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

Contents

Chapter 1
Statistical Estimation Theory

1.1 Statistical Estimation Theory

1.1.1 General view of the estimation problem

The design of an estimator involves creating a real-valued function that, given an input vector, $\mathbf{x}$ of observational variables, makes predictions about a target variable, s.

We will assume some statistical dependency exists between the observations and the target. To do so, we model the observations and the target by means of random variables $\mathbf{X}$ and S, respectively[1]. The observation is a sample from an *observation space* $\mathscr{X}$ which, in general, will be a subset of $\mathbb{R}^n$. We will typically assume that the target variable is real, although the general formulation can be applied to multidimensional cases. A schematic view of the estimation problem is depicted in Fig. 1.1.

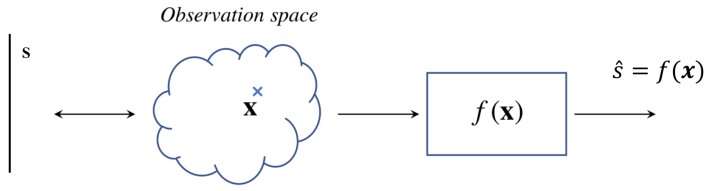


Fig. 1.1 Diagram block of estimation problems.

The estimation module applies a real output function $f(\cdot)$ that is commonly referred to as the *estimator*, and its output, $\hat{S} = f(\mathbf{X})$, as the *estimation* or *prediction*. The estimator is a deterministic function, meaning that for a given value $\mathbf{x}$, it will consistently produce the same output. Although $f(\cdot)$ is deterministic, if the input $\mathbf{X}$ is a random vector, the prediction $\hat{S}$ is a random variable.

The estimator is likely to incur some estimation error that will be quantified by means of a cost (or, alternatively, a reward) function. Designing our estimator will require minimizing (or maximizing) the expected value of this cost (reward).

We identify two main types of problems related to estimation:

- Analysis: given an estimator, evaluate its performance using a specific measure (a cost or a reward function).
- Design: find an estimator $f()$ that optimizes a predefined goal.

1.1.2 Probability model

The statistical relation between the observations and the target variable is described by the **joint** probability density function (pdf) of $\mathbf{X}$ and S: $p_{\mathbf{X},S}(\mathbf{x},s)$, or some distribution related to it.

[1] Note that we use capital letters to model the random variables, and lowercase letters to denote an arbitrary realization of them.

The joint pdf can be factorized as a product of conditional and marginal pdfs:

$$p_{\mathbf{X},S}(\mathbf{x},s) = p_{\mathbf{X}|S}(\mathbf{x}|s) \cdot p_S(s) = p_{S|\mathbf{X}}(s|\mathbf{x}) \cdot p_{\mathbf{X}}(\mathbf{x}) \tag{1.1}$$

In the context of estimation theory, these factors receive specific names:

- The **likelihood** of $S = s$ for observation $\mathbf{x}$, $p_{\mathbf{X}|S}(\mathbf{x}|s)$: it characterizes the generation of observations for each value of the target variable.
- The **prior (or *a priori*) distribution** of S, $p_S(s)$: it describes how much is known (or unknown) about the target variable before observing $\mathbf{X}$.
- The **posterior (or *a posteriori*) distribution** of S given $\mathbf{X} = \mathbf{x}$, $p_{S|\mathbf{X}}(s|\mathbf{x})$: it describes the knowledge (or the uncertainty) about S after observing $\mathbf{X}$.
- The **evidence** or marginal distribution of $\mathbf{X}$, $p_{\mathbf{X}}(\mathbf{x})$.

The information available to design the estimator may depend on the application. In some cases, the likelihood function is known, because it can be related to the physical generative process of the observations. If additionally, a prior distribution is available, the design can be grounded on the posterior distribution $p_{S|\mathbf{X}}(s|\mathbf{x})$, which can be calculated by means of Bayes' Theorem,

$$p_{S|\mathbf{X}}(s|\mathbf{x}) = \frac{p_{\mathbf{X},S}(\mathbf{x},s)}{p_{\mathbf{X}}(\mathbf{x})} = \frac{p_{\mathbf{X}|S}(\mathbf{x}|s)p_S(s)}{\int p_{\mathbf{X}|S}(\mathbf{x}|s')p_S(s')ds'} \tag{1.2}$$

1.1.3 Cost functions for estimation problems

The evaluation and design of an estimator require some objective criteria. In some cases, we will consider that this criterion materializes in the form of a **cost function** whose value we seek to minimize.

A cost function $c(s,\hat{s})$ is any measure of the discrepancy between the target variable and the estimation. It is generally non-negative, $c(s,\hat{s}) \geq 0$, with equality for $s = \hat{s}$. In some cases, the cost function can be expressed as a function of the estimation error $e = s - \hat{s}$ and we will write[2] $c(s,\hat{s}) = c(s-\hat{s}) = c(e)$. Some frequently used cost functions are:

- Quadratic cost: $c(e) = e^2$.
- Absolute error: $c(e) = |e|$.
- Relative quadratic error: $c(s,\hat{s}) = \frac{(s-\hat{s})^2}{s^2}$
- Cross Entropy: $c(s,\hat{s}) = -s\ln\hat{s} - (1-s)\ln(1-\hat{s})$, for $s,\hat{s} \in [0,1]$

Since the target variable is unknown, the prediction cannot be computed by directly minimising the cost $c(s,\hat{s})$, and we have to work with expectations. The expected value of the cost is usually referred as the **risk** of an estimator $\hat{s} = f(\mathbf{x})$:

$$R_f = \mathbb{E}\{c(S,\hat{S})\} = \int_{\mathbf{x}}\int_{s} c(s,f(\mathbf{x}))p_{S,\mathbf{X}}(s,\mathbf{x})dsd\mathbf{x} \tag{1.3}$$

[2] Note that the cost function is denoted with a lowercase letter, c, because it is a deterministic function, i.e., for fixed values of s and $\hat{s}$ the cost always takes the same value. However, as with the estimation function, the application of that function to random variables will result in another random variable, i.e., $C = c(S,\hat{S})$.

By the Law of Large Numbers, this is the average cost we can expect from a given estimator, after a large number of predictions.

The **conditional risk** is the conditional mean for a given observation

$$R(\hat{s},\mathbf{x}) = \mathbb{E}\{c(S,\hat{s})|\mathbf{x}\} = \int_s c(s,\hat{s}) p_{S|\mathbf{X}}(s|\mathbf{x}) ds \tag{1.4}$$

Example 1.1 (Evaluation of estimators 1)
Given the joint distribution

$$p_{S,X}(s,x) = \begin{cases} \frac{1}{x}, & 0 \le s \le x \text{ and } 0 < x \le 1 \\ 0, & \text{otherwise} \end{cases}, \tag{1.5}$$

consider the estimators $\hat{S}_1 = \frac{1}{2}X$ and $\hat{S}_2 = X$. Which is the best estimator from the point of view of the quadratic cost? To find out, we'll calculate the mean quadratic error for both estimators. Knowing that, for any w,

$$\begin{aligned} \mathbb{E}\{(S-wX)^2\} &= \int_0^1 \int_0^x (s-wx)^2 p_{S,X}(s,x) ds dx = \int_0^1 \int_0^x (s-wx)^2 \frac{1}{x} ds dx \\ &= \int_0^1 \left(\frac{1}{3} - w + w^2\right) x^2 dx = \frac{1}{3}\left(\frac{1}{3} - w + w^2\right) \end{aligned} \tag{1.6}$$

Taking $w = 1/2$ and $w = 1$ we get, respectively,

$$\mathbb{E}\left\{(S-\hat{S}_1)^2\right\} = \mathbb{E}\left\{\left(S - \frac{1}{2}X\right)^2\right\} = \frac{1}{3}\left(\frac{1}{3} - \frac{1}{2} + \frac{1}{4}\right) = \frac{1}{36} \tag{1.7}$$

$$\mathbb{E}\{(S-\hat{S}_2)^2\} = \mathbb{E}\{(S-X)^2\} = \frac{1}{3}\left(\frac{1}{3} - 1 + 1\right) = \frac{1}{9} \tag{1.8}$$

Therefore, from the point of view of the square error, $\hat{S}_1$ is a better estimator than $\hat{S}_2$.

Example 1.2 (Evaluation of estimators 2) Assume that S is a random variable of mean 0 and variance 1, and X is a noisy observation of S,

$$X = S + R \tag{1.9}$$

where R is a random Gaussian variable, independent of S, of mean 0 variance v. We will compute the risk for the estimator $\hat{S} = X$ and different cost functions. For the quadratic error:

$$\mathbb{E}\{(S-\hat{S})^2\} = \mathbb{E}\{(S-X)^2\} = \mathbb{E}\{R^2\} = v \tag{1.10}$$

For the mean absolute error

$$\begin{aligned} \mathbb{E}\{|S-\hat{S}|\} = \mathbb{E}\{|R|\} &= \int_{-\infty}^{\infty} |r| \frac{1}{\sqrt{2\pi v}} \exp\left(-\frac{r^2}{2v}\right) dr \\ &= 2\int_0^{\infty} r \frac{1}{\sqrt{2\pi v}} \exp\left(-\frac{r^2}{2v}\right) dr = \sqrt{\frac{2v}{\pi}} \end{aligned} \tag{1.11}$$

1.1.4 Bias and variance

The bias of estimator $\hat{S}$ for a true target $S = s$ is defined as

$$B(s) = \mathbb{E}\{\hat{S}|S = s\} - s \tag{1.12}$$

and it accounts for the expected deviation of the estimator from the true value of the target variable.

Note that, in general, the bias depends on the target. If a prior model for S, is available, we can compute the expected value to get

$$B = \mathbb{E}\{B(S)\} = \mathbb{E}\{\hat{S}\} - \mathbb{E}\{S\} \tag{1.13}$$

The variance of estimator $\hat{S}$ for a true target $S = s$ is defined as

$$V = \text{var}\{\hat{S}|S = s\} \tag{1.14}$$

The variance accounts for the spread of the sampling distribution, or in other words, it quantifies how much the estimates vary from one sample $\mathbf{x}$ to another, for the same realization of S. Unlike bias, which assesses systematic deviation from the true parameter value, variance captures the randomness inherent in the estimation process due to sampling variability.

1.2 Design of estimators

1.2.1 Maximum Likelihood (ML) estimation

The maximum likelihood estimator (ML) uses the likelihood as the reward function to be maximized:

$$\hat{s}_{\mathrm{ML}} = \underset{s}{\operatorname{argmax}}\, p_{\mathbf{X}|S}(\mathbf{x}|s) = \underset{s}{\operatorname{argmax}} \ln(p_{\mathbf{X}|S}(\mathbf{x}|s)) \tag{1.15}$$

The ML estimator selects the value of the parameter s that maximizes the likelihood of observing $\mathbf{x}$ when $S = s$. Loosely speaking, observing $\mathbf{x}$ when $S = \hat{s}_{\mathrm{ML}}$ is less unexpected than if S takes any other value. Note that $p_{\mathbf{X}|S}(\mathbf{x}|s)$, which is a density function over random variable $\mathbf{X}$, is not maximized with respect to $\mathbf{x}$, but s.

Example 1.3 (ML Estimation)

We want to estimate the value of a random variable S from an observation X statistically related to it. For the design of the estimator, only the likelihood of S is known, which is given by

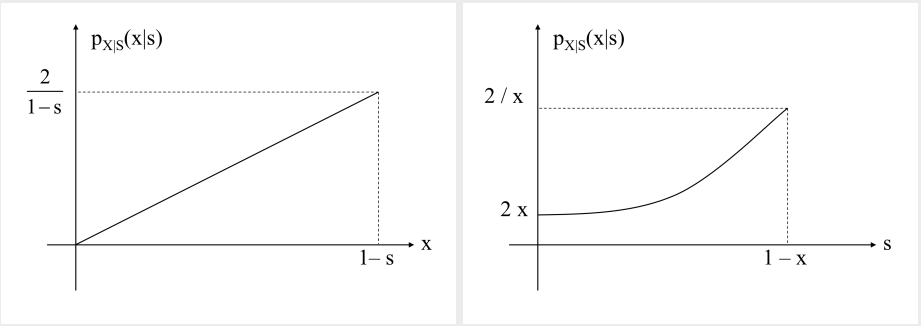
$$p_{X|S}(x|s) = \frac{2x}{(1-s)^2},\ 0 \le x \le 1-s,\ 0 \le s \le 1 \tag{1.16}$$

Given the available statistical information, it is decided to construct the ML estimator of S. The likelihood function represents the probability density of the random variable X, normalized to have unit area, as represented in Figure 1.2(a). However, to carry out the maximization, representing this likelihood as a function[a] of s (Fig.1.2(b)) is more useful, as it shows that the estimator is

$$\hat{s}_{\mathrm{ML}} = 1 - x$$

or, alternatively, if we consider the application of the estimation function on the random variable X instead of on a specific value of it,

$$\hat{S}_{\mathrm{ML}} = 1 - X$$



(a) (b)

Fig. 1.2 Representation of the likelihood distribution of the example 1.3 as a function of x and s.

[a] Note that the integral with respect to s of $p_{X|S}(x|s)$ will not generally be the unit, since this function does not constitute a probability density of S.

Note that the second equality in (1.15) states that the maximization of the likelihood is equivalent to the maximization of its logarithm (the **log-likelihood** function). Since the logarithm function is strictly increasing, $p_{\mathbf{X}|S}(\mathbf{x}|s_1) > p_{\mathbf{X}|S}(\mathbf{x}|s_2)$ implies $\ln(p_{\mathbf{X}|S}(\mathbf{x}|s_1)) >$

$\ln(p_{\mathbf{X}|S}(\mathbf{x}|s_2)))$ and, thus, the logarithm does not alter the outcome of the maximization. The logarithm is used by practical reasons when the likelihood is a product of several factors or an exponential function, as it will transform products into sums and an exponential into the exponents. In this way, the maximization process can be simplified considerably.

Note that the maximum likelihood does not need any probability model about the target variable, S, which is treated as a deterministic parameter. This is useful in situations where only the likelihood function is known.

1.2.2 Maximum a posteriori (MAP) estimation

We define the maximum a posteriori (MAP) estimator as the mode of the posterior distribution, that is

$$\hat{s}_{\text{MAP}} = \underset{s}{\operatorname{argmax}}\, p_{S|\mathbf{X}}(s|\mathbf{x}) = \underset{s}{\operatorname{argmax}} \ln(p_{S|\mathbf{X}}(s|\mathbf{x})) \tag{1.17}$$

Using the definition of conditional pdf and Bayes' rule, it is easy to see that the MAP estimator can be computed as

$$\hat{s}_{\text{MAP}} = \underset{s}{\operatorname{argmax}}\, p_{S,\mathbf{X}}(s,\mathbf{x}) \tag{1.18}$$

$$= \underset{s}{\operatorname{argmax}} \{p_{\mathbf{X}|S}(\mathbf{x}|s)p_S(s)\} \tag{1.19}$$

Eq. (1.19) demonstrates that the MAP estimator seeks to maximize the likelihood function, modulated by the prior distribution. This integration of the prior distribution allows the MAP estimator to incorporate existing knowledge or assumptions about the target S before observing the data, $\mathbf{x}$. When the prior distribution, is uniform across the entire range of possible values of S, the distinction between the MAP and ML estimators vanishes, as the MAP estimator effectively reduces to the ML estimator.

However, when the prior distribution is not uniform, the MAP estimator is biased towards values of the target variable with higher prior probabilities. This shift illustrates the MAP estimator's sensitivity to prior knowledge.

Beyond their mathematical formulations, the MAP and ML estimators embody fundamentally different inference philosophies. The ML estimator optimizes the likelihood function, which models the probability of the observed data under various values of the target, treating s as a fixed but unknown parameter. This approach aligns with the **frequentist paradigm**, which interprets probability as the long-run frequency of events and does not incorporate prior information about S.

Conversely, the MAP estimator embraces a **Bayesian framework**, treating the target as a random variable. This perspective allows the incorporation of prior knowledge or beliefs about S through the prior distribution, $p_S(s)$, and the posterior distribution, $p_{S|\mathbf{X}}(s|\mathbf{x})$, updates this knowledge based on new evidence from the data. The Bayesian approach, therefore, provides a probabilistic framework for updating beliefs about uncertain parameters in light of new data.

Example 1.4 (Estimation MAP) Considering that

$$p_{S|X}(s|x) = \frac{1}{x^2} s \exp\left(-\frac{s}{x}\right), \qquad x \geq 0, \quad s \geq 0 \tag{1.20}$$

the MAP estimator can be computed by maximizing

$$\ln(p_{S|X}(s|x)) = -2\ln(x) + \ln(s) - \frac{s}{x}, \qquad x \geq 0, \quad s \geq 0, \tag{1.21}$$

Since $\ln(p_{S|X}(s|x))$ tends to $-\infty$ around $s = 0$ and $s = \infty$, its maximum must be at some intermediate point with zero derivative. Deriving respect to s results in

$$\frac{\partial}{\partial s} \ln p_{S|X}(s|x)\Big|_{s=\hat{s}_{\text{MAP}}} = \frac{1}{\hat{s}_{\text{MAP}}} - \frac{1}{x} = 0, \qquad x \geq 0, \quad s \geq 0 \tag{1.22}$$

Thus,

$$\hat{s}_{\text{MAP}} = x \tag{1.23}$$

1.2.3 Minimum risk estimators

When a cost function is used to evaluate the quality of an estimation for a given estimation problem, one may wonder if we can find a mathematical expression for the estimator minimizing the mean value of the cost, that is, the risk.

Taking back the formula of the risk, and applying the total expectation theorem, we find

$$R_f = \mathbb{E}\{c(S,\hat{S})\} = \int_{\mathbf{x}} \mathbb{E}\{c(S,\hat{s})|\mathbf{X} = \mathbf{x}\} p_{\mathbf{X}}(\mathbf{x}) \mathbf{dx} \tag{1.24}$$

where $\hat{s} = f(\mathbf{x})$. That is, the risk is the integral of the conditional risk, and, thus, the estimator minimizing the risk will be such that it minimizes the conditional risk for each observation $\mathbf{x}$,

$$\hat{s}^* = \underset{\hat{s}}{\operatorname{argmin}}\ \mathbb{E}\{c(S,\hat{s})|\mathbf{X} = \mathbf{x}\} \tag{1.25}$$

We will refer to this estimator as the **Bayesian estimator** associated with cost function $c()$.

Example 1.5 (Calculation of a minimum mean square error estimator)
Following the example 1.1, we can calculate the posterior distribution of S through

$$p_{S|X}(s|x) = \frac{p_{S,X}(s,x)}{p_X(x)}. \tag{1.26}$$

Knowing that

$$p_X(x) = \int_0^1 p_{S,X}(s,x)ds = \int_0^x \frac{1}{x} ds = 1, \qquad 0 \leq x \leq 1 \tag{1.27}$$

we obtain

$$p_{S|X}(s|x) = \begin{cases} \frac{1}{x}, & 0 < s < x < 1 \\ 0, & \text{otherwise} \end{cases} \tag{1.28}$$

The conditional risk will be given by

$$\begin{aligned}\mathbb{E}\{c(S,\hat{s})|X=x\} &= \mathbb{E}\{(S-\hat{s})^2|X=x\} \\ &= \int_0^1 (s-\hat{s})^2 p_{S|X}(s|x)ds \\ &= \frac{1}{x}\int_0^x (s-\hat{s})^2 ds = \frac{1}{x}\left(\frac{(x-\hat{s})^3}{3}+\frac{\hat{s}^3}{3}\right) \\ &= \frac{1}{3}x^2 - \hat{s}x + \hat{s}^2. \end{aligned} \tag{1.29}$$

As a function of $\hat{s}$, conditional risk is a second-degree polynomial, whose minimum can be calculated through differentiation. Since

$$\frac{d}{d\hat{s}}\mathbb{E}\{c(S,\hat{s})|X=x\} = -x+2\hat{s}, \tag{1.30}$$

the Bayesian estimator associated with the quadratic error is

$$\hat{s}^* = \frac{1}{2}x, \tag{1.31}$$

which matches the estimator $\hat{S}_1$ from the example 1.1. Therefore, $\hat{S}_1$ is the best possible estimator from the point of view of the mean square error.

Based on (1.25) we can conclude that, regardless of the cost to be minimized, the knowledge of the posterior distribution of S given $\mathbf{X}$, $p_{S|\mathbf{X}}(s|\mathbf{x})$, is sufficient to design the Bayesian estimator for a given cost. As mentioned above, this distribution is often calculated from the likelihood of S and its a priori distribution using the Bayes Theorem, which is in fact the origin of the denomination of these estimators.

1.3 Common Bayesian estimators

This section presents some of the most commonly used Bayesian estimators. For their calculation, we will proceed to minimize the mean cost given $\mathbf{X}$ (posterior mean cost) for different cost functions.

1.3.1 Minimum Mean Squared Error estimator (MSE)

The minimum mean squared error (MSE) estimator is the Bayesian estimator associated with the cost function $c(e) = e^2 = (s - \hat{s})^2$, and therefore is given by

$$\begin{aligned} \hat{s}_{\text{MSE}} &= \underset{\hat{s}}{\operatorname{argmin}}\ \mathbb{E}\{c(S,\hat{s})|\mathbf{X}=\mathbf{x}\} \\ &= \underset{\hat{s}}{\operatorname{argmin}}\ \mathbb{E}\{(S-\hat{s})^2|\mathbf{X}=\mathbf{x}\} \end{aligned} \tag{1.32}$$

Figure 1.3 illustrates the minimum MSE estimation problem. The risk can be obtained by integrating (with respect to s) the product of the square error and the posterior pdf of S. The argument for minimization is $\hat{s}$, which allows shifting the graph corresponding to the cost function (represented with discontinuous stroke) so that the result of that integral is minimal.

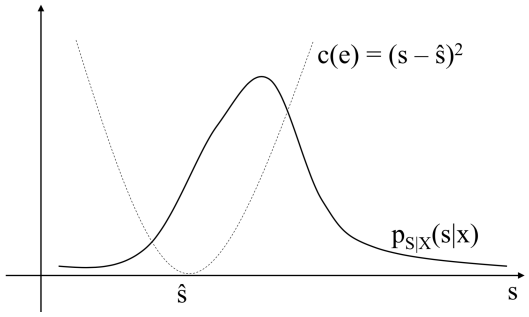


Fig. 1.3 Graphical representation of the process of calculating the posterior mean for a generic value $\hat{s}$.

