally yields

$$\frac{P_{\mathbf{X}|H}(\mathbf{x}|1)}{P_{\mathbf{X}|H}(\mathbf{x}|0)} \underset{D=0}{\overset{D=1}{\gtrless}} \frac{c_{10}-c_{00}}{c_{01}-c_{11}} \frac{P_H(0)}{P_H(1)},$$

and corresponds to yet another point of the ROC curve of the LRT.

5.3.3 Non-Bayesian detectors

Non-Bayesian detectors are those that do not ground on a probability model for the hypothesis. Their design depends on the likelihood functions only. This is the case, for instance, of the ML detector. Other non-Bayesian detectors, in the binary case can be expressed as the LRT in (5.9) for different values of η. The Neyman-Pearson detector is a classical example.

5.3.3.1 Binary case: Neyman-Pearson detector

The Neyman-Pearson (NP) detector is a well known detector for binary problems, which maximizes the probability of detection while it provides a bound on the probability of false alarm. Before proceeding with the derivation, let us recall the definitions of probability of false alarm and detection

$$P_{\text{FA}} = \int_{\mathscr{X}_1} p_{\mathbf{X}|H}(\mathbf{x}|0)d\mathbf{x},$$
$$P_{\text{D}} = \int_{\mathscr{X}_1} p_{\mathbf{X}|H}(\mathbf{x}|1)d\mathbf{x}.$$

Now, the NP detector can be derived as the solution to

$$\text{maximize } P_{\text{D}}, \quad \text{subject to } P_{\text{FA}} \leq \alpha,$$

which is an optimization problem with constraints. The solution to this kind of problems is obtained from the Lagrangian, which is given by

$$\begin{aligned}\mathscr{L}(\mathscr{X}_1,\eta) &= P_{\text{D}} - \eta(P_{\text{FA}}-\alpha) \\ &= \int_{\mathscr{X}_1} p_{\mathbf{X}|H}(\mathbf{x}|1)d\mathbf{x} - \eta\left(\int_{\mathscr{X}_1} p_{\mathbf{X}|H}(\mathbf{x}|0)d\mathbf{x}-\alpha\right) \\ &= \int_{\mathscr{X}_1} \left(p_{\mathbf{X}|H}(\mathbf{x}|1) - \eta\, p_{\mathbf{X}|H}(\mathbf{x}|0)\right) d\mathbf{x} + \eta\,\alpha.\end{aligned}$$

Note, that the optimization variable is the region where we decide $d=1$. Next, we need to maximize the Lagrangian, and therefore the P_{D}, which is achieved by maximizing the above integral. To do so, and taken into account that an integral may be seen as a sum, we need to design $\mathscr{X}_1$ such that the integrand is positive, i.e.

$$\mathscr{X}_1 = \{\mathbf{x}|p_{\mathbf{X}|H}(\mathbf{x}|1) - \eta\, p_{\mathbf{X}|H}(\mathbf{x}|0) \geq 0\} \Rightarrow \frac{p_{\mathbf{X}|H}(\mathbf{x}|1)}{p_{\mathbf{X}|H}(\mathbf{x}|0)} \underset{D=0}{\overset{D=1}{\gtrless}} \eta,$$

and η is selected to achieve the desired probability of false alarm.

5.3.3.2 Minimax classifiers

Minimax classifiers are designed in such a way that their error probability is independent on the prior probabilities of the hypothesis. For binary decision problems, they are given by the LRT such that P_{FA} and P_{M} are the same

$$P_{\text{FA}} = P_{\text{M}} \tag{5.12}$$

5.4 Gaussian models

In this section, we will derive the likelihood ratio test for Gaussian observations under several assumptions. Then, depending on the threshold, we would obtain the different detectors: NP, ML, MAP, and minimum cost.

Before proceeding, we introduce the multivariate real Gaussian probability density function (PDF), which is given by

$$P_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{N/2}|\mathbf{V}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{m})^T\mathbf{V}^{-1}(\mathbf{x}-\mathbf{m})\right),$$

where $\mathbf{x}$ is an N-dimensional vector, $\mathbf{m}$ is the mean vector, and $\mathbf{V}$ is the cross-covariance matrix. Then, under hypothesis $h = 0$, the likelihood is

$$P_{\mathbf{X}|H}(\mathbf{x}|0) = \frac{1}{(2\pi)^{N/2}|\mathbf{V}_0|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{m}_0)^T\mathbf{V}_0^{-1}(\mathbf{x}-\mathbf{m}_0)\right),$$

whereas it is

$$P_{\mathbf{X}|H}(\mathbf{x}|1) = \frac{1}{(2\pi)^{N/2}|\mathbf{V}_1|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{m}_1)^T\mathbf{V}_1^{-1}(\mathbf{x}-\mathbf{m}_1)\right),$$

under hypothesis $h = 1$. For this hypothesis test, the LLRT in (5.7) becomes

$$-\frac{1}{2}\log|\mathbf{V}_1| - \frac{1}{2}(\mathbf{x}-\mathbf{m}_1)^T\mathbf{V}_1^{-1}(\mathbf{x}-\mathbf{m}_1) + \frac{1}{2}\log|\mathbf{V}_0| + \frac{1}{2}(\mathbf{x}-\mathbf{m}_0)^T\mathbf{V}_0^{-1}(\mathbf{x}-\mathbf{m}_0) \underset{D=0}{\overset{D=1}{\gtrless}} \log(\eta)$$

or, equivalently,

$$(\mathbf{x}-\mathbf{m}_0)^T\mathbf{V}_0^{-1}(\mathbf{x}-\mathbf{m}_0) - (\mathbf{x}-\mathbf{m}_1)^T\mathbf{V}_1^{-1}(\mathbf{x}-\mathbf{m}_1) \underset{D=0}{\overset{D=1}{\gtrless}} \mu \tag{5.13}$$

where

$$\mu = 2\log(\eta) + \log|\mathbf{V}_1| - \log|\mathbf{V}_0|,$$

with η being a threshold selected according to the performance criterion.

After a careful look at (5.13), it can be shown that the optimal detector in the Gaussian case is given by a second-order polynomial function. Hence, the decision boundaries[2] are quadratic surfaces. For instance, for 2D problems ($N = 2$), these boundaries are hyperbolas, parabolas, ellipses or straight lines.

In the following sections, we consider a few particular cases, and we conclude this section with two examples.

Example 5.6 Figure 5.4 shows the decision boundaries for the ML detector ($\eta = 1$ in (5.13)), for a detection problem with 2D Gaussian observations with the following means and cross-covariance matrices:

$$\mathbf{m}_0 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \mathbf{V}_0 = \begin{pmatrix} 1.2 & 0.43 \\ 0.43 & 1.75 \end{pmatrix},$$

and

$$\mathbf{m}_1 = \begin{pmatrix} 3 \\ 3 \end{pmatrix}, \mathbf{V}_1 = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}.$$

In this figure, the gray color gradient represents the value of the likelihoods $P_{\mathbf{X}|H}(\mathbf{x}|0)$ and $P_{\mathbf{X}|H}(\mathbf{x}|1)$, where darker colors denote larger values. Moreover, the white curves are the iso-probability lines and the black curve is the decision boundary, which in this case is a hyperbola (the symmetric part is not shown in this figure).