For the square error, the conditional risk in (1.32) becomes

$$\mathbb{E}\{(S-\hat{s})^2|\mathbf{X}=\mathbf{x}\} = \mathbb{E}\{S^2|\mathbf{X}=\mathbf{x}\} - 2\mathbb{E}\{S|\mathbf{X}=\mathbf{x}\}\hat{s} + \hat{s}^2 \tag{1.33}$$

This is a second-degree polynomial that can be minimized by differentiation to get

$$\hat{s}_{\text{MSE}} = \mathbb{E}\{S|\mathbf{X}=\mathbf{x}\} = \int s\, p_{S|\mathbf{X}}(s|x)ds \tag{1.34}$$

In other words, the minimum MSE estimator of S is the posterior mean of S given $\mathbf{X}$.

Example 1.6 (Straightforward calculation of the MSE estimator) According to (1.34), minimum mean squared error estimator obtained in 1.1 can alternatively be derived as follows

$$\hat{s}_{\text{MSE}} = \int_0^1 s p_{S|X}(s|x) ds = \int_0^x \frac{s}{x} ds = \frac{1}{2}x \tag{1.35}$$

which is consistent with (1.31).

1.3.2 Minimum Mean Absolute Deviation Estimator (MAD)

In the same way, as we have proceeded in the case of the estimator $\hat{s}_{\text{MSE}}$, we can calculate the estimator associated with the absolute deviation of the estimation error, $c(e) = |e| = |s - \hat{s}|$. This estimator, which we will refer to as the Mean Absolute Deviation (MAD), is characterized by

$$\hat{s}_{\text{MAD}} = \underset{\hat{s}}{\text{argmin}}\ \mathbb{E}\{|S - \hat{s}|\ |\mathbf{X} = \mathbf{x}\} = \tag{1.36}$$

$$= \underset{\hat{s}}{\text{argmin}} \int_s |s - \hat{s}|\ p_{S|\mathbf{X}}(s|\mathbf{x}) ds \tag{1.37}$$

Again, it is simple to illustrate the process of calculating the posterior mean cost by overlapping on the same axes the cost expressed as a function of s and the posterior distribution of the variable to be estimated (see Fig. 1.4). This representation also suggests the convenience of splitting the integral into two parts corresponding to the two slopes of the cost function:

$$\mathbb{E}\{|S - \hat{s}|\ |\mathbf{X} = \mathbf{x}\} = \int_{-\infty}^{\hat{s}} (\hat{s} - s)\ p_{S|\mathbf{X}}(s|\mathbf{x}) ds + \int_{\hat{s}}^{\infty} (s - \hat{s})\ p_{S|\mathbf{X}}(s|\mathbf{x}) ds \tag{1.38}$$

$$= \hat{s} \left[\int_{-\infty}^{\hat{s}} p_{S|\mathbf{X}}(s|\mathbf{x}) ds - \int_{\hat{s}}^{\infty} p_{S|\mathbf{X}}(s|\mathbf{x}) ds \right] \tag{1.39}$$

$$+ \int_{\hat{s}}^{\infty} s\ p_{S|\mathbf{X}}(s|\mathbf{x}) ds - \int_{-\infty}^{\hat{s}} s\ p_{S|\mathbf{X}}(s|\mathbf{x}) ds \tag{1.40}$$

The fundamental theorem of calculus[3] allows us to obtain the derivative of the conditional risk as

$$\frac{d\mathbb{E}\{|S - \hat{s}|\ |\mathbf{X} = \mathbf{x}\}}{d\hat{s}} = 2F_{S|\mathbf{X}}(\hat{s}|\mathbf{x}) - 1 \tag{1.41}$$

where $F_{S|\mathbf{X}}(s|\mathbf{x})$ is the posterior distribution function of S given $\mathbf{X}$. Since this derivative must vanish at the minimum, we get $F_{S|\mathbf{X}}(\hat{s}_{\text{MAD}}|\mathbf{x}) = 1/2$. In other words, the minimum MAD estimator is given by the median of $p_{S|\mathbf{X}}(s|\mathbf{x})$:

[3] $\frac{d}{dx} \int_{t_0}^{x} g(t) dt = g(x)$.

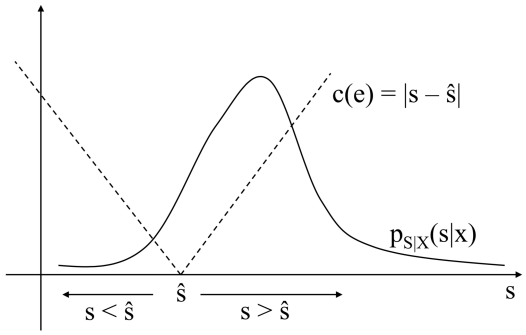


Fig. 1.4 Calculation of the posterior mean absolute error for a generic value $\hat{s}$.

$$\hat{s}_{\text{MAD}} = \text{median}\{S|\mathbf{X} = \mathbf{x}\} \tag{1.42}$$

Remember that the median of a distribution is the point that separates that distribution into two regions that have the same probability, so the minimum MAD estimator satisfies

$$P\{S > \hat{s}_{\text{MAD}}|\mathbf{x}\} = P\{S < \hat{s}_{\text{MAD}}|\mathbf{x}\} \tag{1.43}$$

In practice, this can be computed as the solution of

$$\int_{-\infty}^{\hat{s}_{\text{MAD}}} p_{S|\mathbf{X}}(s|\mathbf{x})ds = \frac{1}{2} \tag{1.44}$$

Example 1.7 (Design of a Minimum Mean Absolute Deviation Estimator)
In the scenario of the example 1.1, the posterior distribution of S given X is uniform between 0 and x, the median of which is $x/2$. Thus,

$$\hat{s}_{\text{MAD}} = \frac{1}{2}x \tag{1.45}$$

Note that, in this case, the MAD estimator matches the MSE obtained at (1.31). This is a consequence of the symmetry of the a posterior distribution.

1.4 Estimation with constrains

1.4.1 General principles

Occasionally, it might be beneficial to prescribe a specific parametric form for the estimator, denoted as $\hat{S} = f_{\mathbf{w}}(\mathbf{X})$, where $\mathbf{w}$ represents a vector of parameters. For instance, in scenarios involving two observations, $\mathbf{X} = [X_1, X_2]^T$, a design constraint might necessitate limiting the search for an estimator to the family of quadratic estimators, characterized by $\hat{S} = w_0 + w_1 X_1^2 + w_2 X_2^2$. In such situations, the task of designing an estimator consists of identifying the optimal parameter vector $\mathbf{w}^*$ that minimizes the risk, while adhering to the specified constraints on the estimator structure:

$$\begin{aligned}\mathbf{w}^* &= \underset{\mathbf{w}}{\operatorname{argmin}}\, \mathbb{E}\{c(S,\hat{S})\} = \underset{\mathbf{w}}{\operatorname{argmin}}\, \mathbb{E}\{c(S, f_{\mathbf{w}}(\mathbf{X}))\} \\ &= \underset{\mathbf{w}}{\operatorname{argmin}} \int_{\mathbf{x}} \int_{s} c(s, f_{\mathbf{w}}(\mathbf{x})) p_{S,\mathbf{X}}(s,\mathbf{x}) ds d\mathbf{x}. \end{aligned} \tag{1.46}$$

Restricting the estimator to a specific analytic form typically results in a higher risk than what could be achieved with a Bayesian estimator tailored to the same cost function. The exception to this rule occurs when the imposed constraints align with the optimal estimator form, essentially when the Bayesian estimator inherently fits within the specified constraints. Despite potentially incurring higher costs, practical considerations may justify opting for such constrained estimators, such as simplification in design or implementation. An exploration of this concept is presented in Section 1.4.2, focusing on linear estimators that achieve minimum MSE.

Example 1.8 (Calculating an Estimator with Constrains)

Continuing the example 1.5, we want to calculate the minimum MSE estimator that has the form $\hat{s} = wx^2$. Starting from the conditional risk calculated in (1.29), the expression of the global average cost can be obtained as

$$\mathbb{E}\{c(S,\hat{S})\} = \int_x \mathbb{E}\{c(S,\hat{s})|X = x\}\, p_X(x) dx \tag{1.47}$$

$$= \int_x \left(\frac{1}{3}x^2 - \hat{s}x + \hat{s}^2 \right) p_X(x) dx \tag{1.48}$$

Forcing $\hat{s} = wx^2$ and taking into account that $p_X(x) = 1$ for $0 < x < 1$, we get the MSE as a function of w.

$$\mathbb{E}\{c(S,wX^2)\} = \int_0^1 \left(\frac{1}{3}x^2 - wx^3 + w^2x^4 \right) dx \tag{1.49}$$

$$= \frac{1}{9} - \frac{1}{4}w + \frac{1}{5}w^2 \tag{1.50}$$

The value w^* that optimizes (1.50) can be calculated by differentiation:

$$\frac{d}{d\hat{w}} \mathbb{E}\{c(S,w\mathbf{X}^2)\}\Big|_{w=w^*} = -\frac{1}{4} + \frac{2}{5}w^* = 0, \tag{1.51}$$

$$w^* = \frac{5}{8}, \tag{1.52}$$

and therefore the estimator is $\hat{s} = \frac{5}{8}x^2$.

1.4.2 Linear (in the parameters) estimation of minimum MSE

In this section we will focus on the study of estimators that are a linear combination of variables related to the observations, using the minimization of the MSE as a design criterion. More specifically, we will consider estimators given by the general expression

$$\hat{S} = \mathbf{w}^\intercal \mathbf{Z} \tag{1.53}$$

where

$$\mathbf{Z} = \phi(\mathbf{X}) \tag{1.54}$$

is some known transformation of the observations. The nature of this transformation may depend on the application, but there are some cases of particular interest:

- **Linear estimation**: in this case, ϕ is the identity function, so that $\mathbf{Z} = \mathbf{X}$, and the estimator a linear function of the observations

$$\hat{S} = w_0 X_0 + w_1 X_1 + \cdots + w_{N-1} X_{N-1} \tag{1.55}$$

- **Linear estimation with independent term**: in this case, $\mathbf{Z} = (1, \mathbf{X}^\intercal)^\intercal$ and, thus the estimator includes a constant term w_0

$$\hat{S} = w_0 + w_1 X_0 + \cdots + w_N X_{N-1} \tag{1.56}$$

- **Polynomial estimation**: in this case, the components of $\mathbf{Z}$ are monomials of the observational variables. For instance, for a scalar observation $X \in \mathbb{R}$, we can take $\mathbf{Z} = (1, X, X^2, \ldots X^M)$, for some $M > 1$, and the estimation becomes a polynomial of the observation with degree M.

$$\hat{S} = w_0 + w_1 X + w_2 X^2 \cdots + w_M X^M \tag{1.57}$$

Note that in general, the dimensions of $\mathbf{X}$ and $\mathbf{Z}$ may be different. Note, also, that, despite the estimator may be a non-linear function of the observation, all estimators given by (1.53) are linear functions of the parameters. For this reason, we will refer to them as estimators that are *linear in the parameters*.

By imposing a restriction on the analytic form of the estimator, linear-in-the-parameter estimators will generally obtain lower performance than the optimal Bayesian estimator. However, the interest in linear estimators is justified by their simplicity and ease of design. As we shall see, the linear estimator of minimum MSE depends exclusively on first and second-order statistical moments (means and covariances) associated with the target variable and the transformed observation, $\mathbf{Z}$.

1.4.2.1 Minimization of the mean squared error.

We will consider as design criteria the squared error, $c(e) = (s - \hat{s})^2$, so the optimal weight vector will be the one that minimizes the MSE risk:

$$\mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{argmin}}\ \mathbb{E}\{(S - \hat{S})^2\} = \underset{\mathbf{w}}{\operatorname{argmin}}\ \mathbb{E}\{(S - \mathbf{w}^\intercal \mathbf{Z})^2\} \tag{1.58}$$

and we will refer to the estimator associated with the weight vector as $\hat{S}_{\text{LMSE}}$:

$$\hat{S}_{\text{LMSE}} = \mathbf{w}^{*T}\mathbf{Z}$$

The MSE can be expanded as

$$\begin{aligned} MSE &= \mathbb{E}\{(S - \hat{S})^2\} = \mathbb{E}\{(S - \mathbf{w}^\intercal \mathbf{Z})^2\} \\ &= \mathbb{E}\{S^2\} - 2\mathbb{E}\{\mathbf{w}^\intercal \mathbf{Z} S\} + \mathbb{E}\{(\mathbf{w}^\intercal \mathbf{Z})^2\} \\ &= \mathbb{E}\{S^2\} - 2\mathbb{E}\{S\mathbf{Z}\}^\intercal \mathbf{w} + \mathbf{w}^\intercal \mathbb{E}\{\mathbf{Z}\mathbf{Z}^\intercal\}\mathbf{w} \\ &= \mathbb{E}\{S^2\} - 2\mathbf{r}_{S\mathbf{Z}}^\intercal \mathbf{w} + \mathbf{w}^\intercal \mathbf{R}_{\mathbf{Z}}\mathbf{w} \end{aligned} \tag{1.59}$$

where

- $\mathbf{r}_{S\mathbf{Z}} = \mathbb{E}\{S\mathbf{Z}\}$ is the **cross-correlation vector** between S and $\mathbf{Z}$
- $\mathbf{R}_{\mathbf{Z}} = \mathbb{E}\{\mathbf{Z}\mathbf{Z}^\intercal\}$ is the **autocorrelation matrix** of $\mathbf{Z}$.

Thus, the MSE is a second-degree polynomial in $\mathbf{w}$. This is illustrated in Figure 1.5, which depicts the error surface in a case with two observations. Being the function to minimize quadratic in weights (minimization argument), the error surface will take the form of a N dimensional paraboloid.

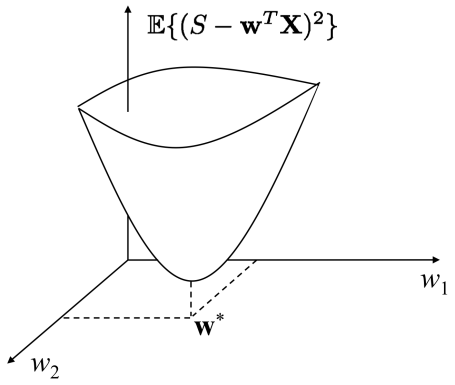


Fig. 1.5 Surface of the MSE for the linear estimator with $\mathbf{Z} = \mathbf{X}$, as a function of the parameters.

Since the MSE is non-negative, it is guaranteed to be a convex function of $\mathbf{w}$ and, thus, its minimum must be located at a point with zero gradient[4] (with respect to $\mathbf{w}$):

$$\nabla_{\mathbf{w}} MSE|_{\mathbf{w}=\mathbf{w}^*} = -2\mathbf{r}_{S\mathbf{Z}} + 2\mathbf{R}_{\mathbf{Z}}\mathbf{w}^* = \mathbf{0} \tag{1.60}$$

therefore, the optimal weight vector is any solution of

$$\mathbf{R}_{\mathbf{Z}}\mathbf{w}^* = \mathbf{r}_{S\mathbf{Z}} \tag{1.61}$$

If the autocorrelation matrix is invertible, we get

$$\boxed{\mathbf{w}^* = \mathbf{R}_{\mathbf{Z}}^{-1}\mathbf{r}_{S\mathbf{Z}}} \tag{1.62}$$

1.4.2.2 Alternative expression for the linear case

We can obtain an alternative expression for the optimal weights in the case of the linear estimator with the independent term in (1.56), which is given by $\mathbf{Z} = (1, \mathbf{X}^\intercal)^\intercal$ and can thus be written as

$$\hat{S} = w_0 + \mathbf{w}_{1:}^\intercal \mathbf{X} \tag{1.63}$$

where $\mathbf{w}_{1:}$ is the weight vector after removing the first component, w_0. To do so, we can re-express (1.61) as the equivalent pair of equations

$$\mathbb{E}\{w_0^* + \mathbf{w}_{1:}^{*\intercal}\mathbf{X}\} = \mathbb{E}\{S\} \tag{1.64}$$

$$\mathbb{E}\{\mathbf{X}\}w_0^* + \mathbb{E}\{\mathbf{X}\mathbf{X}^\intercal\}\mathbf{w}_{1:}^* = \mathbb{E}\{S\mathbf{X}\} \tag{1.65}$$

that is

$$w_0^* = m_S - \mathbf{w}_{1:}^{*\intercal}\mathbf{m}_{\mathbf{x}} \tag{1.66}$$

$$\mathbb{E}\{\mathbf{X}\mathbf{X}^\intercal\}\mathbf{w}_{1:}^* = \mathbb{E}\{S\mathbf{X}\} - w_0^*\mathbf{m}_{\mathbf{X}} \tag{1.67}$$

where $m_S = \mathbb{E}\{S\}$, $m_{\mathbf{X}} = \mathbb{E}\{\mathbf{X}\}$. Now using the expressions that relate the correlation and covariance of two variables:

$$\mathbb{E}\{S\mathbf{X}\} = \mathbf{v}_{S\mathbf{X}} + m_S\mathbf{m}_{\mathbf{X}} \tag{1.68}$$

$$\mathbb{E}\{\mathbf{X}\mathbf{X}^\intercal\} = \mathbf{V}_{\mathbf{X}} + \mathbf{m}_{\mathbf{X}}\mathbf{m}_{\mathbf{X}}^\intercal \tag{1.69}$$

we get

$$w_0^* = m_S - \mathbf{w}_{1:}^{*\intercal}\mathbf{m}_{\mathbf{x}} \tag{1.70}$$

$$\mathbf{v}_{S\mathbf{X}} + m_S\mathbf{m}_{\mathbf{X}} = (\mathbf{V}_{\mathbf{X}} + \mathbf{m}_{\mathbf{X}}\mathbf{m}_{\mathbf{X}}^T)\mathbf{w}_{1:}^* + w_0^*\mathbf{m}_{\mathbf{X}} \tag{1.71}$$

[4] The gradient of a function scale $f(\mathbf{w})$ with respect to the vector $\mathbf{w}$ is defined as a vector formed by the derivatives of the function with respect to each one of the components of $\mathbf{w}$: $\nabla_{\mathbf{w}} f(\mathbf{w}) = \left[\frac{\partial f}{\partial w_1}, \ldots \frac{\partial f}{\partial w_N}\right]^T$.

Solving these equations for w_0^* and $w_{1:}^*$ we get

$$w_0^* = m_S - \mathbf{w}_{1:}^{*\,T}\mathbf{m}_{\mathbf{x}} \tag{1.72}$$

$$\mathbf{w}_{1:}^* = \mathbf{V}_{\mathbf{X}}^{-1}\,\mathbf{v}_{S,\mathbf{X}} \tag{1.73}$$

The optimal estimator is, thus,

$$\hat{S}_{\text{LMSE}} = m_S + \mathbf{v}_{S\mathbf{X}}^{\intercal}\mathbf{V}_{\mathbf{X}}(\mathbf{X} - \mathbf{m}_{\mathbf{X}}) \tag{1.74}$$

The bias term w_0 compensates for differences between the means of the target variable and the observations. Therefore, when all the variables involved have zero mean, $w_0^* = 0$, and the estimator is purely linear in the observations.

1.4.2.3 Minimum squared mean error

The minimum MSE can be computed by replacing the optimal weights (1.62) into (1.59)

$$\begin{aligned} MSE^* &= \mathbb{E}\{S^2\} - 2\mathbf{r}_{S\mathbf{Z}}^{\intercal}\mathbf{R}_{\mathbf{Z}}^{-1}\mathbf{r}_{S\mathbf{Z}} + (\mathbf{R}_{\mathbf{Z}}^{-1}\mathbf{r}_{S\mathbf{Z}})^{\intercal}\mathbf{R}_{\mathbf{Z}}\mathbf{R}_{\mathbf{Z}}^{-1}\mathbf{r}_{S\mathbf{Z}} \\ &= \mathbb{E}\{S^2\} - \mathbf{r}_{S\mathbf{Z}}^{\intercal}\mathbf{R}_{\mathbf{Z}}^{-1}\mathbf{r}_{S\mathbf{Z}} \end{aligned} \tag{1.75}$$

For the linear estimator (1.74)

$$MSE^* = v_S - \mathbf{w}^{*T}\mathbf{v}_{S\mathbf{X}} \tag{1.76}$$

1.5 Estimation with Gaussian distributions

In this section, we delve into the estimation of random variables within the context where the combined distribution of all involved variables (the target variable along with the observational variables) is a multidimensional Gaussian. This scenario is particularly significant due to the prevalent occurrence of these distributions in signal processing, communications and various other fields.

When the joint distribution $p_{S,\mathbf{X}}(s,\mathbf{x})$ is Gaussian, all marginal and conditional distributions retain a Gaussian form. In particular, the posterior distribution, $p_{S|\mathbf{X}}(s|\mathbf{x})$ is Gaussian. Since the mean, median, and mode of the Gaussian distribution align, $\hat{s}_{\text{MSE}} = \hat{s}_{\text{MAD}} = \hat{s}_{\text{MAP}}$. Thus, our discussion in this section will primarily concentrate on deriving the estimator that minimizes the MSE.

Besides, we will demonstrate that the minimum MSE estimator and, consequently, the MAP and MAD estimators are linear, which will allow us to use the results shown in the previous section for minimum MSE estimation.

1.5.1 One dimensional case

We will consider as a starting point a case with one-dimensional random variables with zero means, in which the joint distribution of X and S has the following form:

$$p_{S,X}(s,x) \sim G\left(\begin{bmatrix} s \\ x \end{bmatrix}, \begin{bmatrix} v_X & \rho \\ \rho & v_S \end{bmatrix}\right) \tag{1.77}$$

where ρ is the covariance between the two random variables.

From this joint distribution, we can obtain any other distribution involving the variables S and X; specifically, the posterior distribution can be obtained as:

$$p_{S|X}(s|x) = \frac{p_{S,X}(s,x)}{p_X(x)} = \frac{\frac{1}{2\pi\sqrt{v_X v_S - \rho^2}} \exp\left[-\frac{1}{2(v_X v_S - \rho^2)} \begin{bmatrix} s \\ x \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} v_X & -\rho \\ -\rho & v_S \end{bmatrix} \begin{bmatrix} s \\ x \end{bmatrix}\right]}{\frac{1}{\sqrt{2\pi v_X}} \exp\left[-\frac{x^2}{2v_X}\right]} \tag{1.78}$$

where it has been necessary to calculate the inverse of the covariance matrix of S and X.