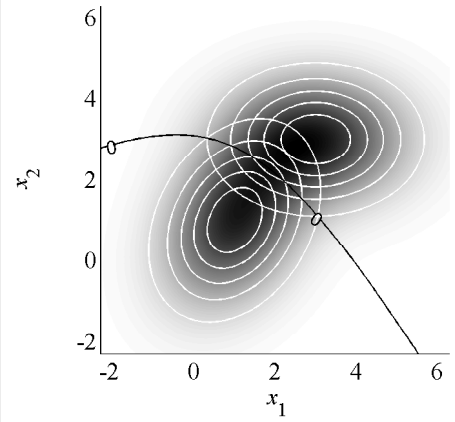


Fig. 5.4 Hyperbolic decision boundary of the ML detector and likelihoods for a Gaussian detection problem with 2D observations.

Example 5.7 Figure 5.5 shows an equivalent figure to that of the previous example, but for a problem with the following means and cross-covariance matrices:

$$\mathbf{m}_0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \mathbf{V}_0 = \begin{pmatrix} 0.7 & 0 \\ 0 & 0.7 \end{pmatrix},$$

[2] We obtain the decision boundaries for the equality in (5.13).

and

$$\mathbf{m}_1 = \begin{pmatrix} 0.2 \\ 0.4 \end{pmatrix}, \mathbf{V}_1 = \begin{pmatrix} 0.5 & 0 \\ 0 & 0.2 \end{pmatrix}.$$

In this case, the decision boundary is an ellipse.

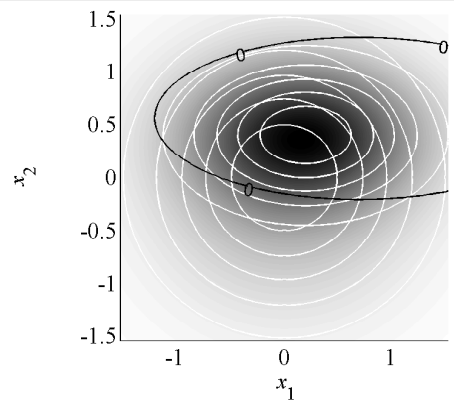


Fig. 5.5 Elliptic decision boundary of the ML detectors and likelihoods for a Gaussian detection problem with 2D observations.

5.4.1 Identical cross-covariance matrices

This section considers the case of $\mathbf{V}_1 = \mathbf{V}_0 = \mathbf{V}$. Then, the LLRT becomes

$$(\mathbf{x}-\mathbf{m}_0)^T\mathbf{V}^{-1}(\mathbf{x}-\mathbf{m}_0) - (\mathbf{x}-\mathbf{m}_1)^T\mathbf{V}^{-1}(\mathbf{x}-\mathbf{m}_1) \underset{D=0}{\overset{D=1}{\gtrless}} \mu.$$

Now, expanding the quadratic forms, the above expression simplifies to

$$(\mathbf{m}_1-\mathbf{m}_0)^T\mathbf{V}^{-1}\mathbf{x} \underset{D=0}{\overset{D=1}{\gtrless}} \tilde{\mu}, \tag{5.14}$$

where $\tilde{\mu} = \mu/2 + \mathbf{m}_1^T\mathbf{V}^{-1}\mathbf{m}_1/2 - \mathbf{m}_0^T\mathbf{V}^{-1}\mathbf{m}_0/2$. In this particular case, the LLRT in (5.14) is a linear function of the observation vector $\mathbf{x}$.

Example 5.8 Figure 5.6 shows three decision boundaries for an example with

$$\mathbf{m}_0 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \mathbf{V}_0 = \begin{pmatrix} 0.44 & 0.32 \\ 0.32 & 0.81 \end{pmatrix}$$

and

$$\mathbf{m}_1 = \begin{pmatrix} 3 \\ 3 \end{pmatrix}, \mathbf{V}_1 = \begin{pmatrix} 0.44 & 0.32 \\ 0.32 & 0.81 \end{pmatrix}.$$

The label of each decision boundary is $\log(\eta)$. Then, $\log(\eta) = 0$ corresponds to the ML detector.

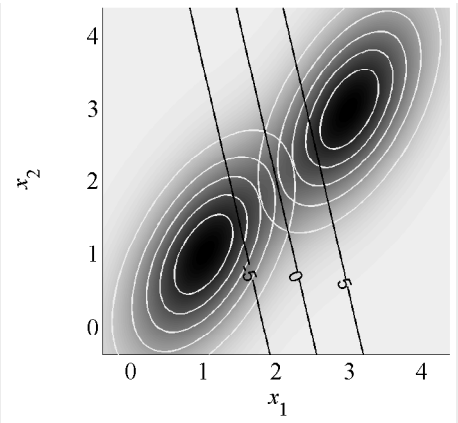


Fig. 5.6 Decision boundaries of the LLRT and likelihoods for a Gaussian detection problem with 2D observations and identical covariance matrices.

Example 5.9 (Matched filter) In this example, we derive one of the most well-known detectors, the matched filter (MF). The MF is the LLRT to the detection of a known signal contaminated by zero-mean Gaussian noise. Concretely, under hypothesis $h = 0$, the observations are given by noise only:

$$x[n] = w[n], \quad n = 0, \ldots, N-1,$$

and under hypothesis $h = 1$, the observations are

$$x[n] = s[n] + w[n], \quad n = 0, \ldots, N-1,$$

where $s[n]$ is a known signal and $w[n]$ is additive white Gaussian noise with zero mean and variance σ^2, i.e., $w[n] \sim \mathcal{N}(0, \sigma^2)$. To use the LLRT already derived in this section, we must first define the vector

$$\mathbf{x} = \big(x[0]\ x[1]\ \cdots\ x[N-1]\big)^T = \mathbf{s} + \mathbf{w},$$

with $\mathbf{s} = \big(s[0]\ s[1]\ \cdots\ s[N-1]\big)^T$ and $\mathbf{w} = \big(w[0]\ w[1]\ \cdots\ w[N-1]\big)^T$, and obtain the distributions of $\mathbf{x}$ under both hypothesis. Under hypothesis $h = 0$, the observation vector $\mathbf{x}$ collects samples of a Gaussian process, which makes it also Gaussian. Hence, only the mean and cross-covariance matrices are required:

$$\mathbf{m}_0 = \mathbb{E}\{\mathbf{x}|0\} = \mathbb{E}\{\mathbf{w}\} = \big(\mathbb{E}\{w[0]\}\ \mathbb{E}\{w[1]\}\ \cdots\ \mathbb{E}\{w[N-1]\}\big)^T = \mathbf{0},$$

and

$$\begin{aligned}
\mathbf{V}_0 &= \mathbb{E}\left\{(\mathbf{x}-\mathbf{m}_0)(\mathbf{x}-\mathbf{m}_0)^T|0\right\} = \mathbb{E}\left\{\mathbf{w}\mathbf{w}^T\right\} \\
&= \mathbb{E}\left\{\big(w[0]\ w[1]\ \cdots\ w[N-1]\big)^T \big(w[0]\ w[1]\ \cdots\ w[N-1]\big)\right\} \\
&= \begin{pmatrix}
\mathbb{E}\left\{w^2[0]\right\} & \mathbb{E}\{w[0]w[1]\} & \cdots & \mathbb{E}\{w[0]w[N-1]\} \\
\mathbb{E}\{w[1]w[0]\} & \mathbb{E}\left\{w^2[1]\right\} & \cdots & \mathbb{E}\{w[1]w[N-1]\} \\
\vdots & \vdots & \ddots & \vdots \\
\mathbb{E}\{w[N-1]w[0]\} & \mathbb{E}\{w[N-1]w[1]\} & \cdots & \mathbb{E}\left\{w^2[N-1]\right\}
\end{pmatrix}.
\end{aligned}$$