Noting that, as a function of s, $p_{S|X}(s|x)$ differs from $p_{S,X}(s,x)$ in the scale factor $p_X(x)$, which does not depend on s, $p_{S|X}(s|x)$ should be a Gaussian pdf too. Therefore, it must be expressed as

$$p_{S|X}(s|x) = \frac{1}{\sqrt{2\pi v_{S|X}}} \exp\left[-\frac{(s - m_{S|X})^2}{2v_{S|X}}\right] \tag{1.79}$$

where $m_{S|X}$ and $v_{S|X}$ are the posterior mean and variance, respectively, to be determined.

Joining (1.78) and (1.79), we can write

$$\frac{2\pi\sqrt{v_X v_S-\rho^2}}{\sqrt{2\pi v_{S|X}}\sqrt{2\pi v_X}}\exp\left[-\frac{(s-m_{S|X})^2}{2v_{S|X}}+\frac{1}{2(v_X v_S-\rho^2)}\begin{bmatrix}s\\x\end{bmatrix}^T\begin{bmatrix}v_X & -\rho\\-\rho & v_S\end{bmatrix}\begin{bmatrix}s\\x\end{bmatrix}-\frac{x^2}{2v_X}\right]=1 \tag{1.80}$$

which can be simplified to

$$\frac{\sqrt{v_X v_S-\rho^2}}{\sqrt{v_{S|X}v_X}}\exp\left[-\frac{s^2-2m_{S|X}s+m_{S|X}^2}{2v_{S|X}}+\frac{v_X s^2-2\rho xs+v_S x^2}{2(v_X v_S-\rho^2)}-\frac{x^2}{2v_X}\right]=1 \tag{1.81}$$

Note that the equation above must be satisfied for any $s\in\mathbb{R}$. Since the right-hand side is constant, and the exponent on the left-hand side is a polynomial function of s, the coefficients multiplying s^2 and s must be zero. Thus

$$\frac{m_{S|X}}{v_{S|X}}=\frac{\rho x}{v_X v_S-\rho^2} \tag{1.82}$$

$$\frac{1}{v_{S|X}}=\frac{v_X}{v_X v_S-\rho^2} \tag{1.83}$$

From (1.83) we get the posterior variance

$$v_{S|X}=v_S-\frac{\rho^2}{v_X} \tag{1.84}$$

and, replacing (1.84) into (1.82) we get the posterior mean, which is the minimum MSE estimate

$$\hat{s}_{\text{MSE}}=\hat{s}_{\text{MAD}}=\hat{s}_{\text{MAP}}=m_{S|X}=\frac{\rho}{v_X}x \tag{1.85}$$

Note that the estimator is a linear function of the observation.

Exercise 1.1 Generalize the above result for the case where the variables S and X have (non-zero) means m_S and m_X, respectively. Demonstrate that in such a case, the estimator is

$$\hat{s}_{\text{MSE}}=m_S+\frac{\rho}{v_X}(x-m_X) \tag{1.86}$$

Example 1.9 (Estimation of a Gaussian signal contaminated by Gaussian noise)

In this example, we will consider the case in which the observation is the sum of the target variable and an independent noise component: $X=S+R$. Both the target and the noise are zero-mean Gaussian random variables with variances v_S and v_R, respectively.

Figure (1.6) represents the situation described for a case with $v_S<v_R$.

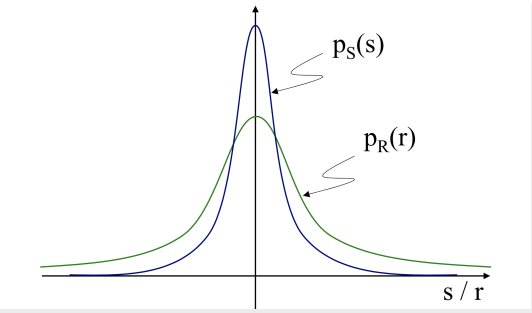


Fig. 1.6 Estimation of Gaussian random variable S contaminated by Gaussian noise R.

According to (1.85), for the resolution of the problem, we must find the variance of X and the covariance between S and X (ρ). The variance v_X is obtained simply as the sum of v_S and v_R because both are independent variables. For the covariance calculation,

$$\rho = \mathbb{E}\{(X - m_X)(S - m_S)\} = \mathbb{E}\{X\,S\} = \mathbb{E}\{(S+R)S\} = \mathbb{E}\{S^2\} + \mathbb{E}\{S\,R\} = v_S \tag{1.87}$$

where independence of S and R has been used, and the fact that all variables (including X) have zero means.

Replacing these results in (1.85) we get

$$\hat{s}_{\text{MSE}} = \frac{v_S}{v_S + v_R} x \tag{1.88}$$

This result can be interpreted quite intuitively: when the variance of the noise is much lower than that of the signal (high Signal to Noise Ratio (SNR), $v_S \gg v_R$) we get $\hat{s}_{\text{MSE}} \to x$, which makes sense since the effect of the noise component, in this case, is not very significant; on the contrary, when the SNR is very small ($v_S \ll v_R$), the observation barely provides information about the S value in each experiment, so the estimator keeps the mean value of the signal component, $\hat{s}_{\text{MSE}} \to 0$.

1.5.2 Case with multidimensional variables

In a general multidimensional case, $\mathbf{S}$ and $\mathbf{X}$ can be random vectors of dimensions N and M, respectively, with joint Gaussian distribution.

$$p_{\mathbf{S},\mathbf{X}}(\mathbf{s},\mathbf{x}) \sim G\left(\begin{bmatrix} \mathbf{m_S} \\ \mathbf{m_X} \end{bmatrix}, \begin{bmatrix} \mathbf{V_S} & \mathbf{V_{SX}} \\ \mathbf{V_{SX}}^T & \mathbf{V_X} \end{bmatrix} \right) \tag{1.89}$$

being $\mathbf{m_S}$ and $\mathbf{m_X}$ the means of $\mathbf{S}$ and $\mathbf{X}$, respectively, $\mathbf{V_S}$ and $\mathbf{V_X}$ the covariance matrix of $\mathbf{S}$ and $\mathbf{X}$, respectively, and $\mathbf{V_{SX}}$ the matrix of crossed covariances of $\mathbf{S}$ and $\mathbf{X}$, that is,

$$\mathbf{V_S} = \mathbb{E}\{(\mathbf{S} - \mathbf{m_S})(\mathbf{S} - \mathbf{m_S})^\intercal\} \tag{1.90}$$

$$\mathbf{V_X} = \mathbb{E}\{(\mathbf{X} - \mathbf{m_X})(\mathbf{X} - \mathbf{m_X})^\intercal\} \tag{1.91}$$

$$\mathbf{V_{SX}} = \mathbb{E}\{(\mathbf{S} - \mathbf{m_S})(\mathbf{X} - \mathbf{m_X})^\intercal\} \tag{1.92}$$

The calculation of the posterior distribution of $\mathbf{S}$ given $\mathbf{X}$ is more complex than in the one-dimensional case, but it follows a similar procedure, which we will omit here. It can be shown that the posterior distribution is Gaussian with mean

$$\mathbf{m}_{\mathbf{S}|\mathbf{X}} = \mathbf{m}_{\mathbf{S}} + \mathbf{V}_{\mathbf{SX}}\mathbf{V}_{\mathbf{X}}{}^{-1}(\mathbf{x} - \mathbf{m}_{\mathbf{X}}) \tag{1.93}$$

and covariance

$$\mathbf{V}_{\mathbf{S}|\mathbf{X}} = \mathbf{V}_{\mathbf{S}} - \mathbf{V}_{\mathbf{SX}}\mathbf{V}_{\mathbf{X}}{}^{-1}\mathbf{V}_{\mathbf{SX}}^{T} \tag{1.94}$$

Since the minimum MSE estimator of $\mathbf{S}$ given $\mathbf{X}$ is precisely the posterior mean, we can write

$$\hat{\mathbf{s}}_{\text{MSE}} = \hat{\mathbf{s}}_{\text{MAD}} = \hat{\mathbf{s}}_{\text{MAP}} = \mathbf{m}_{\mathbf{S}|\mathbf{X}} = \mathbf{m}_{\mathbf{S}} + \mathbf{V}_{\mathbf{SX}}\mathbf{V}_{\mathbf{X}}{}^{-1}(\mathbf{x} - \mathbf{m}_{\mathbf{X}}) \tag{1.95}$$

1.5.3 Linear estimation and Gaussian estimation

Note that, if the target variable is scalar, (1.95) is identical to (1.74). This is not coincidental: if the minimum MSE estimator in the Gaussian case is linear, it must be equal to the best linear estimator, which is given by (1.95).

1.6 ML estimation of probability distributions parameters

Sometimes we may be interested in estimating the parameters of a probability distribution, such as the mean or variance of a Gaussian distribution, the decay parameter that characterizes an exponential distribution, or values a and b delimiting the interval in which a uniform distribution is defined.

In these cases, the prior distribution of these variables is not usually known, even more, in many cases, these parameters are said to be deterministic and are not treated as random parameters. However, if a set of observations generated from these distributions is available, we can obtain the likelihood of these variables and estimate their values with maximum likelihood criteria.

Example 1.10 (ML estimate of the mean and variance of a one-dimensional Gaussian distribution)

The weight of individuals in a family of molluscs is known to obey a Gaussian distribution, but the mean and variance are unknown. The weight of l individuals taken independently, $\{X^{(k)}\}_{k=1}^{l}$, is available.

The likelihood of the mean and the variance for a single observation x, in this case, is given by:

$$p_X(x) = p_{X|m,v}(x|m,v) = \frac{1}{\sqrt{2\pi v}} \exp\left[-\frac{(x-m)^2}{2v}\right] \tag{1.96}$$

Since we must construct the estimator based on the joint observation of l observations, we will need to calculate the joint distribution of all of them which, being independent observations, is obtained as the product of individual observations:

$$\begin{aligned} p_{\{X^{(k)}\}|m,v}(\{x^{(k)}\}|m,v) &= \prod_{k=1}^{l} p_{X|m,v}(x^{(k)}|m,v) \\ &= \frac{1}{(2\pi v)^{l/2}} \prod_{k=1}^{l} \exp\left[-\frac{(x^{(k)}-m)^2}{2v}\right] \end{aligned} \tag{1.97}$$

The ML estimators of m and v will be the values maximizing (1.97). The analytical form of this function suggests the use of the logarithm to simplify the maximization:

$$L = \ln\left[p_{\{X^{(k)}\}|m,v}(\{x^{(k)}\}|m,v)\right] = -\frac{l}{2}\ln(2\pi v) - \frac{1}{2v}\sum_{k=1}^{l}(x^{(k)}-m)^2 \tag{1.98}$$

Differentiating (1.98) with respect to m and v, we get the system of equations to solve

$$\begin{aligned} \left.\frac{d\,L}{d\,m}\right|_{\substack{m=\hat{m}_{\mathrm{ML}} \\ v=\hat{v}_{\mathrm{ML}}}} &= -\frac{1}{v}\sum_{k=1}^{l}(x^{(k)}-m)\Bigg|_{\substack{m=\hat{m}_{\mathrm{ML}} \\ v=\hat{v}_{\mathrm{ML}}}} = 0 \\ \left.\frac{d\,L}{d\,v}\right|_{\substack{m=\hat{m}_{\mathrm{ML}} \\ v=\hat{v}_{\mathrm{ML}}}} &= -\frac{l}{2v} + \frac{1}{2v^2}\sum_{k=1}^{l}(x^{(k)}-m)^2\Bigg|_{\substack{m=\hat{m}_{\mathrm{ML}} \\ v=\hat{v}_{\mathrm{ML}}}} = 0 \end{aligned} \tag{1.99}$$

Solving for m the first equation we get

$$\hat{m}_{\mathrm{ML}} = \frac{1}{l}\sum_{k=1}^{l} x^{(k)} \tag{1.100}$$

which is the sample average of the observations. On the other hand, we can solve the second equation for v to get

$$\hat{v}_{\mathrm{ML}} = \frac{1}{l}\sum_{k=1}^{l} (x^{(k)} - \hat{m}_{\mathrm{ML}})^2 \tag{1.101}$$

which is the sample variance. Note that, if instead of applying the estimation function (of m or v) on some specific observations we did it on generic values $\{X^{(k)}\}$, the estimators could be treated as random variables, i.e.,

$$\hat{M}_{\mathrm{ML}} = \frac{1}{l}\sum_{k=1}^{l} X^{(k)} \tag{1.102}$$

$$\hat{V}_{\mathrm{ML}} = \frac{1}{l}\sum_{k=1}^{l} [X^{(k)} - \hat{M}_{\mathrm{ML}}]^2 \tag{1.103}$$

1.7 Problems

1.1 The posterior distribution of S given X is

$$p_{S|X}(s|x) = x^2 \exp(-x^2 s), \qquad s \geq 0$$

Compute estimators $\hat{S}_{\text{MMSE}}$, $\hat{S}_{\text{MAD}}$ y $\hat{S}_{\text{MAP}}$.

1.2 Consider an estimation problem given by the following posterior distribution:

$$p_{S|X}(s|x) = x \exp(-xs), \;\; s > 0 \tag{1.104}$$

Compute estimators $\hat{S}_{\text{MMSE}}$, $\hat{S}_{\text{MAD}}$ y $\hat{S}_{\text{MAP}}$.

1.3 A r.v. S must be estimated from the observation of another r.v. X by means of a linear mean square error estimator given by:

$$\hat{S}_{\text{LMSE}} = w_0 + w_1 X$$

Knowing that $\mathbb{E}\{X\} = 1$, $\mathbb{E}\{S\} = 0$, $\mathbb{E}\{X^2\} = 2$, $\mathbb{E}\{S^2\} = 1$ y $\mathbb{E}\{SX\} = 1/2$, compute:

a) The values for w_0 y w_1.
b) The mean square error of the estimator, $\mathbb{E}\left\{\left(S - \hat{S}_{\text{LMSE}}\right)^2\right\}$.

1.4 (Linear estimation of minimum mean squared error) We want to construct a linear estimator of minimum mean squared error that will allow us to estimate the random variable S from the random variables X_1 and X_2. Knowing that

$$\begin{array}{lll} \mathbb{E}\{S\} = 1/2 & \mathbb{E}\{X_1\} = 1 & \mathbb{E}\{X_2\} = 0 \\ \mathbb{E}\{S^2\} = 4 & \mathbb{E}\{X_1^2\} = 3/2 & \mathbb{E}\{X_2^2\} = 2 \\ \mathbb{E}\{SX_1\} = 1 & \mathbb{E}\{SX_2\} = 2 & \mathbb{E}\{X_1 X_2\} = 1/2 \end{array}$$

get the weights from the desired estimator and calculate its squared mean error. Calculate the estimated value for the observation vector: $[X_1, X_2] = [3, 1]$.

1.5 Let X and S be two random variables with joint pdf

$$p_{X,S}(x,s) \begin{cases} 2 \; 0 < x < 1, 0 < s < x \\ 0 \text{ resto} \end{cases}$$

a) Compute the minimum mean square error estimate of S given X, $\hat{S}_{\text{MMSE}}$.
b) Compute the risk of estimator $\hat{S}_{\text{MMSE}}$.

1.6 A digitized image of dimensions $8x8$ is available, whose luminance values are statistically independent and evenly distributed between 0 (white) and 1 (black); the image has been modified by applying a transformation of the form $Y = X^r$ on each pixel; $r > 0$, where X is the r.v. associated with the pixels of the original image and Y is associated with the transformed image. Obtain the expression that allows estimating r by maximum likelihood given the 64 pixel values of the transformed image $\{y^{(k)}\}_{k=1}^{64}$, without knowing the original image.

1.7 For the design of a communication system, it is desired to estimate the signal attenuation between the transmitter and the receiver, as well as the noise power introduced by the channel when this noise is Gaussian of zero mean and independent of the transmitted signal. For this, the transmitter sends a signal with a constant amplitude of 1 and the receiver collects a set of K observations available at its input.

a) Estimate the channel attenuation, α, and the noise variance, v_r, by maximum likelihood, when the available observations on the receiver are

$$\{0.55, 0.68, 0.27, 0.58, 0.53, 0.37, 0.45, 0.53, 0.86, 0.78\}.$$

b) If the system is to be used for the transmission of digital signals with unipolar coding (a A signal level is used to transmit a bit 1 and the signal level is maintained at 0 for the transmission of bit 0), considering equiprobability between symbols, indicate the minimum level of signal that should be used in the coding, $A_{\min}$, to guarantee a SNR level in the receiver of 3 dB.

1.8 Company *Like2Call* offers hosting services for call centers. In order to dimension the staff of operators the company is designing a statistical model to characterize the activity in the hosted call centers. One of the components of such a model relies on the well-known fact that the times between incoming calls follow an exponential distribution

$$p_{X|S}(x|s) = s\ \exp(-s\,x), \qquad x > 0$$

where random variable X represents the time before a new call arrives, and S is the parameter of such distribution, that depends on the time of the day and each particular call-center service (e.g., attention to the clients of an insurance company, customers of an online bank, etc).

For random variable S, the following *a priori* model is assumed:

$$p_S(s) = \exp(-s), \qquad s > 0.$$

With this information, we would like to design an estimator of S that is based on the first K incoming calls for each implemented service and time interval, i.e., the observation vector is given by $\mathbf{x} = \left[x^{(0)}, x^{(1)}, \cdots, x^{(K-1)}\right]$, where all observations in the vector are assumed i.i.d.

a) Obtain the maximum likelihood estimator or S based on the observation vector $\mathbf{X}$ and verify that it depends just on the sum of all observations, $z = \sum_{k=0}^{K-1} x^{(k)}$.
b) Calculate the posterior distribution of S given $\mathbf{X}$, $p_{S|\mathbf{X}}(s|\mathbf{x})$.
c) Obtain the maximum *a posteriori* estimator of S given $\mathbf{X}$, $\hat{s}_{\text{MAP}}$.
d) Obtain the minimum mean square error estimator of S given $\mathbf{X}$, $\hat{s}_{\text{MSE}}$.
e) Calculate the mean square error given $\mathbf{X}$ of a generic estimator $\hat{S}$, and particularize the result for estimators of the following analytical shape $\hat{s}_c = \frac{c}{z+1}$.
f) Find expressions for the following probability density functions: $p_{Z|S}(z|s)$, $p_{Z,S}(z,s)$, and $p_Z(z)$.
g) Calculate the mean square error of a generic estimator $\hat{s}_c = \frac{c}{z+1}$. Study how the result changes with c and K.

You can use the following results:

i.
$$\int_0^\infty x^N \exp(-x)dx = N!$$

ii. If $f(x) = a\, exp(-a\, x),\ x > 0$ then

$$\underbrace{f(x) * f(x) * \cdots * f(x)}_{N \text{ times}} = \frac{a^N\, x^{N-1}}{(N-1)!} exp(-a\, x),\ x > 0$$

iii. For K an integer
$$\int_0^\infty \frac{K\, x^{K-1}}{(x+1)^{K+3}} dx = \frac{2}{(K+2)(K+1)}$$

Solution 1.1

a)

$$\begin{aligned}
&p_{\mathbf{X}|S}(\mathbf{x}|s) = s^K\ \exp(-s\, z), \qquad z > 0\\
&\ln p_{\mathbf{X}|S}(\mathbf{x}|s) = K \ln s - s\, z\\
&\frac{d}{ds} \ln p_{\mathbf{X}|S}(\mathbf{x}|s) = \frac{K}{s} - z\\
&\hat{s}_{\mathrm{ML}} = \frac{K}{z}
\end{aligned}$$

b)

$$\begin{aligned}
&p_{\mathbf{X},S}(\mathbf{x}, s) = p_{\mathbf{X}|S}(\mathbf{x}|s) p_S(s) = s^K\ \exp[-s(z+1)]\\
&\text{(note the expression above is not the joint pdf of } Z \text{ and } S)\\
&p_{\mathbf{X}}(\mathbf{x}) = \int p_{\mathbf{X},S}(\mathbf{x}, s)\, ds = \int_0^\infty s^K\ \exp[-s(z+1)]\, ds
\end{aligned}$$

With the change of variable $s' = s(z+1)$ the previous integral can be simplified using expression (i), and we get

$$p_{S|\mathbf{X}}(s|\mathbf{x}) = \frac{p_{\mathbf{X},S}(\mathbf{x}, s)}{p_{\mathbf{X}}(\mathbf{x})} = \frac{(z+1)^{K+1}\, p_{\mathbf{X},S}(\mathbf{x}, s)}{K!} = \frac{s^K (z+1)^{K+1}\ \exp[-s(z+1)]}{K!}$$

c)

$$\begin{aligned}
&\hat{s}_{\mathrm{MAP}} = \arg\max_s p_{S|\mathbf{X}}(s|\mathbf{x}) = \arg\max_s p_{\mathbf{X},S}(\mathbf{x}, s)\\
&\ln p_{\mathbf{X},S}(\mathbf{x}, s) = K \ln s - s\ (z+1)\\
&\frac{d}{ds} \ln p_{\mathbf{X},S}(\mathbf{x}, s) = \frac{K}{s} - (z+1)\\
&\hat{s}_{\mathrm{MAP}} = \frac{K}{z+1}
\end{aligned}$$

d)

$$\hat{s}_{\text{MSE}} = \mathbb{E}\{S|\mathbf{x}\} = \int s\, p_{S|\mathbf{X}}(s|\mathbf{x})\, ds = \frac{(z+1)^{K+1}}{K!} \int_0^\infty s^{K+1}\, \exp[-s(z+1)]\, ds$$

Replacing again $s' = s(z+1)$ and using expression (i), we get

$$\hat{s}_{\text{MSE}} = \frac{K+1}{z+1}$$

e) The calculation is somehow tedious, but can be summarized as follows:

$$\begin{aligned}
\mathbb{E}\left\{(S-\hat{s})^2|X\right\} &= \int_0^\infty (s-\hat{s})^2\, p_{S|X}(s|x) ds \\
&= \frac{(z+1)^{K+1}}{K!}\left[\frac{(K+2)!}{(z+1)^{K+3}} + \hat{s}^2\frac{K!}{(z+1)^{K+1}} - 2\hat{s}\frac{(K+1)!}{(z+1)^{K+2}}\right] \\
&= \frac{(K+2)(K+1)+c^2-2c(K+1)}{(z+1)^2}
\end{aligned}$$

For the MAP and MSE estimators, the expressions are substantially simplified:

$$\begin{aligned}
\mathbb{E}\left\{(S-\hat{s}_{MAP})^2|z\right\} &= \frac{K+2}{(z+1)^2} \\
E\left\{(S-\hat{s}_{MSE})^2|z\right\} &= \frac{K+1}{(z+1)^2}
\end{aligned}$$

f) Using the fact that Z is the sum of K i.i.d. variables (given S):

$$p_{Z|S}(z|s) = \underbrace{[s\, \exp(-s\, z)] * \cdots * [s\, \exp(-s\, z)]}_{K \text{ times}} = \frac{s^K\, z^{K-1}}{(K-1)!} \exp(-s\, z), \qquad z > 0$$

The joint pdf of Z and S can now be obtained as

$$p_{Z,S}(z,s) = p_{Z|S}(z|s)p_S(s) = \frac{s^K\, z^{K-1}}{(K-1)!} \exp[-s\,(z+1)], \qquad s,z > 0$$

Finally, integrating s out, we have

$$p_Z(z) = \int p_{Z,S}(z,s)ds = \frac{z^{K-1}}{(K-1)!}\int_0^\infty s^K \exp[-s\,(z+1)] = \frac{K\, z^{K-1}}{(z+1)^{K+1}},\ \ z > 0$$

g)

$$\mathbb{E}\left\{(S-\hat{S}_c)^2\right\} = \int \mathbb{E}\left\{(S-\hat{s}_c)^2|z\right\}\, p_Z(z)dz$$

Using the results from the previous two sections we can obtain an expression that depends on the value of an integral over z:

$$\mathbb{E}\left\{(S-\hat{S}_c)^2\right\} = \left[(K+2)(K+1)+c^2-2c(K+1)\right]\int_0^\infty \frac{K\, z^{K-1}}{(z+1)^{K+3}} dz$$

The value of the integral is given in (iii). Simplifying also for the MAP and MSE estimators:

$$\mathbb{E}\left\{(S-\hat{S}_{MAP})^2\right\} = \frac{2}{K+1}$$
$$E\left\{(S-\hat{S}_{MSE})^2\right\} = \frac{2}{K+2}$$

Chapter 2
Linear Filtering

2.1 Introduction

A common challenge in estimation involves determining the coefficients of a linear filter with M coefficients based solely on the observation of its inputs and outputs. This task, along with related ones, falls under the generic term 'linear filtering.' In this chapter, we will demonstrate how the techniques described in Chapter **??** can be applied to design ML, MAP, MAD and MMSE estimators for the coefficients of the aforementioned filter. This will also include the estimation of future filter outputs, provided that the corresponding inputs are known.

2.2 The filtering problem

Assume that a finite impulse response filter (FIR) $s[n]$, with $s[n] = 0$, for n other than $0, 1, \ldots, M-1$ is used to filter a signal[1] $u[n]$. The result is then added to a certain Gaussian noise $\varepsilon[n]$, which is an IID zero-mean stochastic process with variance σ_ε^2 and independent of $s[n]$, giving rise to an observation $x[n]$ (see Figure 2.1). That is, the corresponding entries are

$$\begin{aligned} x[n] &= u[n] * s[n] + \varepsilon[n] \\ &= u[n]s[0] + u[n-1]s[1] + \ldots + u[n-M+1]s[M-1] + \varepsilon[n]. \end{aligned} \tag{2.1}$$

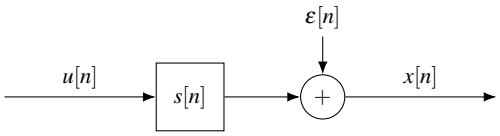


Fig. 2.1 The signal model used in this chapter.

For simplicity, we will assume that the input signal starts at $n = 0$, that is, $u[n] = 0$ for $n < 0$. Also, we will assume that it is known; thus, a satistical model for $u[n]$ is unnecessary for our analysis).

The filtering problem consists in estimating the filter $s[n]$ from the input signal and a finite set of N samples from the output, $x[0], \ldots, x[N-1]$. Also, we will tackle the output prediction problem: predicting unobserved values of the output for a given input signal $u[n]$.

Our goal is to solve the filtering problem using the tools from estimation theory developed in Chapter **??**. To do so, we will first represent the target variables (the filter coefficients) and the observations (the output signal) in vector form, and the relation between them using a vector equation. Thus, we will join the nonzero coefficients in an M-dimensional vector

[1] Although the notation $u[n]$ is frequently used to refer to the step function, in this chapter it is used to denote an arbitrary input signal.

$$\mathbf{s} = \begin{bmatrix} s[0] \\ s[1] \\ \vdots \\ s[M-1] \end{bmatrix}_{M\times 1} . \tag{2.2}$$

Also, we will represent any M-length window of consecutive input values in vector form as

$$\mathbf{u}[n] = \begin{bmatrix} u[n] \\ u[n-1] \\ \vdots \\ u[n-M+1] \end{bmatrix}_{M\times 1} , \tag{2.3}$$

we can write

$$x[n] = \mathbf{u}[n]^\top \mathbf{s} + \varepsilon[n]. \tag{2.4}$$

Note that, for any n, $\mathbf{u}[n]$ contains only the input values that are relevant to compute $x[n]$.

Taking $n = 0, 1, \ldots, N-1$ in Eq. (2.4), we get a system of N linear equations relating the filter coefficients and the observations. We can join all of them into a single matrix equation by defining the observation vector

$$\mathbf{x} = \begin{bmatrix} x[0] \\ x[1] \\ \vdots \\ x[N-1] \end{bmatrix}_{N\times 1} , \tag{2.5}$$

the input matrix

$$\begin{aligned} \mathbf{U} &= [\mathbf{u}[0] \;\; \mathbf{u}[1] \;\; \ldots \;\; \mathbf{u}[M-1] \;\; \ldots \;\; \mathbf{u}[N-1]] \\ &= \begin{bmatrix} u[0] & u[1] & \ldots & u[M-1] & \ldots & u[N-1] \\ 0 & u[0] & \ldots & u[M-2] & \ldots & u[N-2] \\ \vdots & \vdots & \ddots & \vdots & \ldots & \vdots \\ 0 & 0 & \ldots & u[0] & \ldots & u[N-M] \end{bmatrix}_{M\times N} , \end{aligned} \tag{2.6}$$

and the noise vector

$$\boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon[0] \\ \varepsilon[1] \\ \vdots \\ \varepsilon[N-1] \end{bmatrix}_{N\times 1} , \tag{2.7}$$

Using (2.5), (2.6) (2.7), we can write the signal model (2.4) as

$$\boxed{\mathbf{x} = \mathbf{U}^\top \mathbf{s} + \boldsymbol{\varepsilon}} \tag{2.8}$$

The filtering problem reduces to the problem of estimating $\mathbf{s}$ given $\mathbf{x}$ knowing (2.8) and the noise statistics.

2.3 ML solution

Eq. (2.4) shows that, given $\mathbf{s}$, $x[n]$ is Gaussian with mean

$$\mathbb{E}\{x[n] \mid \mathbf{s}\} = \mathbf{u}[n]^\top \mathbf{s} + \mathbb{E}\{\varepsilon[n]\} = \mathbf{u}[n]^\top \mathbf{s} \tag{2.9}$$

and variance

$$\mathbb{E}\{(x[n] - \mathbf{u}[n]^\top \mathbf{s})^2 \mid \mathbf{s}\} = \sigma_\varepsilon^2 \tag{2.10}$$

that is, the likelihood function is

$$p(x[n] \mid \mathbf{s}) = \mathcal{N}(x[n] \mid \mathbf{u}[n]^\top \mathbf{s}, \sigma_\varepsilon^2), \tag{2.11}$$

where the notation $\mathcal{N}(x \mid \mu, v)$ is used to refer to the *normal* (Gaussian) pdf of a random variable with mean μ and variance v, evaluated at x.

Given $\mathbf{s}$, $x[n]$ depends on $\varepsilon[n]$ only, which is an IID process. Thus, all samples from $x[n]$ are independent given $\mathbf{s}$, that is,

$$p_{\mathbf{X}|\mathbf{S}}(\mathbf{x} \mid \mathbf{s}) = \prod_{n=0}^{N-1} \mathcal{N}(x[n] \mid \mathbf{u}[n]^\top \mathbf{s}, \sigma_\varepsilon^2) = \mathcal{N}(\mathbf{x} \mid \mathbf{U}^\top \mathbf{s}, \sigma_\varepsilon^2 \mathbf{I}). \tag{2.12}$$

The value of $\mathbf{s}$ that maximizes $p(\mathbf{x}|\mathbf{s})$ is

$$\begin{aligned} \hat{\mathbf{s}}_{\mathrm{ML}} &= \operatorname*{argmax}_{\mathbf{s}} p_{\mathbf{X}|\mathbf{S}}(\mathbf{x} \mid \mathbf{s}) = \operatorname*{argmax}_{\mathbf{s}} \log p_{\mathbf{X}|\mathbf{S}}(\mathbf{x} \mid \mathbf{s}) \\ &= \operatorname*{argmin}_{\mathbf{s}} \frac{1}{2}(\mathbf{x} - \mathbf{U}^\top \mathbf{s})^\top (\sigma_\varepsilon^2 \mathbf{I})^{-1} (\mathbf{x} - \mathbf{U}^\top \mathbf{s}) + \frac{1}{2} \log |\sigma_\varepsilon^2 \mathbf{I}| + \frac{N}{2} \log(2\pi) \\ &= \operatorname*{argmin}_{\mathbf{s}} ||\mathbf{x} - \mathbf{U}^\top \mathbf{s}||^2 \end{aligned} \tag{2.13}$$

This minimum can be easily obtained by taking the gradient with respect to $\mathbf{s}$, equalizing to zero and clearing, leading to

$$\hat{\mathbf{s}}_{\mathrm{ML}} = (\mathbf{U}\mathbf{U}^\top)^{-1} \mathbf{U}\mathbf{x}. \tag{2.14}$$

2.4 Bayesian Solution

To obtain a Bayesian estimator of $\mathbf{s}$ it is necessary to know its prior probability distribution, $p(\mathbf{s})$. A common choice for this distribution is

$$p_{\mathbf{S}}(\mathbf{s}) = \mathcal{N}(\mathbf{s}|\mathbf{0}, \sigma_s^2\mathbf{I}), \tag{2.15}$$

since it accommodates any set of real coefficients and assumes they have a zero mean and a dispersion determined by σ_s^2. It is also possible to set $\sigma_s^2 \to \infty$ to approach a uniform distribution. In any case, the use of this prior distribution enables the analytic derivation of the posterior distribution.

Given the likelihood, $p(\mathbf{x}|\mathbf{s})$, and the prior distribution $p(\mathbf{s})$, the posterior distribution $p(\mathbf{s}|\mathbf{x})$ can be obtained. While it is feasible to directly apply Bayes' theorem and simplify the expression as much as possible, this process can be quite tedious. Instead, we will achieve the result in two steps.

First we will determine the joint pdf of $\mathbf{s}$ and $\mathbf{x}$. A simple way to do this is to observe that

$$\begin{bmatrix}\mathbf{s}\\ \mathbf{x}\end{bmatrix} = \begin{bmatrix}\mathbf{I}\\ \mathbf{U}^\top\end{bmatrix}\mathbf{s} + \begin{bmatrix}\mathbf{0}\\ \boldsymbol{\varepsilon}\end{bmatrix} \tag{2.16}$$

that is, vector $[\mathbf{s}^\top\ \mathbf{x}^\top]^\top$ is a linear combination of two Gaussian random vectors and, thus, is jointly Gaussian:

$$p_{\mathbf{S},\mathbf{X}}(\mathbf{s},\mathbf{x}) = \mathcal{N}\left(\begin{bmatrix}\mathbf{s}\\ \mathbf{x}\end{bmatrix} \middle| \begin{bmatrix}\mathbf{m}_\mathbf{S}\\ \mathbf{m}_\mathbf{X}\end{bmatrix}, \begin{bmatrix}\mathbf{V}_\mathbf{S} & \mathbf{V}_{\mathbf{SX}}\\ \mathbf{V}_{\mathbf{SX}}^\top & \mathbf{V}_\mathbf{X}\end{bmatrix}\right) \tag{2.17}$$

where

$$\mathbf{m}_\mathbf{S} = \mathbb{E}\{\mathbf{s}\} = \mathbf{0} \tag{2.18}$$

$$\mathbf{m}_\mathbf{X} = \mathbb{E}\{\mathbf{x}\} = \mathbf{U}^\top\mathbb{E}\{\mathbf{s}\} + \mathbb{E}\{\boldsymbol{\varepsilon}\} = \mathbf{0} \tag{2.19}$$

$$\mathbf{V}_\mathbf{S} = \mathrm{Var}\{\mathbf{s}\} = \sigma_s^2\mathbf{I} \tag{2.20}$$

$$\mathbf{V}_{\mathbf{SX}} = \mathbb{E}\{\mathbf{s}\mathbf{x}^\top\} = \mathbb{E}\{\mathbf{s}\mathbf{s}^\top\}\mathbf{U} + \mathbb{E}\{\mathbf{s}\boldsymbol{\varepsilon}^\top\} = \sigma_s^2\mathbf{U} \tag{2.21}$$

$$\mathbf{V}_\mathbf{X} = \mathbb{E}\{\mathbf{x}\mathbf{x}^\top\} = \mathbf{U}^\top\mathbb{E}\{\mathbf{s}\mathbf{s}^\top\}\mathbf{U} + \mathbb{E}\{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\} = \sigma_s^2\mathbf{U}^\top\mathbf{U} + \sigma_\varepsilon^2\mathbf{I} \tag{2.22}$$

Therefore,

$$p_{\mathbf{S},\mathbf{X}}(\mathbf{s},\mathbf{x}) = \mathcal{N}\left(\begin{bmatrix}\mathbf{s}\\ \mathbf{x}\end{bmatrix} \middle| \begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix}, \begin{bmatrix}\sigma_s^2\mathbf{I} & \sigma_s^2\mathbf{U}\\ \sigma_s^2\mathbf{U}^\top & \sigma_s^2\mathbf{U}^\top\mathbf{U} + \sigma_\varepsilon^2\mathbf{I}\end{bmatrix}\right) \tag{2.23}$$

From previous chapter, we know that if $p_{\mathbf{S},\mathbf{X}}(\mathbf{s},\mathbf{x})$ is Gaussian, the conditional distribution $p_{\mathbf{S}|\mathbf{X}}(\mathbf{s} \mid \mathbf{x})$ is also Gaussian, with mean and covariance

$$\begin{aligned}\mathbf{m}_{\mathbf{S}|\mathbf{X}} &= \mathbf{m}_{\mathbf{S}} + \mathbf{V}_{\mathbf{SX}}\mathbf{V}_{\mathbf{X}}{}^{-1}(\mathbf{x}-\mathbf{m}_{\mathbf{X}}) \\ &= \mathbf{U}\left(\mathbf{U}^\top\mathbf{U} + \frac{\sigma_\varepsilon^2}{\sigma_s^2}\mathbf{I}\right)^{-1}\mathbf{x} \qquad (2.24)\\ \mathbf{V}_{\mathbf{S}|\mathbf{X}} &= \mathbf{V}_{\mathbf{S}} - \mathbf{V}_{\mathbf{SX}}\mathbf{V}_{\mathbf{X}}{}^{-1}\mathbf{V}_{\mathbf{SX}}^\top \\ &= \sigma_s^2\mathbf{I} - \sigma_s^2\mathbf{U}\left(\mathbf{U}^\top\mathbf{U} + \frac{\sigma_\varepsilon^2}{\sigma_s^2}\mathbf{I}\right)^{-1}\mathbf{U}^\top \qquad (2.25)\end{aligned}$$

The above expressions involve the computation of the inverse of an $N \times N$ matrix, which may be not feasible for large signal records (the computational cost is $\mathcal{O}(N^3)$. However, using the *matrix inversion lemma* (see the Appendix in Sec. 2.7), we can obtain the following alternative expressions:

$$\mathbf{m}_{\mathbf{S}|\mathbf{X}} = \mathbf{P}\mathbf{U}\mathbf{x} \qquad (2.26)$$

$$\mathbf{V}_{\mathbf{S}|\mathbf{X}} = \sigma_\varepsilon^2\mathbf{P} \qquad (2.27)$$

where

$$\mathbf{P} = (\mathbf{U}\mathbf{U}^\top + \tfrac{\sigma_\varepsilon^2}{\sigma_s^2}\mathbf{I})^{-1}. \qquad (2.28)$$

Eq. (2.28) involves the inversion of a matrix $M \times M$, which is usually much smaller than $N \times N$.

Using these expressions, the MMSE, MAP and MAD estimates of $\mathbf{s}$ are:

$$\hat{\mathbf{s}}_{\text{MSE}} = \hat{\mathbf{s}}_{\text{MAP}} = \hat{\mathbf{s}}_{\text{MAD}} = \mathbf{P}\mathbf{U}\mathbf{x} \qquad (2.29)$$

Also, note that taking $\sigma_\varepsilon^2 \to 0$ (negligible noise) and/or $\sigma_s^2 \to \infty$ (which can be interpreted as assuming an infinitely wide uniform prior) these Bayesian solutions become equivalent to the ML in (2.14).

2.4.1 Probabilistic prediction of the filter output

Once we have resolved several estimators of filter $\mathbf{s}$, we now begin to consider the problem of predicting a new output $x[k]$ at some time $k > N$. Continuing with the Bayesian perspective, we will obtain the posterior pdf of the target variable, $x[k]$, in light of the outputs already observed, $\mathbf{x}$. That is, we aim to calculate $p(x[k] \mid \mathbf{x})$.

First, it should be noted that $\mathbf{x}, x[k]$ and $\mathbf{s}$ are jointly Gaussian. This follows from Eq. (2.23), which can be extended to any arbitrary number of outputs, including $x[k]$. This

necessarily implies that $\mathbf{x}$ and $x[k]$ are jointly Gaussian (when marginalizing $\mathbf{s}$) and finally that $p(x[k]|\mathbf{x})$ must be Gaussian. Given that

$$x[k] = \mathbf{u}[k]^\top \mathbf{s} + \varepsilon[k] \tag{2.30}$$

is a linear transformation of $\mathbf{s}$ with independent white noise, we can easily compute the posterior mean and variance, respectively, as follows:

$$\mathbb{E}\{x[k] \mid \mathbf{x}\} = \mathbf{u}[k]^\top \mathbb{E}\{\mathbf{s} \mid \mathbf{x}\} + \mathbb{E}\{\varepsilon[k] \mid \mathbf{x}\} = \mathbf{u}[k]^\top \hat{\mathbf{s}}_{\text{MSE}} \tag{2.31}$$

$$\begin{aligned}
\text{Var}\{x[k] \mid \mathbf{x}\} &= \mathbb{E}\left\{ \left(\mathbf{u}[k]^\top (\mathbf{s} - \hat{\mathbf{s}}_{\text{MSE}}) + \varepsilon[k] \right)^2 \mid \mathbf{x} \right\} \\
&= \mathbf{u}[k]^\top \mathbb{E}\{ (\mathbf{s} - \hat{\mathbf{s}}_{\text{MSE}}) (\mathbf{s} - \hat{\mathbf{s}}_{\text{MSE}})^\top \mid \mathbf{x} \} \mathbf{u}[k] + \mathbb{E}\{\varepsilon[k]^2 \mid \mathbf{x}\} \\
&= \mathbf{u}[k]^\top \mathbf{V}_{\mathbf{S}|\mathbf{X}} \mathbf{u}[k] + \sigma_\varepsilon^2 \\
&= \sigma_\varepsilon^2 \mathbf{u}[k]^\top \mathbf{P} \mathbf{u}[k] + \sigma_\varepsilon^2
\end{aligned} \tag{2.32}$$

Therefore, the Bayesian prediction of the output is succinctly given by

$$\hat{x}_{\text{MSE}}[k] = \hat{x}_{\text{MAP}}[k] = \hat{x}_{\text{MAD}}[k] = \mathbf{u}[k]^\top \hat{\mathbf{s}}_{\text{MSE}}. \tag{2.33}$$

Eq. (2.33) shows that the optimal prediction of the output at any time $k > N$ can be computed by passing the input signal $u[n]$ though the filter with the coefficients in the Bayesian estimation of $\mathbf{s}$ (see Fig. 2.2).

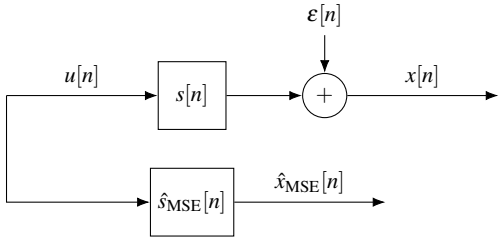


Fig. 2.2 The best prediction of future values of the output (knowing the input) can be computed by passing the input signal through a filter with the coefficients of the Bayesian estimation of $\mathbf{s}$.

2.5 Online calculus

It is possible to obtain the above solutions online, that is, as new input-output pairs are obtained. While complete calculations could be repeated each time a new sample arrives, there are often more efficient ways to do this.

Note that estimating $\mathbf{s}$ using Eqs. (2.14) or (2.29) requires inverting an $M \times M$ matrix. This has a cost $\mathscr{O}(M^3)$, that is, if we double the size of the filter, M we multiply its computational cost by eight. Suppose now that you want to estimate $\mathbf{s}$ as new input-output pairs

are received, that is, we are given first $\{u[0],x[0]\}$, then $\{u[1],x[1]\}$ and so on. In this case, we could reuse the results of the previous estimate to calculate the new updated estimate of $\mathbf{s}$, thus reducing the cost $\mathscr{O}(M^3)$ that would have a *naive* method that simply recalculates everything again every time a sample arrives.

2.5.1 Bayesian solution

$\hat{\mathbf{s}}_{\text{MSE}}$ can be obtained exactly as more samples are available (i.e. as N increases) without redoing all calculations, by reusing the previous solution. To do this, it is defined

$$\mathbf{P}_N = (\mathbf{U}\mathbf{U}^\top + \frac{\sigma_\varepsilon^2}{\sigma_s^2}\mathbf{I})^{-1}, \tag{2.34}$$

$$\mathbf{r}_N = \mathbf{U}\mathbf{x} \tag{2.35}$$

and the following recursive calculation is used (the first equation corresponds to the direct application of the matrix inversion lemma to the $\mathbf{P}$ update):

$$\begin{aligned}
\mathbf{P}_{N+1} &= \mathbf{P}_N - \frac{\mathbf{P}_N\mathbf{u}[N+1]\mathbf{u}[N+1]^\top\mathbf{P}_N}{1+\mathbf{u}[N+1]^\top\mathbf{P}_N\mathbf{u}[N+1]} \\
\mathbf{r}_{N+1} &= \mathbf{r}_N + \mathbf{u}[N+1]x[N+1] \\
\mathbf{s}_{N+1} &= \mathbf{P}_{N+1}\mathbf{r}_{N+1},
\end{aligned}$$

which only has a cost $\mathscr{O}(M^2)$ per step (as opposed to applying the complete original equation at each step, which would cost $\mathscr{O}(M^3)$). This algorithm is called *recursive least squares* (RLS).