The cross-covariance matrix $\mathbf{V}_0$ can be simplified taking into account that the noise is white, i.e., $\mathbb{E}\{w[n]w[n-m]\} = \sigma^2\delta[m]$, which yields

$$\mathbf{V}_0 = \begin{pmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{pmatrix} = \sigma^2 \mathbf{I}.$$

Similarly, under hypothesis $h = 1$, the observations are Gaussian with mean

$$\mathbf{m}_1 = \mathbb{E}\{\mathbf{x}|1\} = \mathbb{E}\{\mathbf{s}+\mathbf{w}\} = \mathbb{E}\{\mathbf{s}\} + \mathbb{E}\{\mathbf{w}\} = \mathbf{s},$$

since $\mathbf{s}$ is deterministic, and cross-covariance matrix

$$\mathbf{V}_1 = \mathbb{E}\left\{(\mathbf{x}-\mathbf{m}_1)(\mathbf{x}-\mathbf{m}_1)^T|1\right\} = \mathbb{E}\left\{(\mathbf{s}+\mathbf{w}-\mathbf{s})(\mathbf{s}+\mathbf{w}-\mathbf{s})^T\right\} = \mathbb{E}\left\{\mathbf{w}\mathbf{w}^T\right\} = \sigma^2\mathbf{I}.$$

Hence, the detection problem is that of Gaussian observations with identical covariance matrices, for which the LLRT is

$$(\mathbf{m}_1 - \mathbf{m}_0)^T \mathbf{V}^{-1}\mathbf{x} = \frac{1}{\sigma^2}\mathbf{s}^T\mathbf{x} \underset{D=0}{\overset{D=1}{\gtrless}} \tilde{\mu} \Rightarrow \underbrace{\sum_{n=0}^{N-1} s[n]x[n]}_{MF} \underset{D=0}{\overset{D=1}{\gtrless}} \sigma^2\tilde{\mu}.$$

Alternatively, and the motivation for the term matched filter, is because the above detector can be rewritten as a filtering of the signal $x[n]$ with the filter $h[n] = s[N-1-n]$, followed by sampling every N samples. Finally, we also would like to point out that the matched filter is a filter that maximizes the signal-to-noise ratio.

5.4.2 Zero means

We consider now that $\mathbf{m}_0 = \mathbf{m}_1 = \mathbf{0}$, which yields

$$\mathbf{x}^T\left(\mathbf{V}_0^{-1} - \mathbf{V}_1^{-1}\right)\mathbf{x} \underset{D=0}{\overset{D=1}{\gtrless}} \mu.$$

Example 5.10 Figure 5.7 shows the ML decision boundary for 2D Gaussian observations with

$$\mathbf{m}_0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \mathbf{V}_0 = \begin{pmatrix} 0.62 & -0.22 \\ -0.22 & 0.37 \end{pmatrix},$$

and

$$\mathbf{m}_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \mathbf{V}_1 = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}.$$

The region $\mathscr{X}_0$ is given by the interior of the ellipse. Moreover, since the variance of the observations in every direction is larger under hypothesis $h = 1$, points further away from the origin should be assigned $d = 1$.

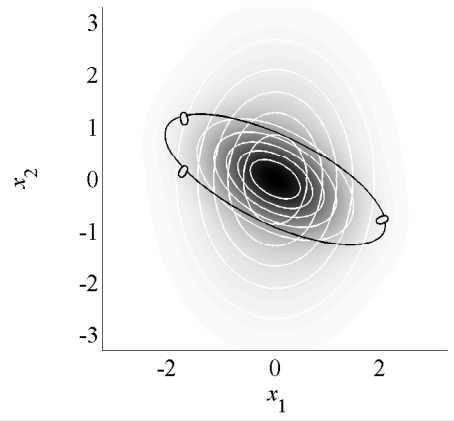


Fig. 5.7 Elliptic decision boundary for a 2D Gaussian problem with zero means.

Example 5.11 Figure 5.8 shows the ML decision boundary for 2D Gaussian observations with

$$\mathbf{m}_0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \mathbf{V}_0 = \begin{pmatrix} 0.33 & 0.39 \\ 0.39 & 0.77 \end{pmatrix}$$

and

$$\mathbf{m}_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \mathbf{V}_1 = \begin{pmatrix} 0.39 & -0.19 \\ -0.19 & 0.16 \end{pmatrix}.$$

In this example, the variance under hypothesis $h = 1$ is larger only along dimension 1, whereas it is smaller along dimension 2. Hence, as a consequence, the boundary is a hyperbola.

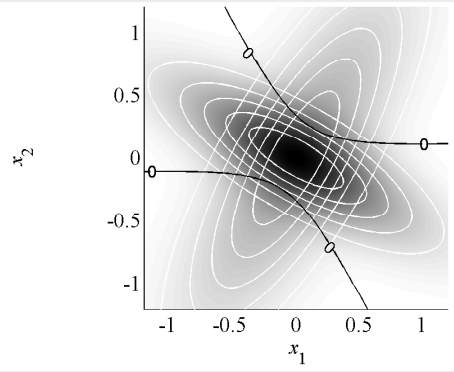


Fig. 5.8 Hyperbolic decision boundary for a 2D Gaussian problem with zero means.

5.5 Problems

5.1 Consider the decision problem with three hypotheses given by the observation $\mathbf{x} = (x_1, x_2) \in [0,1]^2$ and likelihoods:

$$p_{\mathbf{X}|H}(\mathbf{x}|0) = 2(1 - x_1) \tag{5.15}$$

$$p_{\mathbf{X}|H}(\mathbf{x}|1) = 2x_1 \tag{5.16}$$

$$p_{\mathbf{X}|H}(\mathbf{x}|2) = 2x_2 \tag{5.17}$$

a) Determine the ML (Maximum Likelihood) classifier
b) Represent the decision regions.

5.2 Consider the decision problem given by the observation $x \in [0,1]$, likelihoods:

$$p_{X|H}(x|0) = 2(1-x) \tag{5.18}$$

$$p_{X|H}(x|1) = 1 \tag{5.19}$$

and prior probability $P_H(1) = 1/4$.

a) Determine the ML classifier.
b) Determine the MAP (Maximum A Posteriori) classifier.
c) Given that $c_{01} = 2$, $c_{10} = 1$, $c_{11} = c_{00} = 0$, determine the decision-maker of minimum risk.
d) Consider a threshold detector over x in the form:

$$x \underset{D=0}{\overset{D=1}{\gtrless}} \eta \tag{5.20}$$

Calculate the probabilities of false alarm, miss and error, as a function of η.
e) Apply the result to the previous three decision-makers. Verify that the MAP classifier obtains the minimum probability of error.
f) Determine the risk for the previous three decision-makers, using the cost parameters from part (c), and verify that the decision-maker obtained in said part achieves the lowest risk.