2.5.2 ML solution

An online approximation to $\hat{\mathbf{s}}_{\text{ML}}$ with computational cost $\mathscr{O}(M)$ can be obtained just by noting that

$$\hat{\mathbf{s}}_{\text{ML}} = \underset{\mathbf{s}}{\operatorname{argmax}}\, p(\mathbf{x}|\mathbf{s}) = \underset{\mathbf{s}}{\operatorname{argmin}}\, ||\mathbf{x}-\mathbf{U}^\top\mathbf{s}||^2 \tag{2.36}$$

and then use stochastic gradient to minimize $||\mathbf{x}-\mathbf{U}^\top\mathbf{s}||^2$.

Notice that

$$||\mathbf{x}-\mathbf{U}^\top\mathbf{s}||^2 = \sum_{n=0}^{N-1}(x[n]-\mathbf{u}[n]^\top\mathbf{s})^2, \tag{2.37}$$

so a gradient descent method would calculate the gradient of that expression and iteratively shift the estimate of the minimum in the opposite direction of the gradient in each step. A descent by stochastic gradient performs the same operation, but considering only one of the additions of the mentioned sum in each step. So, the updating of coefficients that must be iterated to perform the minimization is in this case

$$\hat{\mathbf{s}}_{n+1} = \hat{\mathbf{s}}_n + \mu \left(x[n] - \mathbf{u}[n]^\top \hat{\mathbf{s}}_n \right) \mathbf{u}[n], \tag{2.38}$$

where μ is an adaptation step that should be "small enough". This algorithm is called *least mean squares* (LMS).

2.6 Problems

2.1 Consider the sequence

$$u[1] \ldots u[7] \equiv 0.7,\ -0.1,\ 0.7,\ -0.2,\ -0.1,\ 1.5,\ -1.1$$

which is fed as input to a linear filter of three coefficients, $\mathbf{s} = [s_1, s_2, s_3]^\top$. The following elements of the output sequence are known, (corrupted with Gaussian noise of variance 0.25):

$$x[1] \ldots x[6] \equiv -0.60,\ 1.13,\ 0.57,\ 0.42,\ 1.25,\ -2.58$$

a) What is the ML estimate of $\mathbf{s}$ given the data?
b) Use the obtained filter to predict $x[7]$, $\hat{x}_{\mathrm{ML}}$.
c) Calculate the MSE, MAP and MAD estimates of $\mathbf{s}$ assuming that the prior pdf of its components is $s_i \sim \mathcal{N}(0,1)$ and that they are independent.
d) Get the MSE estimate of $x[7]$, $\hat{x}_{\mathrm{MSE}}$.
e) Calculate the MSE in prediction b). (That is, the mean of $(\hat{x}_{\mathrm{ML}} - x[7])^2$ given the data).
f) Calculate the MSE in prediction d). (That is, the mean of $(\hat{x}_{\mathrm{MMSE}} - x[6])^2$ given the data)

2.7 Appendix: the matrix inversion lemma

In this section we apply the matrix inversion lemma (also known as the Woodbury matrix identity) to obtain expression. This lemma states that, for any matrices $\mathbf{U}$, $\mathbf{C}$, bfV and any non-singular matrix $\mathbf{A}$,

$$(\mathbf{A} + \mathbf{VCU})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{V}\left(\mathbf{C}^{-1} + \mathbf{UA}^{-1}\mathbf{V}\right)^{-1}\mathbf{UA}^{-1} \tag{2.39}$$

Proving this equality is not difficult: multiplying the right-hand side of the equality by $\mathbf{A} + \mathbf{UCV}$, it is easy to see that the result is the identity matrix.

Taking

$$\mathbf{A} = \frac{\sigma_\varepsilon^2}{\sigma_s^2}\mathbf{I} \tag{2.40}$$

$$\mathbf{C} = \mathbf{I} \tag{2.41}$$

$$\mathbf{V} = \mathbf{U}^\top \tag{2.42}$$

we get

$$\begin{aligned}\left(\mathbf{U}^\top\mathbf{U}+\frac{\sigma_\varepsilon^2}{\sigma_s^2}\mathbf{I}\right)^{-1} &= \frac{\sigma_s^2}{\sigma_\varepsilon^2}\mathbf{I}-\left(\frac{\sigma_s^2}{\sigma_\varepsilon^2}\right)^2\mathbf{U}^\top\left(\mathbf{I}+\frac{\sigma_s^2}{\sigma_\varepsilon^2}\mathbf{U}\mathbf{U}^\top\right)^{-1}\mathbf{U}\\ &= \frac{\sigma_s^2}{\sigma_\varepsilon^2}\mathbf{I}-\frac{\sigma_s^2}{\sigma_\varepsilon^2}\mathbf{U}^\top\left(\mathbf{U}\mathbf{U}^\top+\frac{\sigma_\varepsilon^2}{\sigma_s^2}\mathbf{I}\right)^{-1}\mathbf{U}\\ &= \frac{\sigma_s^2}{\sigma_\varepsilon^2}\mathbf{I}-\frac{\sigma_s^2}{\sigma_\varepsilon^2}\mathbf{U}^\top\mathbf{P}\mathbf{U} \qquad (2.43)\end{aligned}$$

Where $\mathbf{P}$ is given by (2.28). Note that, by definition,

$$\left(\mathbf{U}\mathbf{U}^\top+\frac{\sigma_\varepsilon^2}{\sigma_s^2}\mathbf{I}\right)\mathbf{P}=\mathbf{P}\left(\mathbf{U}\mathbf{U}^\top+\frac{\sigma_\varepsilon^2}{\sigma_s^2}\mathbf{I}\right)=\mathbf{I} \qquad (2.44)$$

so that

$$\mathbf{U}\mathbf{U}^\top\mathbf{P}=\mathbf{P}\mathbf{U}\mathbf{U}^\top=\mathbf{I}-\frac{\sigma_\varepsilon^2}{\sigma_s^2}\mathbf{P} \qquad (2.45)$$

We will use these equalities below. Using (2.43) into (2.24), we get

$$\begin{aligned}\mathbf{m}_{\mathbf{S}|\mathbf{X}} &= \mathbf{U}\left(\mathbf{U}^\top\mathbf{U}+\frac{\sigma_\varepsilon^2}{\sigma_s^2}\mathbf{I}\right)^{-1}\mathbf{x}\\ &= \frac{\sigma_s^2}{\sigma_\varepsilon^2}\mathbf{U}\mathbf{x}-\frac{\sigma_s^2}{\sigma_\varepsilon^2}\mathbf{U}\mathbf{U}^\top\mathbf{P}\mathbf{U}\mathbf{x}\\ &= \frac{\sigma_s^2}{\sigma_\varepsilon^2}\mathbf{U}\mathbf{x}-\frac{\sigma_s^2}{\sigma_\varepsilon^2}\left(\mathbf{I}-\frac{\sigma_\varepsilon^2}{\sigma_s^2}\mathbf{P}\right)\mathbf{U}\mathbf{x}\\ &= \frac{\sigma_s^2}{\sigma_\varepsilon^2}\mathbf{U}\mathbf{x}-\frac{\sigma_s^2}{\sigma_\varepsilon^2}\mathbf{U}\mathbf{x}+\mathbf{P}\mathbf{U}\mathbf{x}\\ &= \mathbf{P}\mathbf{U}\mathbf{x} \qquad (2.46)\\ \mathbf{V}_{\mathbf{S}|\mathbf{X}} &= \sigma_s^2\mathbf{I}-\sigma_s^2\mathbf{U}\left(\mathbf{U}^\top\mathbf{U}+\frac{\sigma_\varepsilon^2}{\sigma_s^2}\mathbf{I}\right)^{-1}\mathbf{U}^\top\\ &= \sigma_s^2\left[\mathbf{I}-\mathbf{U}\left(\frac{\sigma_s^2}{\sigma_\varepsilon^2}\mathbf{I}-\frac{\sigma_s^2}{\sigma_\varepsilon^2}\mathbf{U}^\top\mathbf{P}\mathbf{U}\right)\mathbf{U}^\top\right]\\ &= \sigma_s^2\left[\mathbf{I}-\frac{\sigma_s^2}{\sigma_\varepsilon^2}\mathbf{U}\mathbf{U}^\top+\frac{\sigma_s^2}{\sigma_\varepsilon^2}\mathbf{U}\mathbf{U}^\top\mathbf{P}\mathbf{U}\mathbf{U}^\top\right]\\ &= \sigma_s^2\left[\mathbf{I}-\frac{\sigma_s^2}{\sigma_\varepsilon^2}\mathbf{U}\mathbf{U}^\top+\frac{\sigma_s^2}{\sigma_\varepsilon^2}\left(\mathbf{I}-\frac{\sigma_\varepsilon^2}{\sigma_s^2}\mathbf{P}\right)\mathbf{U}\mathbf{U}^\top\right]\\ &= \sigma_\varepsilon^2\mathbf{P} \qquad (2.47)\end{aligned}$$

Chapter 3
Spectral Estimation

3.1 Introduction

This chapter studies a very important estimation problem, which is that of estimating the power spectral density (PSD) of a stationary process. We will consider two families of estimators: 1) classical (or non-parametric) and 2) parametric estimators, which are based on a model for the PSD.

Computing the estimate of $S_x(e^{j\omega})$, which we will denote by $\hat{S}_x(e^{j\omega})$, from an arbitrarily large number of realizations of a stationary process (see Figure 3.1) would be a (relatively) easy task. Of course, this is an idealized scenario as we do not have access to all realizations and, even more, we also do not have access to all time samples of the same realization. Thus, the objective in this chapter is to compute $\hat{S}_x(e^{j\omega})$ from N samples of a single realization of the process $x[n]$.

The spectral estimation problem is defined only for wide-sense stationary (WSS) processes for which the mean function is time-independent, that is, $\mu_x = \mu_x[n] = \mathbb{E}[x[n]]$, and the auto-correlation function depends only on the time difference, i.e., $r_x[m] = r_x[n, n-m] = \mathbb{E}[x[n]x^*[n-m]]$. For non-stationary processes, the usual practice is to apply the estimators to small windows, since on a local scale we can assume that non-stationary processes are WSS. For instance, this is typically done when analyzing speech signals, which are usually described using non-stationary processes. Moreover, since only one realization is available, the process must be ergodic such that expectations can be substituted by time averages.

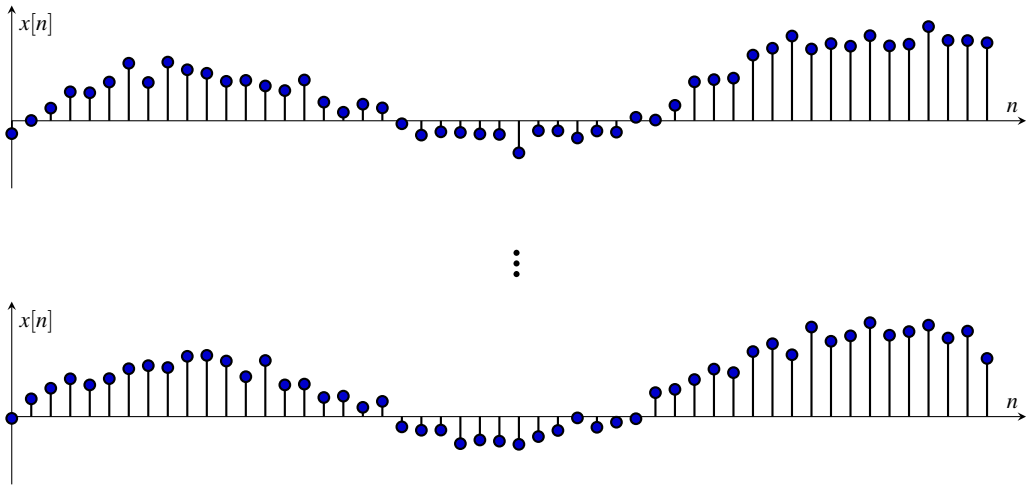


Fig. 3.1 Realizations of a discrete stochastic process

3.2 Preliminaries: Spectral analysis of deterministic signals

Before going into the spectral analysis of stochastic processes, it is convenient to study the case of deterministic signals, which will help us to understand the concept of spectral

resolution. Thus, the problem is to compute the Fourier transform of the deterministic signal $x[n]$. However, this relatively "simple" task has two problems. First, we do not have access to the whole signal $x[n]$, but only to a finite record thereof.

$$x_w[n] = \begin{cases} x[n], & n = 0, \ldots, N-1, \\ 0, & \text{otherwise.} \end{cases}$$

Defining now the window

$$w_{R,N}[n] = \begin{cases} 1, & n = 0, \ldots, N-1, \\ 0, & \text{otherwise,} \end{cases}$$

we may rewrite $x_w[n] = w_{R,N}[n]x[n]$, which allows us to compute the Fourier transform of $x_w[n]$ as[1]

$$X_w(e^{j\omega}) = \mathscr{F}(x_w[n]) = \sum_{n=0}^{N-1} x_w[n]e^{-j\omega N} = \frac{1}{2\pi} W_{R,N}(e^{j\omega}) \circledast X(e^{j\omega}), \tag{3.1}$$

where $\circledast$ denotes the circular convolution. So, the Fourier transform of the windowed signal, $x_w[n]$, is related to that of $x[n]$ through the Fourier transform of the window $w_{R,N}[n]$, which is given by

$$W_{R,N}(e^{j\omega}) = e^{-j\omega(N-1)/2} \frac{\sin\left(\frac{\omega N}{2}\right)}{\sin\left(\frac{\omega}{2}\right)} = e^{-j\omega(N-1)/2} P_N(e^{j\omega}),$$

and its amplitude $|P_N(e^{j\omega})|$ is depicted in Figure 3.2. As this figure shows, the width of the main lobe is $4\pi/N$.

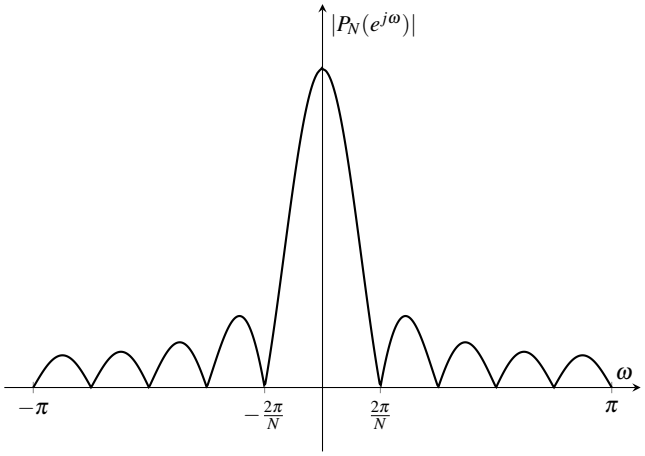


Fig. 3.2 Module of the Fourier transform of the rectangular window

[1] In the "Signals and Systems" parlance, this Fourier transform is named Discrete Time Fourier Transform (DTFT).

The second issue is that the DTFT in (3.1) is a function of a continuous variable. Hence it cannot be computed nor stored in a computer. The solution is simple and consists in discretizing the spectrum, which yields the Discrete Fourier Transform (DFT). Thus, we are only able to compute $X_w(e^{j\omega_k})$, with $\omega_k = 2\pi k/N$ and $k = 0,\ldots,N-1$. The DFT is typically computed using the fast Fourier transform (FFT) algorithm.

The aforementioned procedure based on the DFT/FFT gets only N samples of the spectrum for length-N signals, but we can get more samples by zero-padding the signals, i.e., by simply adding $N_{\text{fft}} - N$ zeros after the N samples. This procedure increases the number of frequencies but it does not increase the resolution as it does not modify the window.

Example 3.1 (Spectral analysis of a complex exponential)

This example considers the spectral analysis of a finite record of a complex exponential, i.e., $x[n] = e^{j\omega_0 n}, n = 0,\ldots,N-1$. Using the DTFT of a complex exponential, given by

$$X(e^{j\omega}) = 2\pi\delta(\omega - \omega_0),$$

and $W_N(e^{j\omega})$, $X_w(e^{j\omega})$ becomes

$$X_w(e^{j\omega}) = e^{-j(\omega-\omega_0)(N-1)/2}\frac{\sin\left(\frac{(\omega-\omega_0)N}{2}\right)}{\sin\left(\frac{\omega-\omega_0}{2}\right)} = e^{-j(\omega-\omega_0)(N-1)/2}P_N(\omega-\omega_0),$$

and its magnitude squared is

$$|X_w(e^{j\omega})|^2 = \left|P_N\left(e^{j(\omega-\omega_0)}\right)\right|^2.$$

Figure 3.3 plots, in logarithmic scale, $|X_w(e^{j\omega})|^2$ and $|X_w(e^{j\omega_k})|^2$ for two different values of N_{fft}.

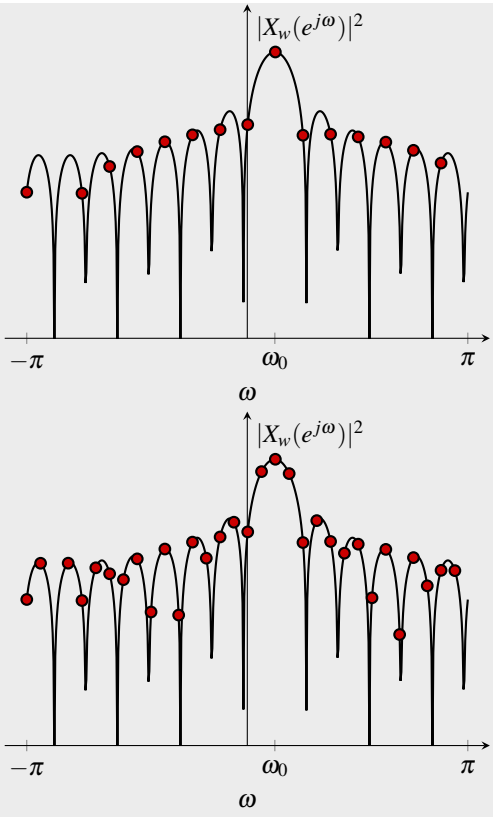


Fig. 3.3 Fourier transform (in logarithmic scale) of a windowed complex exponential

As we have seen in this example, the spectral analysis of deterministic signals depends on two factors. First, the number of available samples, which determines the window and, therefore, the shape of the windowed spectrum. As we have seen in Figure 3.3, the rectangular window has a narrow main lobe at the expense of high secondary lobes. This effect could be reduced by pre-multiplying $x_w[n]$ by a different window, which would reduce the height of the secondary lobes, but it would widen the main lobe. Second, the number of frequencies where the estimated PSD is evaluated.

Example 3.2 (Spectral analysis of two complex exponentials)

This example considers the spectral analysis of a finite record of the sum of two complex exponential, i.e., $x[n] = e^{j\omega_0 n} + e^{j\omega_1 n}, n = 0,\ldots,N-1$, which will help us to understand the concept of resolution. Using (3.1), we have

$$X_w(e^{j\omega}) = e^{-j(\omega-\omega_0)(N-1)/2} P_N(\omega-\omega_0) + e^{-j(\omega-\omega_1)(N-1)/2} P_N(\omega-\omega_1).$$

and its magnitude squared is

$$|X_w(e^{j\omega})|^2 = \left|P_N\left(e^{j(\omega-\omega_0)}\right)\right|^2 + \left|P_N\left(e^{j(\omega-\omega_1)}\right)\right|^2 + 2\cos\left(\frac{(\omega_0-\omega_1)(N-1)}{2}\right) P_N\left(e^{j(\omega-\omega_0)}\right) P_N\left(e^{j(\omega-\omega_1)}\right).$$

Figure 3.4 plots, in logarithmic scale, $|X_w(e^{j\omega})|^2$ for two different separations between the frequencies of the exponential components. As we can see in this figure, for small frequency separations, it is impossible to identify in the spectrum the two complex exponentials.

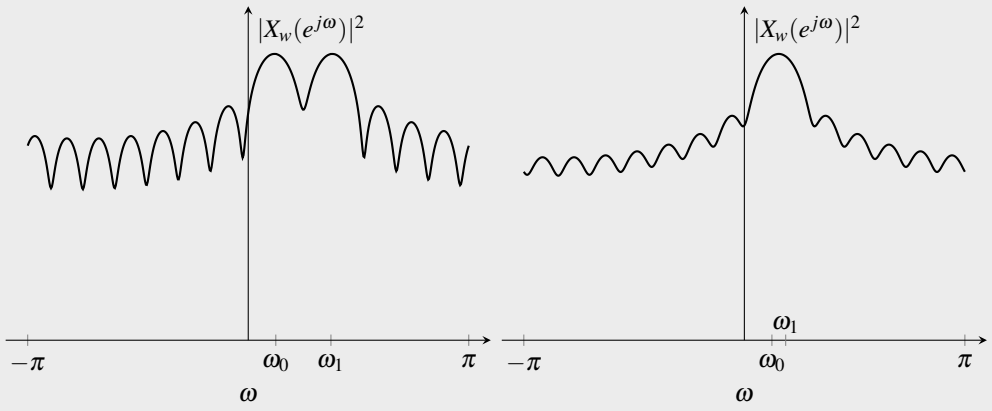


Fig. 3.4 Fourier transform (in logarithmic scale) of the sum of two complex exponentials

3.3 Non-parametric methods in spectral estimation

In this section, we turn our attention to the case of stochastic signals and, in particular, to the development of non-parametric spectral estimation methods. We will therefore study the periodogram and variations thereof.

Before proceeding, let us note that throughout this section, we will only consider DTFTs. However, we have to keep in mind that, in practice, we can only compute the DFT (using the FFT algorithm), as we have seen in Section 3.2.

Remind that the power spectral density, or power spectrum, of a stochastic process $x[n]$, is defined as

$$S_x(e^{j\omega}) = \lim_{N\to\infty} \frac{1}{2N-1}\mathbb{E}\left[\left|\sum_{n=-N+1}^{N-1} x[n]e^{-j\omega n}\right|^2\right]. \tag{3.2}$$

If the process is WSS and the autocorrelation is absolutely summable (the usual case in practice), this definition is equivalent to the Fourier transform of the autocorrelation, i.e.,

$$S_x(e^{j\omega}) = \mathscr{F}(r_x[m]), \tag{3.3}$$

where

$$r_x[m] = \mathbb{E}[x[n]x^*[n-m]], \tag{3.4}$$

These alternative expressions for the power spectrum motivate two different strategies for spectral estimation:

1. Drop the limit in (3.2) and estimate the expectation from an finite sample.
2. Estimate the autocorrelation function and compute its Fourier transform.

Both strategies are closely related, as we will see in the following sections.

3.3.1 The periodogram and the correlogram

The **periodogram**, which is a term coined by Arthur Schuster in 1898, is obtained from on (3.2) by simply dropping the expectation and considering a finite number of samples, i.e.,

$$\hat{S}_x^p(e^{j\omega}) = \frac{1}{N}\left|\sum_{n=0}^{N-1} x[n]e^{-j\omega n}\right|^2 = \frac{1}{N}\left|X_N(e^{j\omega})\right|^2, \tag{3.5}$$

where $X_N(e^{j\omega}) = \mathscr{F}(x_N[n])$ is the Fourier transform of the truncated version of $x[n]$

$$x_N[n] = \begin{cases} x[n], & 0 \le n \le N-1 \\ 0, & \text{otherwise} \end{cases} \tag{3.6}$$

The **correlogram**, based on (3.3), is defined as

$$\hat{S}_x^c(e^{j\omega}) = \mathscr{F}(\hat{r}_x[m]),$$

where

$$\hat{r}_x[m] = \frac{1}{N}\sum_{n=m}^{N-1} x[n]x^*[n-m], \quad m = 0,\ldots,N-1, \tag{3.7}$$

Despite the differences in names and definitions, the correlogram and the periodogram are identical estimators: rewriting $\hat{r}_x[m]$ as

$$\hat{r}_x[m] = \frac{1}{N}\sum_{n=m}^{N-1} x[n]x^*[n-m] = \frac{1}{N}\sum_{n=-\infty}^{\infty} x_N[n]x_N^*[n-m] = \frac{1}{N}x_N[n] * x_N^*[-n],$$

and taking its Fourier transform yields

$$\hat{S}_x^c(e^{j\omega}) = \mathscr{F}(\hat{r}_x[m]) = \frac{1}{N}\mathscr{F}\left(x_N[n] * x_N^*[-n]\right).$$

Finally, applying the properties of the Fourier transform, $\hat{S}_x^c(e^{j\omega})$ simplifies to

$$\hat{S}_x^c(e^{j\omega}) = \frac{1}{N}\mathscr{F}\left(x_N[n]\right)\mathscr{F}\left(x_N^*[-n]\right) = \frac{1}{N}X_N(e^{j\omega})X_N^*(e^{j\omega}) = \frac{1}{N}\left|X_N(e^{j\omega})\right|^2 = \hat{S}_x^p(e^{j\omega}),$$

which is the periodogram in (3.5).

3.3.1.1 Bias and variance of the periodogram

To understand why we need more refined estimators of the power spectral density, now we shall perform the statistical analysis of the periodogram (or correlogram), i.e., we will compute its bias and variance as we would do with any other estimator.

First, note that the autocorrelation estimate in (3.3.1) is biased:

$$\begin{aligned}\mathbb{E}\left[\hat{r}_x[m]\right] &= \frac{1}{N}\sum_{n=m}^{N-1}\mathbb{E}\left[x[n]x^*[n-m]\right] = \frac{1}{N}\sum_{n=m}^{N-1} r_x[m] \\ &= \frac{N-|m|}{N} r_x[m], \end{aligned} \tag{3.8}$$

Note that, for $m = 0$ or any m such that $r_x[m] = 0$, the autocorrelation estimate is unbiased. However,for any other values, $\mathbb{E}\left[\hat{r}_x[m]\right] \neq r_x[m]$, and the relative value of the bias increases with m.

Seemingly, we can easily remove this bias by replacing the factor $1/N$ in (3.7) by $1/(N-|m|)$. In this way, we would get unbiased estimates of the autocorrelation values for any m. Unfortunately, it can be show that the resulting function is, in general, not a feasible autocorrelation function (i.e. its Fourier transform may take negative values at some frequencies).

As a consequence of the autocorrelation bias, the periodogram is also biased. To see it, it is useful to define the triangular, or Barlett window, as

$$w_{T,N}[m] = \begin{cases} \frac{N-|m|}{N}, & |m| \leq N-1, \\ 0, & \text{otherwise}, \end{cases}$$

which is depicted in Figure 3.5.

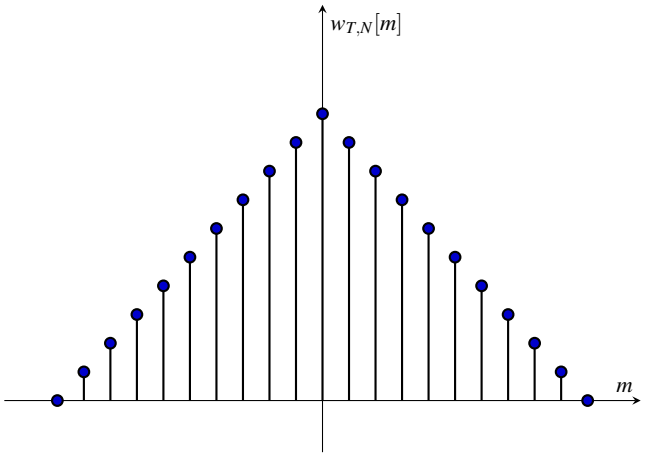


Fig. 3.5 Triangular window

The mean of the autocorrelation in (3.8) can then be written as

$$\mathbb{E}\left[\hat{r}_x[m]\right] = \frac{N-|m|}{N} r_x[m] = w_{T,N}[m] r_x[m], \tag{3.9}$$

Using (3.9), the bias of the periodogram becomes

$$\mathbb{E}\left[\hat{S}_x^p(e^{j\omega})\right] = \mathscr{F}\left(w_{T,N}[m] r_x[m]\right) = \frac{1}{2\pi} W_{T,N}(e^{j\omega}) \circledast S_x(e^{j\omega}), \tag{3.10}$$

where

$$\begin{aligned} W_{T,N}(e^{j\omega}) &= \mathscr{F}\left(w_{T,N}[m]\right) = \frac{1}{N}\mathscr{F}\left(w_{R,N}[m] * w_{R,N}[-m]\right) = |W_{R,N}(e^{j\omega})|^2 \\ &= \frac{1}{N}\frac{\sin^2\left(\frac{\omega N}{2}\right)}{\sin^2\left(\frac{\omega}{2}\right)}, \end{aligned}$$

is the Fourier transform of the triangular window and is depicted in Figure 3.6. Comparing Figures 3.2 and 3.6, it can be seen that the level of secondary lobes is smaller for the triangular window. By analogy with Example 3.2, we can say that the bias of the periodogram is related with its resolution.

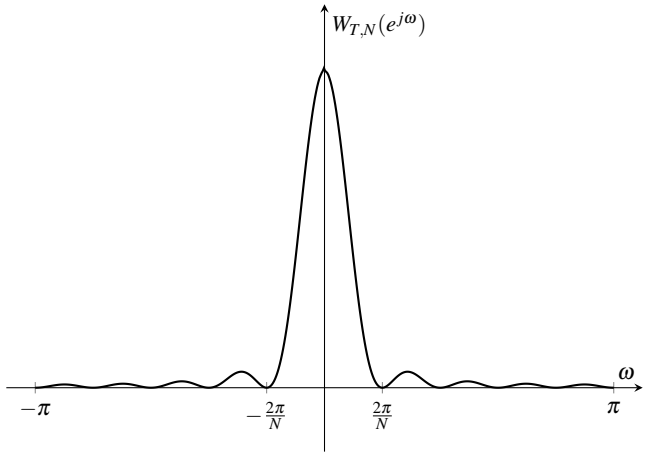


Fig. 3.6 Fourier transform of the triangular window

Note that the width of the main lobe is inversely proportional to N. Therefore, by increasing the sample size we can narrow the main lobe and reduce the bias. This is because the periodogram is asymptotically unbiased: the bias reduces to zero as N grows.

$$\lim_{N\to\infty} \hat{S}_x^p(e^{j\omega}) = S_x(e^{j\omega}).$$

As a special case, note that the autocorrelation function of a white process is zero for all $m \neq 0$, therefore, the estimate of its autocorrelation is unbiased, and so the periodogram.

The analysis of the variance of the periodogram is cumbersome and can only be done in particular cases. For white noise, it can be shown that

$$Var\left(\hat{S}_x^p(e^{j\omega})\right) = S_x^2(e^{j\omega}),$$

and in general we can say that

$$Var\left(\hat{S}_x^p(e^{j\omega})\right) \propto S_x^2(e^{j\omega}),$$

where $\propto$ denotes approximately proportional to. This expression tells us that the variance does not decrease for larger data records. That is, the periodogram is not a consistent estimate of the PSD.

3.3.2 The Blackman-Tukey estimator

One of the reasons for the behavior of the periodogram variance is the poor quality of the estimate $\hat{r}_x[m]$ for values of m close to N. This problem is what the Blackman-Tukey (BT) estimator tries to improve. The idea is to ignore or weight the samples of $\hat{r}_x[m]$ for m close to N. Thus, the BT estimator is

$$\hat{S}_x^{BT}(e^{j\omega}) = \mathscr{F}(w_M[m]\hat{r}_x[m]) = \sum_{m=-N+1}^{N-1} w_M[m]\hat{r}_x[m]e^{-j\omega m}, \tag{3.11}$$

where $w[m]$ is a window that must fulfill

$$w_M[m] = \begin{cases} f(|m|), & |m| \leq M-1, \\ 0, & \text{otherwise}, \end{cases}$$

where $f(|m|)$ is a monotonically decreasing function of $|m|$ and $M \leq N$. This window ignores the lags of the estimated auto-correlation for $|m| > M-1$ and weights the lags for large m. The choice of the window is critical to achieve good performance, but, in any case, it must guarantee that $\hat{S}_x^{BT}(e^{j\omega}) \geq 0$.

Using the properties of the Fourier transform, we may rewrite $\hat{S}_x^{BT}(e^{j\omega})$ as

$$\hat{S}_x^{BT}(e^{j\omega}) = \frac{1}{2\pi} W_M(e^{j\omega}) \circledast \hat{S}_x^p(e^{j\omega}) = \frac{1}{2\pi}\int_{-\pi}^{\pi} W_M(e^{j\psi})\hat{S}_x^p(e^{j(\omega-\psi)})d\psi,$$

where $W_M(e^{j\omega}) = \mathscr{F}(w_M[n])$. Then, the Blackman-Tukey estimator is locally smoothing the periodogram, which reduces its variance. However, there is no free lunch and we will show that this variance reduction translates into lower resolution (or larger bias). Concretely, the bias of the BT estimator is

$$\mathbb{E}\left[\hat{S}_x^{BT}(e^{j\omega})\right] = \frac{1}{2\pi} W_M(e^{j\omega}) \circledast \mathbb{E}\left[\hat{S}_x^p(e^{j\omega})\right] = \frac{1}{2\pi} W_M(e^{j\omega}) \circledast W_{T,N}(e^{j\omega}) \circledast S_x(e^{j\omega}).$$

Finally, since $w_M[n]$ is shorter than $w_{T,N}[n]$, it can be shown that $W_M(e^{j\omega})$ is wider than $W_{T,M}(e^{j\omega})$, which translates into a lower resolution. This behavior is depicted in Figure 3.7 for $w_M[n] = w_{T,M}[n]$. Note that the y-axis is in logarithmic scale.

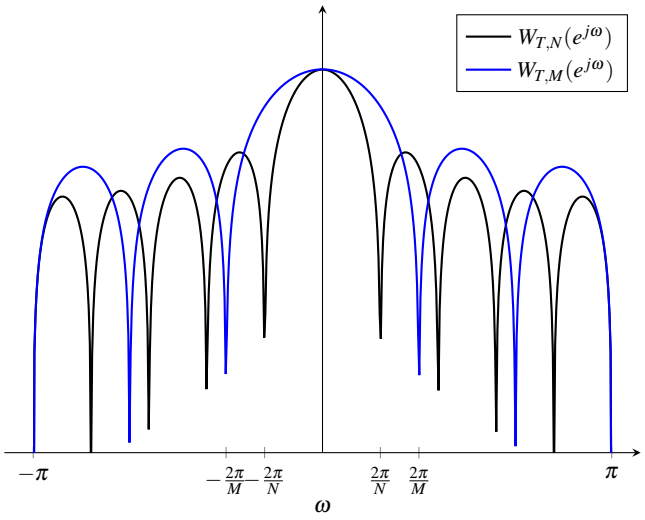


Fig. 3.7 Fourier transform (in logarithmic scale) of two triangular windows of different lengths

3.3.3 Estimators based on the averaged periodogram

The Blackman-Tukey estimator yields a smaller variance than that of the periodogram because, as we have seen, it smooths the periodogram. An alternative to reduce the variance is to average several periodograms. However, the question is: How do we obtain such periodograms? The answer is easy and consists in dividing the N observations into windows of length $M < N$.

The Barlett method is one of the possible estimators based on the averaged periodogram. First, it divides the N observations into L non-overlapping windows of length M as

$$x_l[n] = x[(l-1)M+n], \tag{3.12}$$

where $n = 0,\ldots,M-1$, and $l = 1,\ldots,L$, and computes the periodogram of each window, that is,

$$\hat{S}^p_{x,l}(e^{j\omega}) = \frac{1}{M}\left|\sum_{n=0}^{M-1} x_l[n]e^{-j\omega n}\right|^2 = \frac{1}{M}\left|X_l(e^{j\omega})\right|^2. \tag{3.13}$$

Then, the Barlett estimator is given by simply averaging the individual periodograms

$$\hat{S}^B_x(e^{j\omega}) = \frac{1}{L}\sum_{l=1}^{L}\hat{S}^p_{x,l}(e^{j\omega}). \tag{3.14}$$

Although it is out of the scope of these notes, we must point out that $\hat{S}^B_x(e^{j\omega})$ is somehow related to $\hat{S}^{BT}_x(e^{j\omega})$.

There are two further improvements of the periodogram. The first one is based on the Barlett estimator but substituting the individual periodograms by Blackman-Tukey esti-

mates. The second one is based on dividing the N observations into L overlapping windows. The combination of both improvements is known as the Welch method.

One final question remains: What happens to the bias and variance of these methods. Regarding the bias, it is going to be higher (lower resolution) than that of the periodogram since $M < N$, as also happened to the Blackman-Tukey estimate. As for the variance, it is going to be reduced by a factor of L, the number of windows. That is,

$$Var\left(\hat{S}_x^{ap}(e^{j\omega})\right) \approx \frac{1}{L} Var\left(\hat{S}_x^{p}(e^{j\omega})\right),$$

where $\hat{S}_x^{ap}(e^{j\omega})$ is any averaged periodogram (either Barlett or Welch methods) and $\approx$ is due to the non-independence between the windows. It would be an equality when the windows are independent, i.e., the Barlett method.

3.4 Parametric methods in spectral estimation

The problem of non-parametric methods is that they estimate an infinite number of parameters (the PSD at each frequency) from a sequence of N observations. Clearly, this is an ill-posed problem since there are (many) more parameters to estimate than observations. To overcome this issue, we could postulate a parametric model for the PSD and estimate only the parameters of such model using the N observations. For instance, the model could be $S_x(e^{j\omega}) = a + b\cos^2(\omega)$ and, hence, we only have to estimate a and b.

Parametric approaches, as described above, can provide a significant performance boost if the signal fits the postulated model, otherwise the performance could be even worse than that of non-parametric methods. It is therefore of the utmost importance to select the proper model.

In this chapter we will analyze autorregresive (AR) models, which are particularly amenable for parametric estimation.

3.4.1 Auto-Regressive (AR) models

We say that the stochastic process $x[n]$ follows an auto-regressive model of order p (or, simply, an AR(p) model, if it is the output of a causal, linear and time-invariant filter driven by the recursive relation

$$x[n] = u[n] - \sum_{k=1}^{p} a_k x[n-k], \tag{3.15}$$

when the input, $u[n]$ is an IID Gaussian process with zero mean and variance $\sigma^2 > 0$.

The impulse response of the filter can be represented by means of the recursive relation (3.15), replacing $u[n]$ by $\delta[n]$

$$h[n] = \delta[n] - \sum_{k=1}^{p} a_k h[n-k],$$

Applying the Fourier transform to this equation, we get,

$$H(e^{j\omega}) = 1 - \sum_{k=1}^{p} a_k e^{-j\omega k} H(e^{j\omega}),$$

therefore

$$H(e^{j\omega}) = \frac{1}{1 + \sum_{k=1}^{p} a_k e^{-j\omega k}},$$

3.4.1.1 Power Spectum

Since $x[n]$ is the output of filter $h[n]$ for input $u[n]$,

$$x[n] = u[n] * h[n] \tag{3.16}$$

therefore, the power spectrum is

$$S_x(e^{j\omega}) = S_u(e^{j\omega})|H(e^{j\omega})|^2 = \frac{\sigma^2}{\left|1 + \sum_{k=1}^{p} a_k e^{-j\omega k}\right|^2}. \tag{3.17}$$

According to Weierstrass theorem, for large values of p, the PSD model in (3.17) can approximate arbitrarily close any continuous PSD. Hence, there is a strong interest in this kind of models.

3.4.2 Auto-correlation

Using (3.15), we can express the auto-correlation function of the AR(p) process $x[n]$ as

$$\begin{aligned} r_x[m] &= \mathbb{E}\{x[n]x^*[n-m]\} \\ &= \mathbb{E}\left\{\left(u[n] - \sum_{k=1}^{p} a_k x[n-k]\right) x^*[n-m]\right\} \\ &= \mathbb{E}\{u[n]x^*[n-m]\} - \sum_{k=1}^{p} a_k r_x[m-k] \end{aligned} \tag{3.18}$$

We can develop this expression by analyzing the cases $m > 0$, $m = 0$ and $m < 0$ separately.

- For $m > 0$, since the filter is causal, $x^*[n-m]$ does not depend on $u[n]$, therefore,

$$\mathbb{E}\{u[n]x^*[n-m]\} = \mathbb{E}\{u[n]\}\mathbb{E}\{x^*[n-m]\} = 0, \qquad m > 0 \tag{3.19}$$

therefore, (3.18) simplifies to

$$r_x[m] = -\sum_{k=1}^{p} a_k r_x[m-k], \qquad m > 0$$

- For $m = 0$, we get,

$$\begin{aligned}\mathbb{E}\{u[n]x^*[n-m]\} &= \mathbb{E}\{u[n]x^*[n]\} \\ &= \mathbb{E}\left\{u[n]\left(u^*[n] - \sum_{k=1}^{p} a_k^* x^*[n-k]\right)\right\} \\ &= \mathbb{E}\{u[n]u^*[n]\} = \sigma^2 \end{aligned} \tag{3.20}$$

  therefore

$$r_x[0] = \sigma^2 - \sum_{k=1}^{p} a_k r_x[m-k].$$

- Finally, for $m < 0$, since the autocorrelation function is Hermitian, we have $r_x[m] = r_x^*[-m]$.

  Joining the cases $m > 0$, $m = 0$ and $m < 0$, we get

$$r_x[m] = \begin{cases} -\sum_{k=1}^{p} a_k r_x[m-k] + \sigma^2, & m = 0, \\ -\sum_{k=1}^{p} a_k r_x[m-k], & m > 0, \\ r_x^*[-m], & m < 0. \end{cases} \tag{3.21}$$

We see in (3.21) that the relationship between the model parameters (σ^2, $a_1, \ldots, a_p$) and the auto-correlation is linear. which simplifies the estimation of such parameters. The estimation procedure consists in substituting the theoretical auto-correlation by an estimate and then solving a linear system of equations. The PSD estimate is obtained by substituting the estimated parameters in the corresponding model.

3.4.3 Parameter estimation from the autocorrelation

Since we need to obtain $p+1$ parameters, i.e., $a_1, \ldots, a_p$ and σ^2, we need $p+1$ equations from (3.21), which are known as the Yule-Walker equations:

$$\begin{aligned} r_x[0] &= -a_1 r_x[-1] - a_2 r_x[-2] + \ldots - a_p r_x[-p] + \sigma^2, \\ r_x[1] &= -a_1 r_x[0] - a_2 r_x[-1] + \ldots - a_p r_x[-p+1], \\ &\vdots \\ r_x[p] &= -a_1 r_x[p-1] - a_2 r_x[p-2] + \ldots - a_p r_x[0]. \end{aligned}$$

The last p equations depend only on $a_1, \ldots, a_p$. We can write them in matrix form by using $r_x[-m] = r_x^*[m]$ and defining

$$\mathbf{r}_x = \left[r_x[1]\ r_x[2]\ \cdots\ r_x[p]\right]^T,$$

and

$$\mathbf{R}_x = \begin{bmatrix} r_x[0] & r_x^*[1] & \cdots & r_x^*[p-1] \\ r_x[1] & r_x[0] & \cdots & r_x^*[p-2] \\ \vdots & \vdots & \ddots & \vdots \\ r_x[p-1] & r_x[p-2] & \cdots & r_x[0] \end{bmatrix},$$

to arrive at

$$\begin{bmatrix} \hat{a}_1 \\ \hat{a}_2 \\ \vdots \\ \hat{a}_p \end{bmatrix} = -\mathbf{R}_x^{-1}\mathbf{r}_x,$$

The matrix $\mathbf{R}_x$ has a special structure, namely, it is constant along diagonals, and is therefore known as Toeplitz. This fact is important for solving the system of equations (computing the matrix inverse) as it reduces the complexity from $\mathcal{O}(p^3)$ to $\mathcal{O}(p^2)$. The remaining parameter to be estimated, the variance, is easily obtained as

$$\hat{\sigma}^2 = r_x[0] + \hat{a}_1 r_x^*[1] + \hat{a}_2 r_x^*[2] + \ldots + \hat{a}_p r_x^*[p].$$

Finally, as we have already seen before, in practical scenarios the auto-correlation function is not available and must therefore be replaced by an estimate. The overall procedure consist on three steps:

1. Estimate the autocorrelation function
2. Compute the model parameters solving the Yule-Walker equations
3. Estimate the power spectrum using the parameter estimates.