5.3 Consider a one-dimensional binary decision problem with equiprobable hypotheses, defined by the likelihoods:

$$p_{X|H}(x|0) = \frac{1}{6}, \quad |x| \leq 3,$$

$$p_{X|H}(x|1) = \frac{3}{2}x^2, \quad |x| \leq 1$$

a) Determine the ML (Maximum Likelihood) classifier.
b) Determine the values of P_{FA} (false alarm probability), P_{M} (miss probability), and P_{e} (error probability) for the above classifier.

5.4 Consider the decision problem defined by the observation $x \geq 0$ and likelihoods:

$$p_{X|H}(x|0) = \exp(-x); \tag{5.21}$$

$$p_{X|H}(x|1) = 2\exp(-2x); \tag{5.22}$$

a) Determine the LRT (Likelihood Ratio Test) decision rule.
b) Determine the ROC (Receiver Operating Characteristic).

5.5 Consider a binary decision problem where the likelihood for hypothesis $H = 0$ is uniform over the interval $0 < x < 1$, while $p_{X|H}(x|1) = 2x$, $0 < x < 1$.

a) Obtain the general expression for a Likelihood Ratio Test with parameter η. Graphically represent both likelihoods on the same axes, indicating the decision regions for the ML case.
b) Obtain the analytic expression of the ROC curve for the LRT. Plot this curve indicating the operating points for the ML classifier and the Neyman-Pearson detector with parameter $\alpha = 0.1$ (i.e., $P_{\text{FA}} \leq \alpha = 0.1$).

5.6 Company E manufactures 10,000 units of a product daily. It has been estimated that:

- The sale of a unit in good condition nets a profit of 3 Euros.
- Placing a defective unit on the market causes (on average) a loss of 81 Euros.
- The withdrawal of a unit (whether defective or not) results in a loss of 1 Euro.

An automatic inspection system is available that obtains, for each unit, an observation x_1. Defining $H = 0$ as the hypothesis "the unit is not defective" and $H = 1$ as "the unit is defective",

$$p_{X_1|H}(x_1|0) = \exp(-x_1)u(x_1) \tag{5.23}$$

$$p_{X_1|H}(x_1|1) = \lambda_1 \exp(-\lambda_1 x_1)u(x_1) \tag{5.24}$$

where $\lambda_1 = 1/2$. Additionally, the assembly line produces, on average, one defective unit for every 100 non-defective ones.

The aim is to incorporate an automatic mechanism for withdrawing defective units based on the observation of x_1.

a) Design the detector that provides Company E with the highest expected profit.
b) Determine the maximum expected daily profit that can be achieved.
c) A company offers Company E an innovative inspection device that provides, for each product, in addition to x_1, a new observation x_2, statistically independent from x_1, such that

$$p_{X_2|H}(x_2|0) = \exp(-x_2)u(x_2) \tag{5.25}$$

$$p_{X_2|H}(x_2|1) = \lambda_2 \exp(-\lambda_2 x_2)u(x_2) \tag{5.26}$$

where $\lambda_2 = 1/4$. The cost of this machine is 6000 Euros. Determine an expression for the average time it would take Company E to amortize this machine.

Appendix A
Transformations of random variables

A.1 Change of Random Variable

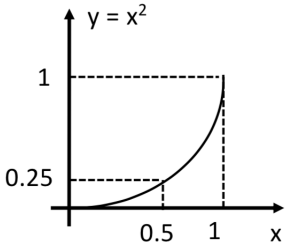
Let's consider we know the probability of a r.v. X, $p_X(x)$, and we now want to compute the probability density function of some variable $Y = f(X)$, that is, we need to calculate $p_Y(y)$.

To understand how this new distribution or **change of random variable** is calculated, let's firstly solve a particular case:

- X is a uniform distribution in the interval $(0,1)$.

$$p_X(x) = \begin{cases} 1 & \text{if} \quad 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

- $Y = X^2$. Note that this change produces this transformation:



x	$y = x^2$
0.1	0.01
0.2	0.04
0.5	0.25
...	...

The transformation function $f(\cdot)$ is strictly increasing. So there exists its inverse function $f^{-1}(\cdot)$.

To solve this change of r.v., we are going to use the fact that:

$$P\{0 < X < 0.1\} = P\{0 < Y < 0.01\}$$
$$P\{0 < X < 0.2\} = P\{0 < Y < 0.04\}$$
$$P\{0 < X < 0.5\} = P\{0 < Y < 0.25\}$$

or, in a general case, for any value of X, x_0, we have

$$P\{0 < X < x_0\} = P\{0 < Y < y_0\}$$

where $y_0 = x_0^2$ or $x_0 = \sqrt{y_0}$

So, we can compute the cumulative distribution function of the r.v. Y as

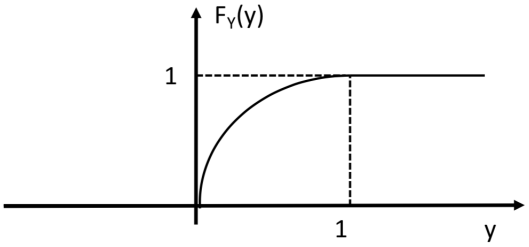
$$F_Y(y_0) = P\{Y < y_0\} = P\{X < \sqrt{y_0}\}$$

Now, as the cumulative function of Y is expressed in terms of the r.v X, we can compute it!!!

$$F_Y(y_0) = P\{X < \sqrt{y_0}\} = \int_{-\infty}^{\sqrt{y_0}} p_X(x)dx = \begin{cases} \int_{-\infty}^{\sqrt{y_0}} 0dx = 0 & \text{if} \quad y_0 < 0 \\ \int_0^{\sqrt{y_0}} 1dx = \sqrt{y_0} & \text{if} \quad 0 < y_0 < 1 \\ \int_0^1 1dx = 1 & \text{if} \quad y_0 > 1 \end{cases}$$

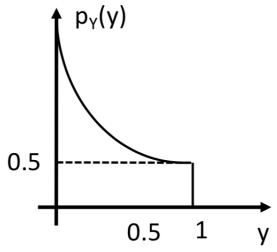
So, we have that

$$F_Y(y_0) = \begin{cases} 0 & \text{if} \quad y_0 < 0 \\ \sqrt{y_0} & \text{if} \quad 0 < y_0 < 1 \\ 1 & \text{if} \quad y_0 > 1 \end{cases}$$



and, finally, we can obtain the density function of Y as

$$p_Y(y) = \frac{dF_Y(y)}{dy} = \begin{cases} \frac{1}{2\sqrt{y}} & \text{if} \quad 0 < y < 1 \\ 0 & \text{otherwise} \end{cases}$$



Now, let's try to generalize this procedure for any transformation

$$Y = f(X)$$

being $f(\cdot)$ a strictly increasing function, so $f^{-1}(\cdot)$ exists.