3.4.4 Maximum Likelihood estimation

We can avoid the estimation of the autocorrelation function by applying estimation theory to compute the model parameters from the signal observations. Assume we have recorded a set of N observations from the process, $\{x[n], 0 \leq n \leq N-1\}$. Our goal is to estimate the model parameters based on these observations. To do so, the following vector notation will be useful:

$$\mathbf{a} = (a_1, a_2, \ldots, a_{p-1})^\top \tag{3.22}$$

$$\mathbf{x} = (x[p], x[p+1], \ldots, x[N-1])^\top \tag{3.23}$$

$$\mathbf{x}_n = (x[n-1], x[n-2], \ldots, x[n-p])^\top \tag{3.24}$$

Note that $\mathbf{x}_n$ contains all values from the signal record that multiply the coefficients to determine $x[n]$, in such a way that the signal model (3.15) can be expressed as

$$x[n] = u[n] - \mathbf{x}_n^\top \mathbf{a}, \tag{3.25}$$

Also, note that vector $\mathbf{x}$ contains all observations from the process starting for $n > p$. We have excluded the observations for $n < p$ because the maximization of the complete likelihood including these values (i.e. $p(\mathbf{x}, \mathbf{x}_p \mid \mathbf{a})$) becomes much more complex. To avoid these complications, we will take $\mathbf{x}_p$ as fixed data, maximizing the likelihood for the rest of observations, that is, we will compute the estimate

$$\hat{\mathbf{a}}_{\mathrm{ML}} = \underset{\mathbf{a}}{\operatorname{argmax}}\, p(\mathbf{x} \mid \mathbf{x}_p, \mathbf{a}) \tag{3.26}$$

To do so, using (3.25), we factorize the likelihood as

$$\begin{aligned} p(\mathbf{x} \mid \mathbf{x}_p, \mathbf{a}) =& p(x[N-1], x[N-2], \ldots, x[p] \mid \mathbf{x}_p, \mathbf{a}) \\ =& p(x[N-1] \mid x[N-2], \ldots, x[p], \mathbf{x}_p, \mathbf{a}) \cdot p(x[N-2] \mid x[N-3], \ldots, x[p], \mathbf{x}_p \mathbf{a}) \\ & \cdot \ldots \cdot p(x[p] \mid \mathbf{x}_p, \mathbf{a}) \\ =& \prod_{n=p}^{N-1} p(x[n] \mid \mathbf{x}_n, \mathbf{a}) \end{aligned} \tag{3.27}$$

Since the input process is IID, zero-mean, Gaussian, we can write

$$p(x[n] \mid \mathbf{x}_n, \mathbf{a}) = \mathcal{N}\left(x[n] - \mathbf{x}_n^\top \mathbf{a}, \sigma^2\right) \tag{3.28}$$

therefore

$$\begin{aligned} p(\mathbf{x} \mid \mathbf{x}_p, \mathbf{a}) =& \prod_{n=p}^{N-1} \mathcal{N}\left(x[n] - \mathbf{x}_n^\top \mathbf{a}, \sigma^2\right) \\ =& \frac{1}{\sigma^{N-p}(2\pi)^{(N-p)/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{n=p}^{N-1} \left(x[n] - \mathbf{x}_n^\top \mathbf{a}\right)^2\right) \end{aligned} \tag{3.29}$$

Therefore, the ML estimate is the solution of a least squares problem

$$\hat{\mathbf{a}}_{\mathrm{ML}} = \underset{\mathbf{a}}{\operatorname{argmin}} \sum_{n=p}^{N-1} \left(x[n] - \mathbf{x}_n^\top \mathbf{a}\right)^2 \tag{3.30}$$

which can be expressed in matrix form by defining the observation matrix

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_p^\top \\ \mathbf{x}_{p+1}^\top \\ \vdots \\ \mathbf{x}_{N-1}^\top \end{pmatrix} \tag{3.31}$$

to arrive at

$$\hat{\mathbf{a}}_{\mathrm{ML}} = \underset{\mathbf{a}}{\operatorname{argmin}} \|\mathbf{x} - \mathbf{X}\mathbf{a}\|^2 = \left(\mathbf{X}\mathbf{X}^\top\right)^{-1} \mathbf{X}^\top \mathbf{x} \tag{3.32}$$

The same approach can be applied to the estimation of the variance, σ^2:

$$\hat{\sigma}^2_{\mathrm{ML}} = \underset{\sigma}{\operatorname{argmax}} \frac{1}{\sigma^{N-p}(2\pi)^{(N-p)/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{n=p}^{N-1}\left(x[n]-\mathbf{x}_n^\top \mathbf{a}\right)^2\right)$$
$$= \frac{1}{N-p}\sum_{n=p}^{N-1}\left(x[n]-\mathbf{x}_n^\top \mathbf{a}\right)^2 \tag{3.33}$$

3.4.5 Signal prediction

The estimation of the parameters and an AR process can be done by simply solving a system of equations. Moreover, from the expression of an AR model

$$x[n] = u[n] - \sum_{k=1}^{p} a_k x[n-k],$$

we note that they can be used to predict future samples by ignoring the input, i.e.,

$$x[n] = -\sum_{k=1}^{p} \hat{a}_k x[n-k],$$

where the coefficients of the model have been replaced by some estimates. That is, from a record of N samples, $x[0],\ldots,x[N-1]$, we can estimate the model parameters and, afterwards, we can predict $x[N], x[N+1], \ldots$.

Chapter 4
Statistical Detection Theory

4.1 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.

4.1.1 Example 1: Binary detection with no observations

Problem 4.1 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 4.1 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 4.1 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 $P_H(1) = 1/6$.

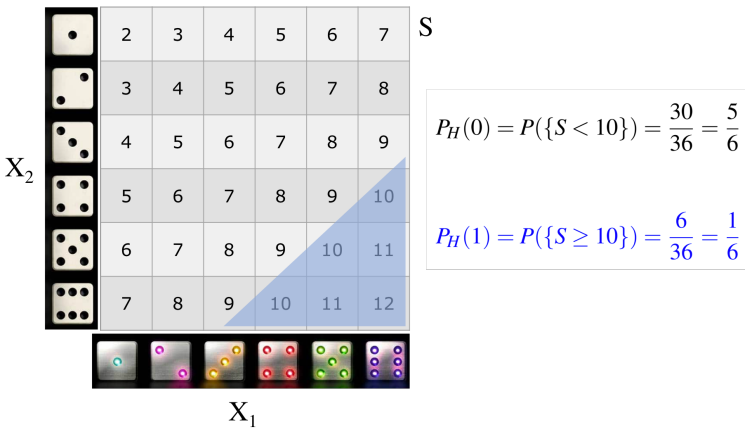


Fig. 4.1 All combinations of of X_1 and X_2 are equally probable, and therefore each of the 36 results represented in the figure have a probability of $1/36$. Counting the number of occurrences of particular values of S or H, the distribution of these variables can be calculated.

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.

4.1.2 Example 2: Binary decision with observations

Problem 4.2 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 4.2 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 4.2 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.

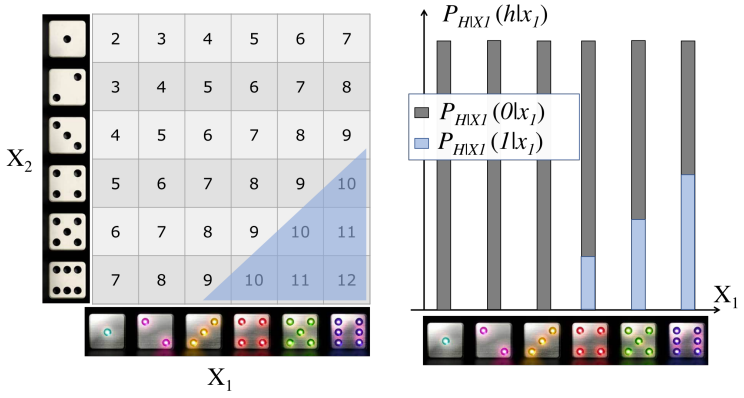


Fig. 4.2 To calculate posterior probabilities of the hypothesis, we need to count how many results in each column correspond to hypothesis 0 and how many correspond to hypothesis 1. Note that $P_{H|X_1}(0|x_1) + P_{H|X_1}(1|x_1) = 1$ for all values of X_1.

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.

4.1.3 Example 3: Working the solution from the likelihoods

Problem 4.3 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 4.3 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).

4.2 Introduction to Decision Theory

In this section we provide a formal presentation of decision theory. Appendix **??** 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.

4.2.1 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. 4.3, 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.

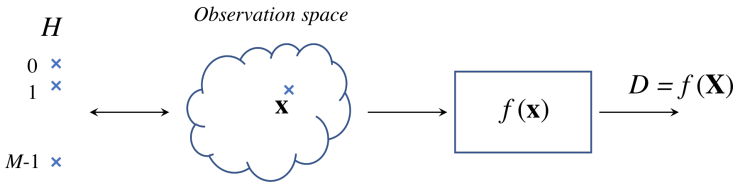


Fig. 4.3 Diagram block of hypothesis testing problems.

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 4.1 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 4.2 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. 4.4.

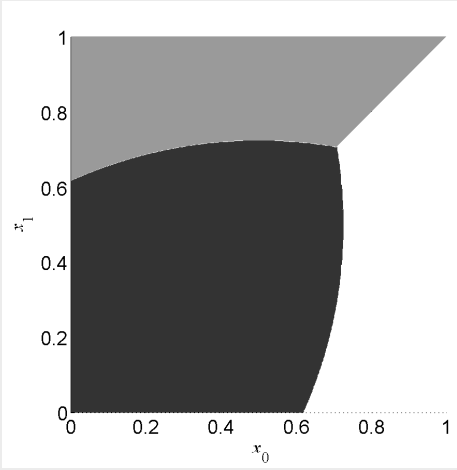


Fig. 4.4 Decision regions for the detector given in Example 4.2: $\mathscr{X}_0$ (black), $\mathscr{X}_1$ (grey), and $\mathscr{X}_2$ (white).

4.2.2 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 4.1.1,

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

- 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.

4.3 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.

4.3.1 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

$$P(H \neq d | \mathbf{x}) = 1 - P(H = d | \mathbf{x}) = 1 - P_{H|\mathbf{X}}(d|\mathbf{x}) = \sum_{\substack{l=0 \\ l \neq d}}^{M-1} P_{H|\mathbf{X}}(l|\mathbf{x}) \tag{4.1}$$

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 $P_{H|\mathbf{X}}(2|\mathbf{x}_o)$.

4.3.1.1 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.

4.3.2 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 4.5 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.

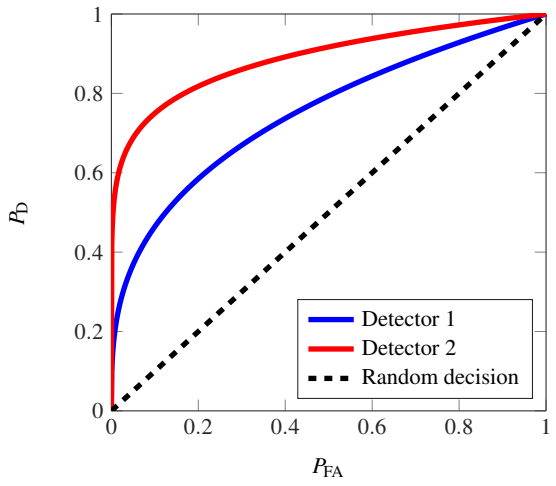


Fig. 4.5 ROC curves for two different detectors.

4.3.3 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,

$$R_\phi = \mathbb{E}\{c_{DH}\} = \sum_{h=0}^{M-1}\sum_{d=0}^{M-1} c_{dh} P_H(h) P_{D|H}(d|h). \tag{4.2}$$

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

$$P_{D|H}(d|h) = P\{X \in \mathscr{X}_d | H = h\} = \int_{\mathscr{X}_d} p_{\mathbf{X}|H}(\mathbf{x}|h) d\mathbf{x} \tag{4.3}$$

therefore

$$R_\phi = \sum_{h=0}^{M-1} P_H(h) \sum_{d=0}^{M-1} c_{dh} \int_{\mathbf{x}\in\mathscr{X}_d} p_{\mathbf{X}|H}(\mathbf{x}|h) d\mathbf{x}. \tag{4.4}$$

Example 4.3 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 (4.2) 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 (4.3)

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

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

$$\mathbb{E}\{c_{dH}|\mathbf{x}\} = \sum_{h=0}^{M-1} c_{dh} P_{H|X}(h|\mathbf{x}). \tag{4.5}$$

Example 4.4 Continuing with Example 4.3, 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 $d=2$:
$$\begin{aligned}\mathbb{E}\{c(2,H)|x\} &= c_{20}P_{H|X}(0|x) + c_{21}P_{H|X}(1|x) + c_{22}P_{H|X}(2|x)\\ &= 1\cdot 0.4 + 1\cdot 0.6(1-x) + 0\cdot 0.6x = 1 - 0.6x\end{aligned}$$

4.3.3.1 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

$$\begin{aligned}R_\phi &= c_{00}P(D=0,H=0) + c_{01}P(D=0,H=1)\\ &= c_{00}P(D=0|H=0)P_H(0) + c_{01}P_{\mathrm{M}}P_H(1) + c_{10}P_{\mathrm{FA}}P_H(0) + c_{11}P_{\mathrm{D}}P_H(1).\\ &= c_{00}(1-P_{\mathrm{FA}})P_H(0) + c_{01}P_{\mathrm{M}}P_H(1) + c_{10}P_{\mathrm{FA}}P_H(0) + c_{11}(1-P_{\mathrm{M}})P_H(1).\\ &= (c_{00}P_H(0) + c_{11}P_H(1)) + (c_{01}-c_{11})P_{\mathrm{M}}P_H(1) + (c_{10}-c_{00})P_{\mathrm{FA}}P_H(0)\end{aligned} \tag{4.6}$$

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 $P_{\mathrm{M}}=0$ and $P_{\mathrm{FA}}=0$ who succeeds with probability 1.
- $(c_{01}-c_{11})P_H(1)P_{\mathrm{M}}$ is the increase in risk caused by miss errors.
- $(c_{10}-c_{00})P_H(0)P_{\mathrm{FA}}$ 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 (4.6) is minimized.

4.4 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.

4.4.1 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 (4.1) 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.

4.4.1.1 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)

$$\log \left[\frac{p_{\mathbf{X}|H}(\mathbf{x}|1)}{p_{\mathbf{X}|H}(\mathbf{x}|0)} \right] = \log p_{\mathbf{X}|H}(\mathbf{x}|1) - \log p_{\mathbf{X}|H}(\mathbf{x}|0) \underset{D=0}{\overset{D=1}{\gtrless}} 0, \tag{4.7}$$

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

$$\frac{p_{\mathbf{X}|H}(\mathbf{x}|1)}{p_{\mathbf{X}|H}(\mathbf{x}|0)} \underset{D=0}{\overset{D=1}{\gtrless}} \frac{P_H(0)}{P_H(1)}. \tag{4.8}$$

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

$$\frac{p_{\mathbf{X}|H}(\mathbf{x}|1)}{p_{\mathbf{X}|H}(\mathbf{x}|0)} \underset{D=0}{\overset{D=1}{\gtrless}} \eta, \tag{4.9}$$

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 4.5.[2]

4.4.2 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 (4.5), which we reproduce here for convenience:

$$\mathbb{E}\{c_{dH}|\mathbf{x}\} = \sum_{h=0}^{M-1} c_{dh} P_{H|\mathbf{X}}(h|\mathbf{x}). \tag{4.10}$$

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 4.5 Continuing with Example 4.3, 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:

[2] 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

$$\begin{aligned}d^\star &= \arg\min_d \sum_{h=0}^{M-1} c_{dh}P_{H|\mathbf{X}}(h|\mathbf{x})\\ &= \arg\min_d\ c\sum_{h\neq d} P_{H|\mathbf{X}}(h|\mathbf{x})\\ &= \arg\min_d 1-P_{H|\mathbf{X}}(d|\mathbf{x})\\ &= \arg\max_d P_{H|\mathbf{X}}(d|\mathbf{x})\\ &= d_{MAP}.\end{aligned} \tag{4.11}$$

4.4.2.1 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 (4.10) 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.

4.4.3 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 (4.9) for different values of η. The Neyman-Pearson detector is a classical example.

4.4.3.1 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.

4.4.3.2 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

$$P_{\text{FA}} = P_{\text{M}} \tag{4.12}$$

4.5 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 (4.7) 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,

$$(\mathbf{x}-\mathbf{m}_0)^T\mathbf{V}_0^{-1}(\mathbf{x}-\mathbf{m}_0) - (\mathbf{x}-\mathbf{m}_1)^T\mathbf{V}_1^{-1}(\mathbf{x}-\mathbf{m}_1) \underset{D=0}{\overset{D=1}{\gtrless}} \mu \tag{4.13}$$

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 (4.13), it can be shown that the optimal detector in the Gaussian case is given by a second-order polynomial function. Hence, the decision boundaries[3] 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 4.6 Figure 4.6 shows the decision boundaries for the ML detector ($\eta = 1$ in (4.13)), 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).

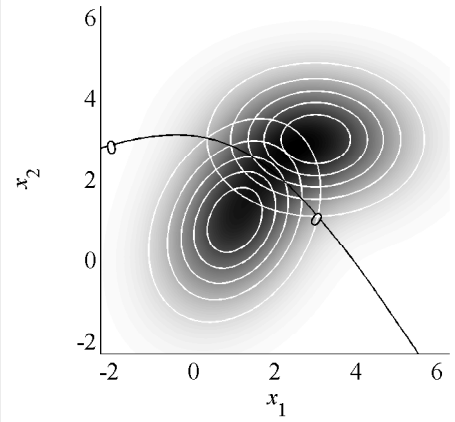


Fig. 4.6 Hyperbolic decision boundary of the ML detector and likelihoods for a Gaussian detection problem with 2D observations.

Example 4.7 Figure 4.7 shows an equivalent figure to that of the previous example, but for a problem with the following means and cross-covariance matrices:

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

[3] We obtain the decision boundaries for the equality in (4.13).

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.

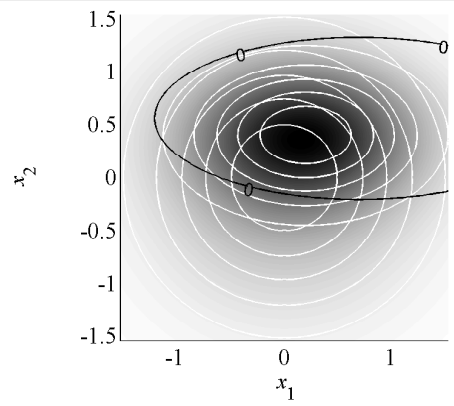


Fig. 4.7 Elliptic decision boundary of the ML detectors and likelihoods for a Gaussian detection problem with 2D observations.

4.5.1 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

$$(\mathbf{m}_1-\mathbf{m}_0)^T\mathbf{V}^{-1}\mathbf{x} \underset{D=0}{\overset{D=1}{\gtrless}} \tilde{\mu}, \tag{4.14}$$

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 (4.14) is a linear function of the observation vector $\mathbf{x}$.

Example 4.8 Figure 4.8 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.

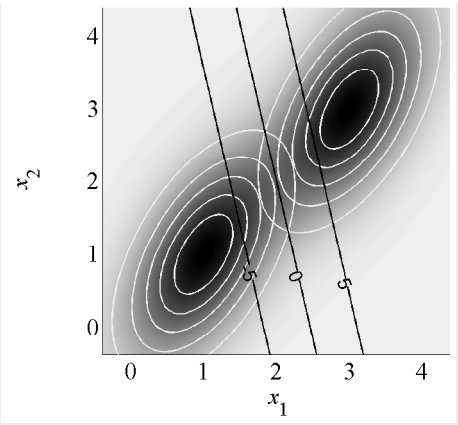


Fig. 4.8 Decision boundaries of the LLRT and likelihoods for a Gaussian detection problem with 2D observations and identical covariance matrices.

Example 4.9 (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

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

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.

4.5.2 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 4.10 Figure 4.9 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 $d = 1$.

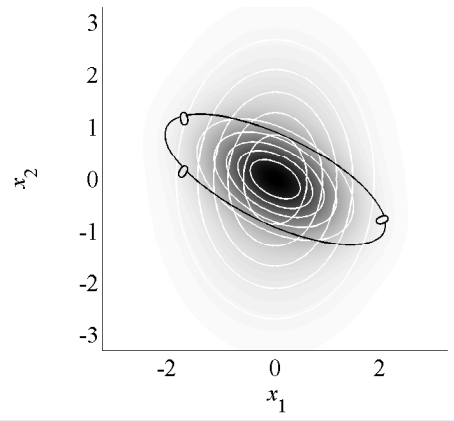


Fig. 4.9 Elliptic decision boundary for a 2D Gaussian problem with zero means.

Example 4.11 Figure 4.10 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.

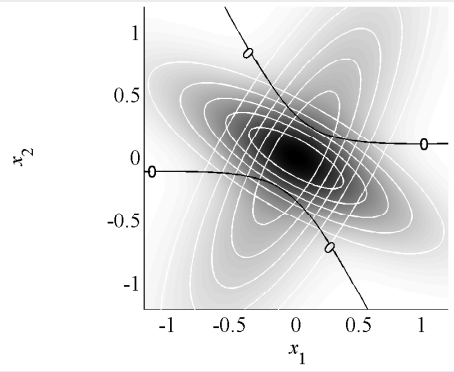


Fig. 4.10 Hyperbolic decision boundary for a 2D Gaussian problem with zero means.

4.6 Problems

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

$$p_{\mathbf{X}|H}(\mathbf{x}|0) = 2(1 - x_1) \tag{4.15}$$

$$p_{\mathbf{X}|H}(\mathbf{x}|1) = 2x_1 \tag{4.16}$$

$$p_{\mathbf{X}|H}(\mathbf{x}|2) = 2x_2 \tag{4.17}$$

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

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

$$p_{X|H}(x|0) = 2(1-x) \tag{4.18}$$

$$p_{X|H}(x|1) = 1 \tag{4.19}$$

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:

$$x \underset{D=0}{\overset{D=1}{\gtrless}} \eta \tag{4.20}$$

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.

4.3 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.

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

$$p_{X|H}(x|0) = \exp(-x); \tag{4.21}$$

$$p_{X|H}(x|1) = 2\exp(-2x); \tag{4.22}$$

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

4.5 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., $P_{\text{FA}} \leq \alpha = 0.1$).

4.6 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",

$$p_{X_1|H}(x_1|0) = \exp(-x_1)u(x_1) \tag{4.23}$$

$$p_{X_1|H}(x_1|1) = \lambda_1 \exp(-\lambda_1 x_1)u(x_1) \tag{4.24}$$

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

$$p_{X_2|H}(x_2|0) = \exp(-x_2)u(x_2) \tag{4.25}$$

$$p_{X_2|H}(x_2|1) = \lambda_2 \exp(-\lambda_2 x_2)u(x_2) \tag{4.26}$$

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.