1. Compute the cumulative function of Y (by means of X)

$$F_Y(y) = P\{Y < y\} = P\{X < f^{-1}(y)\} = \int_{-\infty}^{f^{-1}(y)} p_X(x)dx =$$
$$F_X(f^{-1}(y)) - F_X(-\infty) = F_X(f^{-1}(y))$$

   Note: $F_X(-\infty) = 0$ for any cumulative distribution function
2. Compute the density distribution function (use the chain rule)

$$p_Y(y) = \frac{dF_Y(y)}{dy} = \frac{dF_X(f^{-1}(y))}{dy} = \frac{dF_X(x = f^{-1}(y))}{dx}\frac{dx}{dy} = p_X(x = f^{-1}(y))\frac{dx}{dy}$$

So, we obtain that

$$p_Y(y) = p_X(x = f^{-1}(y))\frac{dx}{dy}$$

This formula for the r.v. change can be generalized for any transformation function $f(\cdot)$ which is monotic (either strictly increasing or decreasing) as follows:

$$p_Y(y) = p_X(x = f^{-1}(y))\left|\frac{dx}{dy}\right| \tag{A.1}$$

In fact, we can now use this formula over the previous example:

$$Y = X^2 \qquad\qquad p_X(x) = \begin{cases} 1 & \text{if} \quad 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

each term of the formula (A.1) is given by:

$$\left|\frac{dx}{dy}\right| = \left|\frac{df^{-1}(y)}{dy}\right| = \left|\frac{d\sqrt{y}}{dy}\right| = \frac{1}{2\sqrt{y}}$$

$$p_X(x = f^{-1}(y)) = p_X(x = \sqrt{y}) = \begin{cases} 1 & \text{if} \quad 0 < \sqrt{y} < 1 \\ 0 & \text{otherwise} \end{cases}$$

So, we get

$$p_Y(y) = \frac{1}{2\sqrt{y}} p_X(x = \sqrt{y}) = \begin{cases} \frac{1}{2\sqrt{y}} & \text{if} \quad 0 < y < 1 \\ 0 & \text{otherwise} \end{cases}$$

In case the transformation function is not monotic, we have to divide the transformation into intervals where we get monotic transformations. That is, we have $Y = f(X)$ and $f(\cdot)$ is not monotic, then redefine the transformation as

$$Y = \begin{cases} f_1(X) & \text{if} x_0 < x < x_1 \\ f_2(X) & \text{if} x_1 < x < x_2 \\ \dots & \\ f_N(X) & \text{if} x_{N-1} < x < x_N \end{cases}$$

where $f_1(\cdot), \dots, f_N(\cdot)$ are monotic. Then, you can compute $p_Y(y)$ as:

$$p_Y(y) = \sum_{n=1}^{N} p_X(x = f_n^{-1}(y)) \left|\frac{df_n^{-1}(y)}{dy}\right|$$

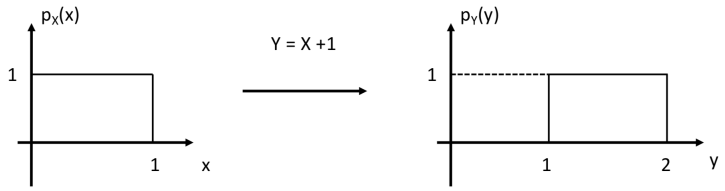
A.1.1 Some usual r.v. changes

The demonstration of these changes is left as homework.

1. SHIFTING of R.V.
   $Y = X + a$, where a is a known constant. Then,

   $$p_Y(y) = p_X(x = y - a)$$

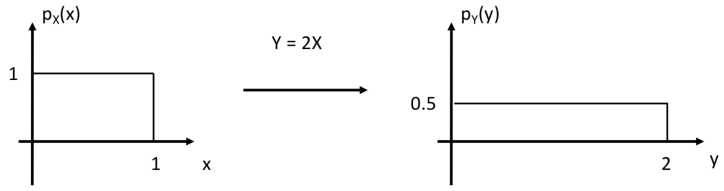
   when we are adding a constant to any r.v., we are shifting the distribution from the origin to the position of the constant



2. RESCALING of R.V.
   $Y = aX$, where a is a known constant. Then,

$$p_Y(y) = \frac{1}{a} p_X(x = \frac{y}{a})$$

   in this case we are modifying both the support of the distribution function and its height.



Appendix B
Introductory examples

B.1 Some introductory examples

The contents of this section provide an introduction to the detection problem in the binary case using some simple examples. Concretely, we will present some basic concepts through these examples. Important concepts, such as hypothesis, their *a priori* and *a posteriori* probabilities, likelihoods, or cost and cost function, will be introduced.

Before proceeding, we would like to point out that detection theory is the term employed by some communities, while some use hypothesis testing and others classification.

B.1.1 Example 1: Binary detection with no observations

Problem B.1 Consider a game in which two dice are rolled and our task consists in deciding whether the sum of both dice is larger than or equal to 10, or smaller thereof. For this problem, you have to answer the following questions:

a) What decision results in fewer errors in the long term?
b) Consider now that not all errors are penalized the same. In particular, let us assume that the errors of wrongly deciding that the sum of the dice is larger than or equal to 10 ($S \geq 10$) are assigned a penalty (or cost) of c, whereas wrongly deciding $S < 10$ results in a unit cost (per wrong guess). What would be in this case the long term cost of both decision strategies?
c) What is the optimal strategy to minimize the expected cost? Provide your answer as a function of c.

Solution B.1 Let us start by introducing some notation for this problem. Note that the design of a detector must always be done according to a criterion "in the long term". In other words, the goal is to analyze the average performance as the number of experiments tends to infinity. Hence, there are certain variables that will take different values in each experiment, and these need to be modeled by random variables.

- We denote by X_1 and X_2 the random variables (r.v.) that represent the result of each die roll. Since we consider fair dice, we have $P_{X_i}(x_i) = \frac{1}{6}$, for $i = 1,2$, and for $x_i \in \{1,2,3,4,5,6\}$.
- The sum of the dice is represented with the random variable $S = X_1 + X_2$.
- Finally, this problem involves two different hypotheses depending on the value of S. Since the true hypothesis can change between experiments, we introduce a discrete random variable H that can take just two values

$$h = 0 \text{ if and only if } \{s < 10\},$$
$$h = 1 \text{ if and only if } \{s \geq 10\}.$$

  Note that, being a function of another random variable, H is also a random variable, and it should be possible to compute its distribution from the distribution of S, which in turn can be calculated from the distributions of X_1 and X_2. Moreover, in this problem, there exists a causal relation between the random variables, which implies that the hypothe-

ses depend on X_1 and X_2. This has certain impact on how we can calculate statistical information, as we will discuss later.

a) We first need to discuss what are the possible decisions that can be implemented. Building a detection system translates into designing a function that takes all available information as input, and outputs the selected hypothesis. Since we only consider deterministic functions, and in this case there are no input features, this implies that only two functions can be considered:

   – A detector (function) that selects all the time hypothesis 0 (i.e., $d = 0$).
   – A detector (function) that selects all the time hypothesis 1 (i.e., $d = 1$).

   The probability of error of these two detectors can be calculated as follows:

   – For the former, $d = 0$:

$$P_e = P(H \neq d) = P(H \neq 0) = P_H(1).$$

   – For the latter, $d = 1$:

$$P_e = P(H \neq d) = P(H \neq 1) = P_H(0).$$

   Therefore, we need to compute the distribution of the r.v. H. To do so, we begin by calculating the probability distribution of S. Figure B.1 shows all possible outcomes of X_1 and X_2 and the corresponding value of S. Since all combinations are equally likely, and there are 36 of them, we can easily compute the distribution of S by counting the number of occurrences of each value and dividing the result by 36. Similarly, we can obtain the *a priori* probability of the two hypotheses by counting the number of occurrences of each hypothesis by 36. As indicated in the figure, we can conclude that $P_H(0) = 5/6$ and $P_H(1) = 1/6$.

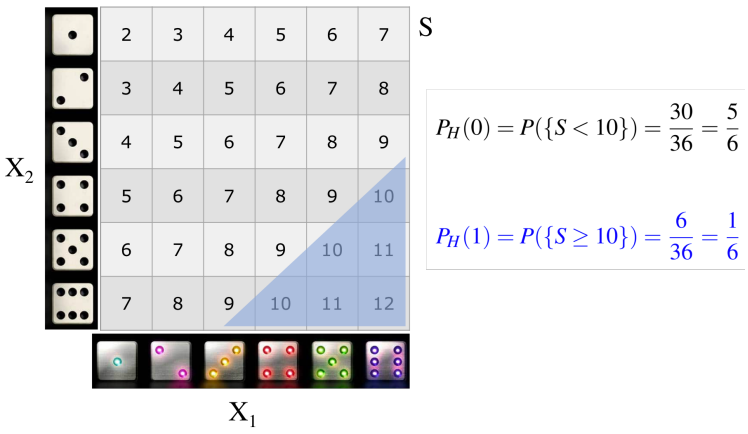


Fig. B.1 All combinations of of X_1 and X_2 are equally probable, and therefore each of the 36 results represented in the figure have a probability of $1/36$. Counting the number of occurrences of particular values of S or H, the distribution of these variables can be calculated.

Since we have to provide the criterion that minimizes the probability of error, we can then conclude that we should always decide in favor of hypothesis 0:

$$d^\star = 0,$$

with a probability of error of $1/6$.

A final remark is in order. Note that the probability of error of each criterion is given by the *a priori* probability of the complementary hypothesis. This implies that, to minimize the probability of error, we have to decide in favor of the hypothesis with a larger *a priori* probability.

b) In real applications, there are scenarios where not all the errors should be given the same importance. Here, we introduce the concept of *cost* to model the penalty that should be assigned to different kinds of errors.[1]

Since different kinds of errors can be observed in different experiments, the cost can also be modeled with a random variable C. In this particular problem, C can take four different values that we will denote as c_{dh}, for $d,h \in \{0,1\}$. That is, c_{dh} is the cost of deciding d when the true hypothesis was h. According to the wording, the costs are:

$$c_{dh} = \begin{cases} c_{00} = c_{11} = 0 \\ c_{01} = 1 \\ c_{10} = c \end{cases}$$

Since C is a function of H, it is also a random variable, for which its distribution could be obtained (from the probability distribution of H, $P_H(h)$). However, in this problem we only need to compute the expected cost of both detectors, that is,

– For the detector $d = 0$:

$$\bar{C} = \mathbb{E}\{c_{dh}\} = \mathbb{E}\{c_{0h}\} = \sum_{h=0}^{1} c_{0h} P_H(h) = c_{00} P_H(0) + c_{01} P_H(1) = \frac{1}{6}.$$

– For the detector $d = 1$:

$$\bar{C} = \mathbb{E}\{c_{dh}\} = \mathbb{E}\{c_{1h}\} = \sum_{h=0}^{1} c_{1h} P_H(h) = c_{10} P_H(0) + c_{11} P_H(1) = \frac{5c}{6}.$$

c) To minimize the expected cost, we have to compare the costs that we calculated in the previous subsection

$$\bar{C}(d=0) \underset{D=0}{\overset{D=1}{\gtrless}} \bar{C}(d=1),$$

which results in

$$c \underset{D=1}{\overset{D=0}{\gtrless}} \frac{1}{5}.$$

Let us check, using our intuition, that this result makes sense. To start with, note that when the penalty given to wrongly deciding $d = 1$ is unitary ($c_{10} = c = 1$), both kinds of

[1] In some cases rather than working with the minimization of a cost we might pursue the maximization of a profit. Both scenarios can be shown to be completely equivalent, but in this course we will always deal with cost functions.

errors are identical. In such case, it can be seen that minimizing the expected cost is the same as minimizing the probability of error, and we should decide $d = 0$ as in part a) of this problem. However, if c_{10} is sufficiently small, deciding $d = 1$ has a very small cost, so it can pay off to decide $d = 1$ even though the number of errors is larger, as it will certainly be the case since hypothesis $H = 0$ appears 5 times more often than hypothesis $H = 1$. Hence, the expression above implies that if $c < 1/5$ then detector $d = 1$ yields a smaller expected cost.

B.1.2 Example 2: Binary decision with observations

Problem B.2 Consider now the scenario described in the previous example, with the difference that, before deciding in favor of one of the hypotheses, we are allowed to see the result of the first die, X_1. In this case, we will therefore be able to take a more informed decision since knowing such value carries information about the value of S.

a) Calculate the probability of error incurred by each possible decision ($d = 0$ and $d = 1$) for each value of X_1.
b) Design the detector that minimizes the probability of error, and compute the probability of error of such detector.
c) Obtain the test statistic that minimizes the cost described in the previous example, for the particular case $c = 1/4$.

Solution B.2 The main difference of the scenario described in this problem with respect to that of the previous example is that, in this case, the detector can be a function of X_1. As a result, the decision may change from experiment to experiment, depending on the value of X_1.

Precisely, when designing a detector our goal is to assign each possible value of the observations to a particular decision. In other words, if the same input is observed twice, the output must be the same in both cases, since the mapping from the observations to the decisions is assumed to be deterministic. We will say more on this later on, but for now, we focus on providing answers to the considered problem.

a) We will follow along the same lines of the previous exercise to compute the probability of error for the two possible decisions. Notice, however, that in this case we will be conditioning these probabilities on the value of X_1.

   – For $x_1 \in \{1,2,3\}$, hypothesis $H = 1$ can never hold. Therefore, in this case it seems obvious that deciding $d = 0$ would guarantee a zero probability of error. More formally:

$$\text{If } x_1 \in \{1,2,3\} \rightarrow \begin{cases} d = 0 \rightarrow P_e = P(d \neq H | x_1 \in \{1,2,3\}) = P_{H|X_1}(1 | x_1 \in \{1,2,3\}) = 0 \\ d = 1 \rightarrow P_e = P(d \neq H | x_1 \in \{1,2,3\}) = P_{H|X_1}(0 | x_1 \in \{1,2,3\}) = 1 \end{cases}$$

   – For $x_1 = 4$, there is only one possibility out of 6 that hypothesis $H = 1$ is correct (for $x_2 = 6$). This allows us to easily compute the error of both criteria. Repeating this for the remaining values of X_1, we obtain the following probabilities of error conditioned on X_1.

$$\text{If } x_1 = 4 \rightarrow \begin{cases} d = 0 \rightarrow P_e = P(d \neq H | x_1 = 4) = P_{H|X_1}(1|x_1 = 4) = \frac{1}{6} \\ d = 1 \rightarrow P_e = P(d \neq H | x_1 = 4) = P_{H|X_1}(0|x_1 = 4) = \frac{5}{6} \end{cases}$$

$$\text{If } x_1 = 5 \rightarrow \begin{cases} d = 0 \rightarrow P_e = P(d \neq H | x_1 = 5) = P_{H|X_1}(1|x_1 = 5) = \frac{2}{6} = \frac{1}{3} \\ d = 1 \rightarrow P_e = P(d \neq H | x_1 = 5) = P_{H|X_1}(0|x_1 = 5) = \frac{4}{6} = \frac{2}{3} \end{cases}$$

$$\text{If } x_1 = 6 \rightarrow \begin{cases} d = 0 \rightarrow P_e = P(d \neq H | x_1 = 6) = P_{H|X_1}(1|x_1 = 6) = \frac{3}{6} = \frac{1}{2} \\ d = 1 \rightarrow P_e = P(d \neq H | x_1 = 6) = P_{H|X_1}(0|x_1 = 6) = \frac{3}{6} = \frac{1}{2} \end{cases}$$

In this case, the probability of error associated to each decision is given by the probability of the complementary hypothesis. The difference is that now we have to use *a posteriori* probabilities of the hypotheses, given that the decision is taken using some information (the value of X_1), and this knowledge refines how likely we can expect the different hypotheses to be. Figure B.2 depicts these probabilities. Note that to compute the probability conditioned on each value of X_1, we need to consider only the values of S that are associated to the corresponding column.

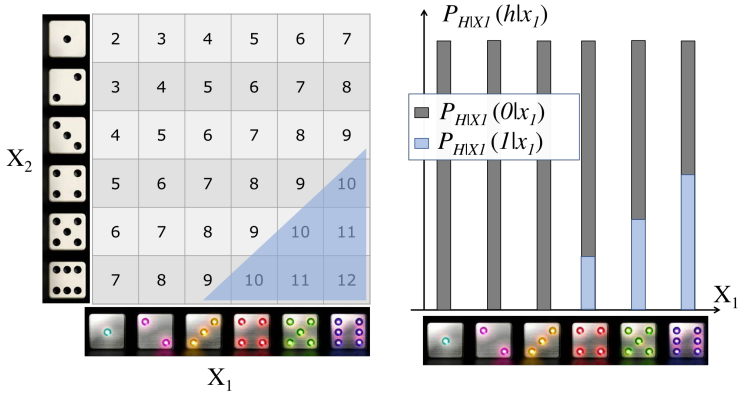


Fig. B.2 To calculate posterior probabilities of the hypothesis, we need to count how many results in each column correspond to hypothesis 0 and how many correspond to hypothesis 1. Note that $P_{H|X_1}(0|x_1) + P_{H|X_1}(1|x_1) = 1$ for all values of X_1.

b) To minimize the probability of error of the detector, it suffices to minimize the conditional probability of error. In this case, since the decision becomes a function of X_1, $D = f(X_1)$, the detector becomes a random variable itself. Designing the detector consists in obtaining such function $f(\cdot)$. In this course, we only consider that $f(\cdot)$ is deterministic, i.e., if the same x_1 is observed twice the detector will produce the same output in both cases. This implies that we can alternatively interpret the goal of designing a detector as partitioning the observation space into as many regions as the number of hypotheses.
Using the results from the previous section, it follows that, to minimize the error at every point, we need to select the hypothesis with the largest *a posteriori* probability, i.e., the test statistic that results in a minimum probability of error is:

$$d(x_1) = \arg\max_i P_{H|X_1}(i|x_1).$$

This expression gives the name to the detection criterion, is known as the *Maximum a Posteriori* (MAP) detector. Actually, maximizing the *a posteriori* probability is the criterion that minimizes the probability of error in general.
Since $P_{H|X_1}(0|x_1=6) = P_{H|X_1}(1|x_1=6)$, for $x_1=6$ deciding in favor of either hypotheses results in the same probability of error $(1/2)$. For the remaining values, $d=0$ should be selected. Finally, using the law of total probability, the probability of error becomes

$$\begin{aligned} P_e = P(D\neq H) &= \sum_{x_1=1}^{6} P(D\neq H|x_1)P_{X_1}(x_1) \\ &= P(D\neq H|x_1=1)P_{X_1}(1)+P(D\neq H|x_1=2)P_{X_1}(2) \\ &\quad + P(D\neq H|x_1=3)P_{X_1}(3)+P(D\neq H|x_1=4)P_{X_1}(4) \\ &\quad + P(D\neq H|x_1=5)P_{X_1}(5)+P(D\neq H|x_1=6)P_{X_1}(6) \\ &= \frac{1}{6}\left[0+0+0+\frac{1}{6}+\frac{1}{3}+\frac{1}{2}\right] = \frac{1}{6}. \end{aligned}$$

c) In this part of the problem we need to minimize the expected cost. Similarly to what we did for the probability of error, we will first compute the expected cost associated to every decision and observation x_1, and then at each point we will simply select the decision criterion that incurs in a minimum expected cost.

$$\text{If } x_1\in\{1,2,3\} \rightarrow \begin{cases} d=0 \rightarrow \mathbb{E}\{C_{0H}|x_1\in\{1,2,3\}\} = 0 \\ d=1 \rightarrow \mathbb{E}\{C_{1H}|x_1\in\{1,2,3\}\} = c_{10}P_{H|X_1}(0|x_1\in\{1,2,3\}) = c_{10} = \frac{1}{4} \end{cases}$$

$$\text{If } x_1=4 \rightarrow \begin{cases} d=0 \rightarrow \mathbb{E}\{C_{0H}|x_1=4)\} = c_{01}P_{H|X_1}(1|x_1=4) = \frac{1}{6} \\ d=1 \rightarrow \mathbb{E}\{C_{1H}|x_1=4)\} = c_{10}P_{H|X_1}(0|x_1=4) = \frac{5}{24} \end{cases}$$

$$\text{If } x_1=5 \rightarrow \begin{cases} d=0 \rightarrow \mathbb{E}\{C_{0H}|x_1=5)\} = c_{01}P_{H|X_1}(1|x_1=5) = \frac{2}{6} \\ d=1 \rightarrow \mathbb{E}\{C_{1H}|x_1=5)\} = c_{10}P_{H|X_1}(0|x_1=5) = \frac{1}{6} \end{cases}$$

$$\text{If } x_1=6 \rightarrow \begin{cases} d=0 \rightarrow \mathbb{E}\{C_{0H}|x_1=6)\} = c_{01}P_{H|X_1}(1|x_1=6) = \frac{1}{2} \\ d=1 \rightarrow \mathbb{E}\{C_{1H}|x_1=6)\} = c_{10}P_{H|X_1}(0|x_1=6) = \frac{1}{8} \end{cases}$$

Then, the detector that minimizes the expected cost is

$$d^\star = \begin{cases} 0, & \text{if } X_1\in\{1,2,3,4\}, \\ 1, & \text{if } X_1\in\{5,6\}, \end{cases}$$

with the expected cost given by

$$\begin{aligned}\mathbb{E}\{C\} &= \sum_{x_1=1}^{6} \mathbb{E}\{C|x_1\} P_{X_1}(x_1) \\ &= \frac{1}{6}[0+0+0+\frac{1}{6}+\frac{1}{6}+\frac{1}{8}] \\ &= \frac{11}{6\cdot 24},\end{aligned}$$

which follows from the law of total probability. One final comment is in order. Using a detector that exploits the value of an observation variable, we were able to reduce the expected cost with respect to the value obtained in the first example.

So far, we have learned that the *a posteriori* probability of H given the observations plays a key role in detection problems. In the first two examples, obtaining such probability was rather straightforward given the inherent mechanism for the generation of the hypotheses: observations take place first, and the hypothesis depends directly on these observations. Now, we will consider the case in which the generation of the hypothesis occurs first, and then observations are drawn according to their probability distribution given the hypothesis. This scenario is frequently encountered in many real problems. When this is the case, one can more easily get access to the *likelihoods* of each hypothesis, and the *a posteriori* probabilities need to be evaluated exploiting Bayes' Theorem.

B.1.3 Example 3: Working the solution from the likelihoods

Problem B.3 Consider now a new game that involves two coins, one of them is fair whereas for the second one, the probability of heads doubles the probability of tails. In this game, a coin is first selected, and the goal is to guess which is the selected coin using as observations the result of flipping the coin n times. Therefore, this problem can also be seen as a hypothesis testing problem, where one has to decide whether the selected coin was the fair one (hypothesis $H = 0$) or the loaded one (hypothesis $H = 1$).

a) Without assuming any other information, design a detector for the aforementioned hypothesis test.
b) Discuss how you would design a detector that minimizes the probability of error, and what additional information you would need for that.

Solution B.3 We denote by $\mathbf{X}$ the vector that contains all the available observations to take the decision, i.e., the result of each coin flipping: $\mathbf{X} = (X^{(1)}, X^{(2)}, \ldots, X^{(n)})^\top$. Each of these variables can be a head or a tail: $X^{(i)} \in \{\circ, \times\}$. We will denote by $n_\circ$ and $n_\times$ the number of observed heads and tails, respectively. Obviously, we have $n = n_\circ + n_\times$.

a) The only statistical information available in this section is the probability of observing a head or a tail for both hypotheses:

$$P_{X^{(i)}|H}(\circ|0) = \frac{1}{2}, \qquad P_{X^{(i)}|H}(\times|0) = \frac{1}{2},$$

and

$$P_{X^{(i)}|H}(\circ|1) = \frac{2}{3}, \qquad P_{X^{(i)}|H}(\times|1) = \frac{1}{3}.$$

Now, since there are available n observations, we can also compute the joint probability of the observation vector $\mathbf{X}$:

$$P_{\mathbf{X}|H}(\mathbf{x}|0) = \left(\frac{1}{2}\right)^n, \qquad P_{\mathbf{X}|H}(\mathbf{x}|1) = \left(\frac{2}{3}\right)^{n_\circ}\left(\frac{1}{3}\right)^{n_\times}.$$

These two expressions above are the joint probabilities of all observed variables given the hypothesis, and are usually referred to as the likelihoods of hypothesis 0 and 1. Essentially, the likelihoods express how well the observed data can be explained by each of the hypotheses.

When the only available information is the likelihoods, a reasonable approach to follow is deciding in favor of the hypothesis that maximizes the likelihood. For this example, the so-called *maximum likelihood* (ML) detector is given by

$$P_{\mathbf{X}|H}(\mathbf{x}|0) \underset{D=1}{\overset{D=0}{\gtrless}} P_{\mathbf{X}|H}(\mathbf{x}|1) \Rightarrow \left(\frac{1}{2}\right)^n \underset{D=1}{\overset{D=0}{\gtrless}} \left(\frac{2}{3}\right)^{n_\circ}\left(\frac{1}{3}\right)^{n_\times}.$$

A convenient way to simplify this expression consists in taking logarithms on both sides of the inequality. Note that, in order to take logarithms, we need to make sure that the arguments thereof are strictly positive, which holds for both sides of the equation above. Then, taking logarithms and simplifying the resulting expression yields

$$(n_\circ + n_\times)\log\frac{1}{2} \underset{D=1}{\overset{D=0}{\gtrless}} n_\circ \log\frac{2}{3} + n_\times \log\frac{1}{3},$$

or, equivalently,

$$\frac{n_\times}{n_\circ} \underset{D=1}{\overset{D=0}{\gtrless}} \frac{\log\frac{2}{3} - \log\frac{1}{2}}{\log\frac{1}{2} - \log\frac{1}{3}}.$$

This equation translates into a partition of the observation space. In fact, we see that the detector does not depend on the value of particular observations, but just on the total number of heads and tails (i.e., the order in which the coin flippings are observed does not matter). Moreover, it also implies that a larger number of observed heads favors the decision $D = 1$, which aligns with the fact that the probability of heads is larger than the probability of tails when $H = 1$.

b) Now, we need to study the minimization of the probability of error, defined as

$$P_e = P(D \neq H) = \sum_{\mathbf{x}} P(d \neq H|\mathbf{X} = \mathbf{x}) P_{\mathbf{X}}(\mathbf{x}).$$

In order to grasp the meaning of P_e, we need to emphasize that for any particular detector, there is a deterministic relation between D and $\mathbf{X}$. Since the probability of error for a given observation vector is $P(d \neq H|\mathbf{X} = \mathbf{x})$, the expectation of this value needs to be taken with respect to $\mathbf{X}$ to obtain the probability of error. The minimization of P_e is equivalent to the minimization of each element in the above summation. That is, for each possible observation vector $\mathbf{x}$ we need to take the decision that minimizes the probability

of error for that particular value of $\mathbf{x}$. Since there are only two hypothesis, the probability of incurring in an error if we decide in favor of one of the hypothesis is the probability of the non-selected hypothesis, i.e.,

$$\text{If we decide } d=0 \quad \rightarrow \quad P(H \neq 0|\mathbf{X}=\mathbf{x}) = P_{H|\mathbf{X}}(1|\mathbf{x}),$$

$$\text{If we decide } d=1 \quad \rightarrow \quad P(H \neq 1|\mathbf{X}=\mathbf{x}) = P_{H|\mathbf{X}}(0|\mathbf{x}).$$

Therefore, in order to minimize the probability of error at each $\mathbf{x}$, and therefore to minimize the overall probability of error, we need to follow the criterion:

$$P_{H|\mathbf{X}}(1|\mathbf{x}) \underset{D=0}{\overset{D=1}{\gtrless}} P_{H|\mathbf{X}}(0|\mathbf{x}),$$

which is, as described above, the *Maximum a posteriori* (MAP) detector. In other words, maximizing the likelihood does not necessarily minimize the probability of error, which is actually minimized by maximizing the *a posteriori* probabilities of each hypotheses. This makes sense, since the likelihood just measures how well the observations fit with a given hypothesis, but ignores the *a priori* probability of the hypotheses. Then, we can decide in favor of a hypotheses with smaller likelihood if its *a priori* probability is sufficiently larger than the probability of the other hypothesis. This can be explicitly quantified by means of Bayes' Theorem, which states that

$$P_{H|\mathbf{X}}(h|\mathbf{x}) = \frac{P_{\mathbf{X}|H}(\mathbf{x}|h)P_H(h)}{P_{\mathbf{X}}(\mathbf{x})}.$$

Bayes' Theorem shows that the maximization of the *a posteriori* probability of each hypothesis (and therefore to minimize the probability of error) requires taking into account both the likelihoods and the *a priori* probabilities of the hypotheses.
In summary, in order to design a detector (or classifier) that minimizes the probability of error, we would need to know the *a priori* probability of each hypothesis. Moreover, if the goal were to minimize a cost function, we would still need to rely on *a posteriori* probabilities.

In the previous examples, we have introduced a number of important concepts in detection problems: hypotheses, *a priori* and *a posteriori* probability, likelihood, probability of error, and (expected) cost. We have also learned that, for the design of detectors when there are available observations, the distribution that provides **the most valuable information is the *a posteriori* distribution of the hypotheses given such observations**. If this distribution is available, we can compute the performance of **any** detector in terms of its probability of error or expected cost (performance analysis problems). Based on these performance metrics, we can also design detectors that minimize each criterion (design problem).