Chapter 5
Sequential Detection

5.1 Some introductory examples

The previous chapter considered detection problems where we are given a set of observations and, based on them, we have to infer (decide) which hypothesis is true. Concretely, we have studied how to design the optimal detector (according to some criterion: maximum likelihood, minimum expected cost, etc.) and how to analyze its performance (false alarm and detection probabilities, error probability or average cost).

In this chapter, we will study a different approach to detection problems, where the observations arrive sequentially and, moreover, we can decide whether we want to acquire more observations to achieve the desired performance. These problems are referred to as sequential detection problems, where the objective is to take a decision as soon as possible (acquiring the smallest amount of observations), while ensuring the required performance.

In the following, we will study the problem of sequential detection in a simple set-up where there are only two hypotheses whose likelihoods are perfectly known, and the observations are independent and identically distributed (i.i.d.). But before we address the problem in a formal manner, this section presents two simple examples to introduce it. Concretely, we consider examples with and without gathering costs.

5.1.1 Example 1: Sequential detection with no gathering cost

Here, we present a simple example to motivate the problem of sequential detection. Consider an experiment in which the observation at time n under hypothesis $H = 0$ follows a zero-mean Gaussian distribution with variance $\sigma^2 = 1$, and under hypothesis $H = 1$, the observation follows a zero-mean Gaussian distribution with variance $\sigma^2 = 4$. That is, the likelihoods are

$$p_{X[n]|H}(x[n]|0) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2[n]}{2}\right), \tag{5.1}$$

and

$$p_{X[n]|H}(x[n]|1) = \frac{1}{\sqrt{8\pi}} \exp\left(-\frac{x^2[n]}{8}\right), \tag{5.2}$$

which are shown in Figure 5.1. Moreover, we assume that the observations are i.i.d.

We shall start by considering no observations, $n = 0$, and derive the detector with minimum average cost and its cost. Concretely, considering $c_{00} = c_{11} = 0$ and $c_{01} = c_{10} = 1$, the minimum expected cost detector optimizes the cost

$$\begin{aligned}
\overline{C}_0 &= \mathbb{E}\{c_{DH}\} \\
&= c_{10}P(D=1,H=0) + c_{01}P(D=0,H=1) + c_{00}P(D=0,H=0) + c_{11}P(D=1,H=1) \\
&= P(D=1,H=0) + P(D=0,H=1) \\
&= P(D=1|H=0)P_H(0) + P(D=0|H=1)P_H(1) \\
&= P(D=1|H=0)p + P(D=0|H=1)(1-p),
\end{aligned}$$

where we have defined, for the sake of notation, $P_H(0) = p$ and $P_H(1) = 1 - p$. The probabilities $P(D = 1|H = 0)$ and $P(D = 0|H = 1)$ are determined by the detector. However,

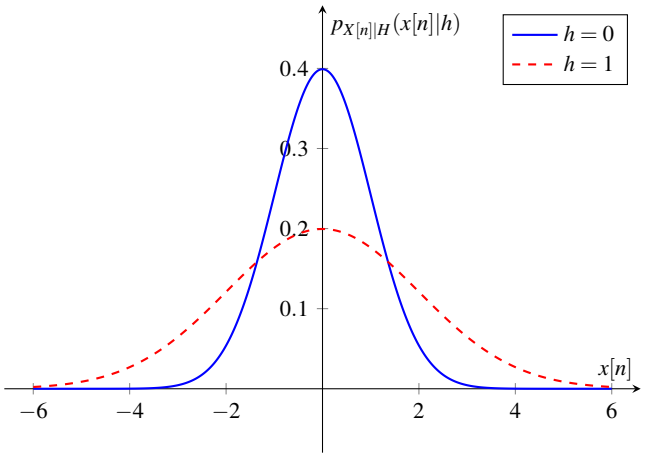


Fig. 5.1 Likelihoods considered in the introductory example

since there are no observations, the detector must always decide the same hypothesis, making one of the aforementioned probabilities one and the other zero. Then, for $0 \leq p \leq 1/2$, we should always decide $D=1$, which yields $P(D=1|H=0)=1$, $P(D=0|H=1)=0$, and

$$\overline{C}_0 = p.$$

If we had decided always $D=0$, we would have $P(D=1|H=0)=0$, $P(D=0|H=1)=1$, and

$$\overline{C}_0 = 1-p,$$

which is obviously larger than $\overline{C}_0 = p$ for $0 \leq p \leq 1/2$. Similarly, for $1/2 < p \leq 1$, we should always decide $d=0$, which implies that $P(D=1|H=0)=0$, $P(D=0|H=1)=1$, and

$$\overline{C}_0 = 1-p.$$

Combining both results, the minimum average cost for varying p is

$$\overline{C}_0(p) = \begin{cases} p, & 0 \leq p \leq 1/2, \\ 1-p, & 1/2 < p \leq 1, \end{cases}$$

where we have explicitly written the dependence of the minimum average cost with p. Figure 5.2 shows $\overline{C}_0(p)$ for varying p.

Let us now consider an arbitrary number of observations n, with $n>0$, and derive again the minimum expected cost detector, which minimizes the expected cost. This cost is denoted by $\overline{C}_n(p)$, to highlight that it depends on the number of observations and the prior probability p. Before proceeding, let us define $\mathbf{x}_n = (x[1],\ldots,x[n])^T$ as the vector that contains all available observations. Now, the sought detector is given by the likelihood ratio test (LRT), that is,

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

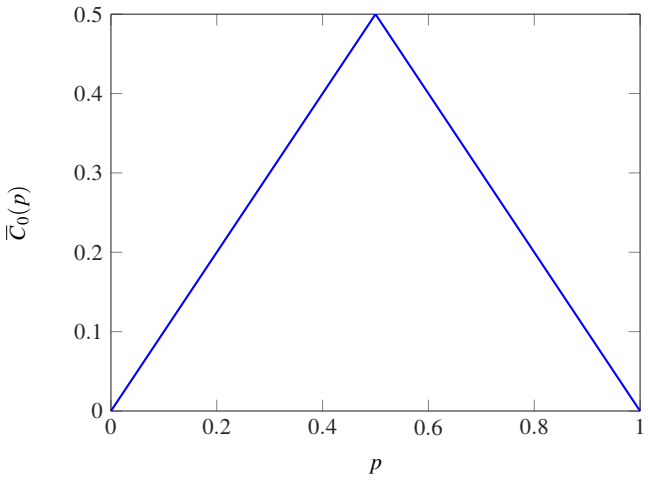


Fig. 5.2 Minimum average cost with no observations

To compute the likelihood of the n observations, we can use the i.i.d. assumption and, therefore,

$$p_{\mathbf{X}_n|H}(\mathbf{x}_n|h) = \prod_{i=1}^{n} p_{X[i]|H}(x[i]|h).$$

Using the likelihoods in (5.1) and (5.2), we get

$$p_{\mathbf{X}_n|H}(\mathbf{x}_n|0) = \prod_{i=1}^{n} p_{X[i]|H}(x[i]|0) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2[i]}{2}\right) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n} x^2[i]\right),$$

and

$$p_{\mathbf{X}_n|H}(\mathbf{x}_n|1) = \prod_{i=1}^{n} p_{X[i]|H}(x[i]|1) = \prod_{i=1}^{n} \frac{1}{\sqrt{8\pi}} \exp\left(-\frac{x^2[i]}{8}\right) = \frac{1}{(8\pi)^{n/2}} \exp\left(-\frac{1}{8}\sum_{i=1}^{n} x^2[i]\right).$$

Then, the LRT becomes

$$\frac{\frac{1}{(8\pi)^{n/2}} \exp\left(-\frac{1}{8}\sum_{i=1}^{n} x^2[i]\right)}{\frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n} x^2[i]\right)} = \frac{1}{2^n} \exp\left(\frac{3}{8}\sum_{i=1}^{n} x^2[i]\right) \underset{D=0}{\overset{D=1}{\gtrless}} \frac{p}{1-p},$$

and the log-likelihood ratio test (LLRT) is

$$\frac{1}{n}\sum_{i=1}^{n} x^2[i] \underset{D=0}{\overset{D=1}{\gtrless}} \frac{8}{3}\left[\frac{1}{n}\log\left(\frac{p}{1-p}\right) + \log 2\right].$$

Essentially, the LLRT compares the estimated variance with a threshold. The decision regions of the LLRT are

$$\mathscr{X}_1 = \left\{ \mathbf{x}_n \in \mathbb{R}^n \left| \frac{1}{n}\sum_{i=1}^{n} x^2[i] > \frac{8}{3}\left[\frac{1}{n}\log\left(\frac{p}{1-p}\right) + \log 2\right] \right. \right\},$$

and

$$\mathscr{X}_0 = \left\{ \mathbf{x}_n \in \mathbb{R}^n \left| \frac{1}{n}\sum_{i=1}^{n} x^2[i] \leq \frac{8}{3}\left[\frac{1}{n}\log\left(\frac{p}{1-p}\right) + \log 2\right] \right. \right\}.$$

For these decision regions, we can compute the minimum expected cost as

$$\begin{aligned}\overline{C}_n &= P(D=1|H=0)p + P(D=0|H=1)(1-p) \\ &= pP_{\text{FA}} + (1-p)P_{\text{M}},\end{aligned}$$

where

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

and

$$P_{\text{M}} = P(D=0|H=1) = \int_{\mathscr{X}_0} P_{\mathbf{X}_n|H}(\mathbf{x}_n|1)d\mathbf{x}_n.$$

Both, P_{FA} and P_{M}, are given by complicated multidimensional integrals with no closed-form solution and that are difficult to evaluate numerically. Hence, we must compute them using a different approach.

We shall start by considering a transformation of the random variables $y[i], i = 1, \ldots, I$, which are Gaussian distributed with zero mean, unit variance, and i.i.d. Concretely, the transformation is

$$Z = \sum_{i=1}^{I} y^2[i],$$

which is distributed as a Chi-squared random variable with I degrees of freedom, denoted as $Z \sim \chi^2_I$. The probability density function of Z is

$$p_Z(z) = \begin{cases} \frac{1}{2^{I/2}\Gamma(I/2)} z^{I/2-1}\exp(-z/2), & z > 0, \\ 0, & \text{otherwise}, \end{cases}$$

and its cumulative distribution function is

$$F_Z(z) = \frac{\gamma(I/2, z/2)}{\Gamma(I/2)},$$

where $\Gamma(\cdot)$ is the gamma function and $\gamma(\cdot,\cdot)$ is the lower incomplete gamma function.[1] Figure 5.3 depicts $p_Z(z)$ and $F_Z(z)$ for a different number of degrees of freedom.

Using the Chi-squared distribution, we can compute P_{FA} and P_{M}. The former is given by

[1] For positive integer values of the argument, the gamma function is given by $\Gamma(a) = (a-1)!$. To compute the incomplete gamma function, it is necessary to resort to (uni-dimensional) numerical integration.

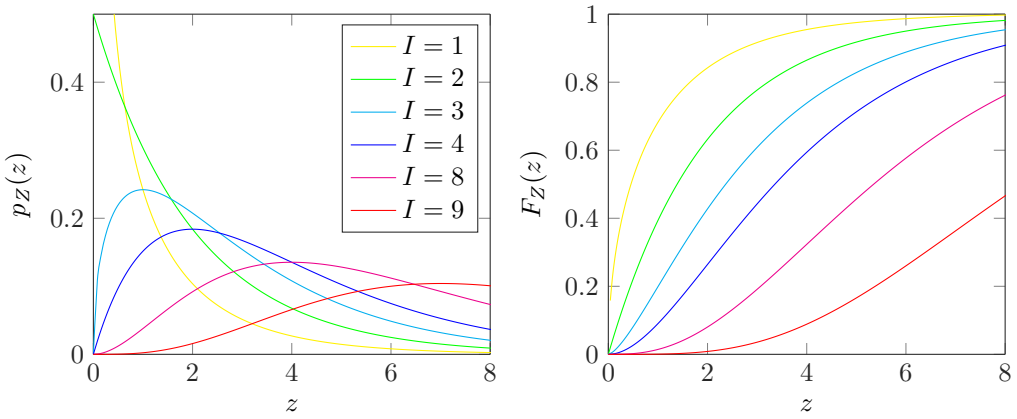


Fig. 5.3 Probability and cumulative density functions of a Chi-squared random variable with I degrees of freedom

$$\begin{aligned} P_{\text{FA}} &= P(D=1|H=0) \\ &= P\left(\frac{1}{n}\sum_{i=1}^{n} x^2[i] > \frac{8}{3}\left[\frac{1}{n}\log\left(\frac{p}{1-p}\right)+\log 2\right]\middle| H=0\right) \\ &= P\left(\sum_{i=1}^{n} x^2[i] > \frac{8}{3}\left[\log\left(\frac{p}{1-p}\right)+n\log 2\right]\middle| H=0\right), \end{aligned}$$

and taking into account that, under $H=0$, $x[i]$ are i.i.d. Gaussian variables with zero mean and unit variance, P_{FA} is the probability that a χ^2_n random variable is larger than $8/3\left[\log\left(p/(1-p)\right)+n\log 2\right]$. That is,

$$P_{\text{FA}} = 1 - \frac{\gamma\left(n/2, \frac{8}{3}\left[\log\left(\frac{p}{1-p}\right)+n\log 2\right]\right)}{\Gamma(n/2)}.$$

We can proceed similarly for P_{M} as follows

$$\begin{aligned} P_{\text{M}} &= P(D=0|H=1) \\ &= P\left(\frac{1}{n}\sum_{i=1}^{n} x^2[i] \le \frac{8}{3}\left[\frac{1}{n}\log\left(\frac{p}{1-p}\right)+\log 2\right]\middle| H=1\right) \\ &= P\left(\sum_{i=1}^{n}\left(\frac{x[i]}{2}\right)^2 \le \frac{2}{3}\left[\log\left(\frac{p}{1-p}\right)+n\log 2\right]\middle| H=1\right). \end{aligned}$$

Under $H=1$, $x[i]/2$ are i.i.d. Gaussian variables with zero mean and unit variance, and P_{M} is therefore the probability that a χ^2_n random variable is smaller than $2/3\left[\log\left(p/(1-p)\right)+n\log 2\right]$, which can be expressed as

$$P_{\text{M}} = \frac{\gamma\left(n/2, \frac{2}{3}\left[\log\left(\frac{p}{1-p}\right)+n\log 2\right]\right)}{\Gamma(n/2)}.$$

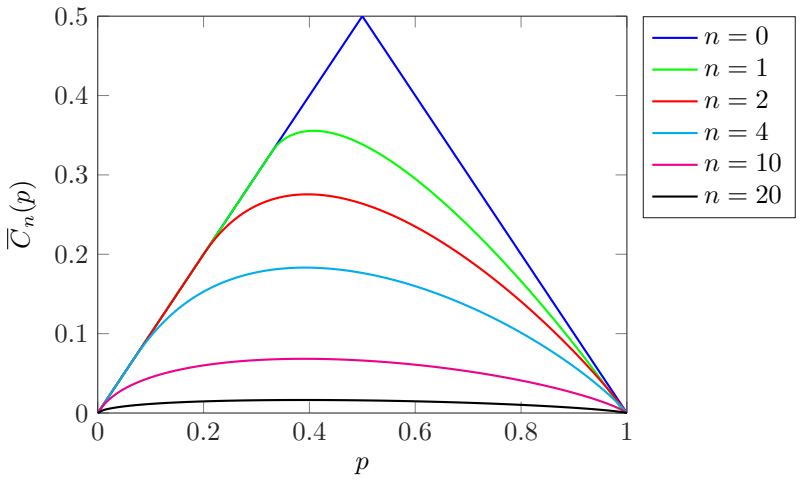


Fig. 5.4 Minimum average cost with n observations

Hence, the expected cost becomes

$$\overline{C}_n(p) = p\left(1 - \frac{\gamma\left(n/2, \frac{8}{3}\left[\log\left(\frac{p}{1-p}\right) + n\log 2\right]\right)}{\Gamma(n/2)}\right) + (1-p)\left(\frac{\gamma\left(n/2, \frac{2}{3}\left[\log\left(\frac{p}{1-p}\right) + n\log 2\right]\right)}{\Gamma(n/2)}\right), \quad (5.3)$$

Let us now point out that the second argument of $\gamma(\cdot,\cdot)$ is negative for

$$p < \frac{1}{2^n+1},$$

making the lower incomplete gamma function zero, which yields

$$\overline{C}_n(p) = p,$$

for $p < \frac{1}{2^n+1}$.

Figure 5.4 shows $\overline{C}_n(p)$ for some values of n, which shows that

$$\overline{C}_0(p) \geq \overline{C}_1(p) \geq \cdots \geq \overline{C}_n(p) \geq \cdots$$

Then, we should keep acquiring samples as long as we can (larger n) and, as a consequence, get a smaller minimum average cost.

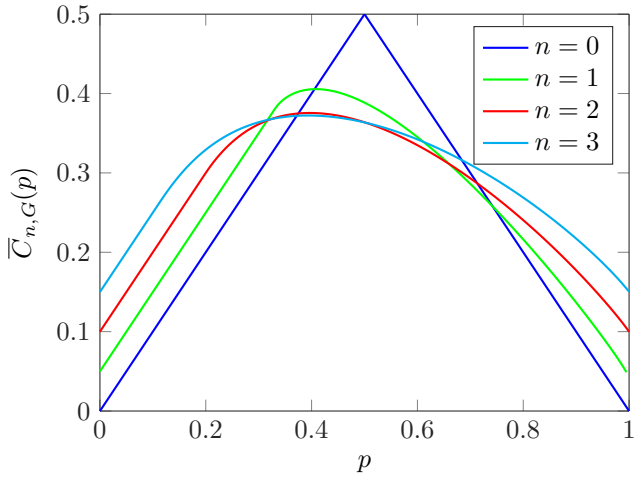


Fig. 5.5 Minimum average cost, including the gathering cost, with n observations

5.1.2 Example 2: Sequential detection with gathering cost

The results of the previous example do not make a lot of sense as we should keep collecting samples forever if we are to minimize the minimum average cost. Actually, for $n \to \infty$ we have $\overline{C}_n(p) \to 0$, regardless of p. Intuitively, we should include a cost every time an observation is collected, which should include a gathering cost related to, i.e., power consumption of the acquisition and transmission devices, and a waiting cost. Let us go back to the previous example and repeat it considering that this gathering (and waiting) cost is $c_G = 0.05$.

Taking into account c_G and $\overline{C}_n(p)$ derived in (5.3), the modified cost is given by

$$\overline{C}_{n,G}(p) = p\left(1 - \frac{\gamma\left(n/2, \frac{8}{3}\left[\log\left(\frac{p}{1-p}\right) + n\log 2\right]\right)}{\Gamma(n/2)}\right) + (1-p)\left(\frac{\gamma\left(n/2, \frac{2}{3}\left[\log\left(\frac{p}{1-p}\right) + n\log 2\right]\right)}{\Gamma(n/2)}\right) + c_G \cdot n,$$

which is depicted in Figure 5.5. From this figure, we can notice that keeping collecting samples does not necessarily improve $\overline{C}_{n,G}(p)$ as it happened in the case of no gathering cost, that is, $\overline{C}_{n,G}(p) \not\geq \overline{C}_{n+1,G}(p), \forall p \in [0,1]$. However, there is a range of values of p, for which it holds that $\overline{C}_{n,G}(p) \geq \overline{C}_{n+1,G}(p)$. Hence, sometimes it will be convenient to acquire an additional samples and sometimes it will not. Precisely this idea is the main ingredient of sequential detection.

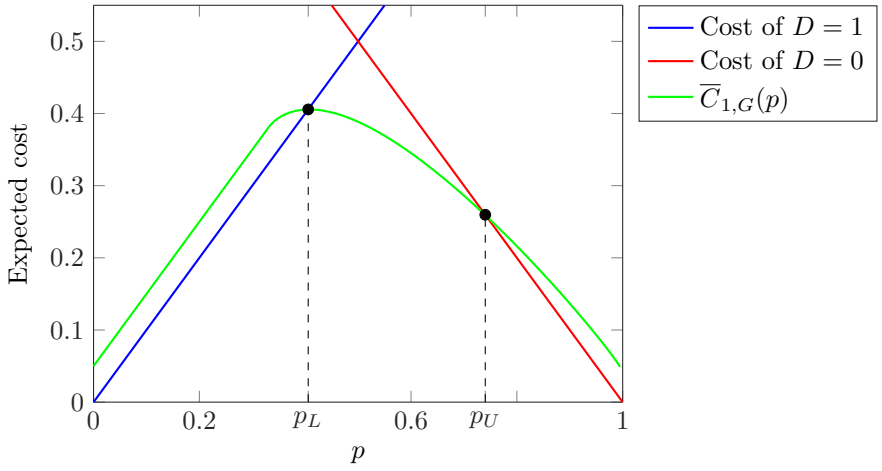


Fig. 5.6 Minimum average cost, including the gathering cost, with n observations

5.2 Sequential test

We have used the previous examples to motivate sequential detection, but such story is not completely accurate because to derive the minimum expected cost at time n, the detector needs to use n samples. However, we need to make the decision every time a new sample is acquired, which makes the story a bit simpler as we shall see.

Consider there are no observations available, that is $n = 0$. At this time instant, we need to decide between $H = 0$, $H = 1$, or take another sample. Since there are no available samples, the cost of always deciding $D = 1$ is p, the prior probability of $H = 0$, the cost of always deciding $D = 0$ is $1 - p$, the prior probability of $H = 1$, and the minimum expected cost with $n = 1$ sample is $\overline{C}_{1,G}(p)$. These three costs are depicted in Figure 5.6, which shows that they intersect at two points. The first of these two points, p_L, can be obtained as the largest value of p for which $\overline{C}_{1,G}(p)$ is still larger than the cost of always deciding $D = 1$. Mathematically, p_L is obtained as

$$p_L = \sup_p \{p \mid \overline{C}_{1,G}(p) > p\}.$$

Similarly, p_U is the smallest value of p where $\overline{C}_{1,G}(p)$ starts to be larger than the cost of always deciding $D = 0$, that is,

$$p_U = \inf_p \{p \mid \overline{C}_{1,G}(p) > 1 - p\}.$$

Hence, for $p \in [0, p_L]$, with $p_L = 0.4057$ in our example, the cost of always deciding $D = 1$ is the smallest, whereas for $p \in [p_U, 1]$, with $p_U = 0.7403$, the cost of always deciding $D = 0$ is the smallest. However, for $p \in (p_L, p_U)$, neither the cost of always deciding $D = 0$, nor the cost of always deciding $D = 1$ is the smallest, and we must take another sample.

Summarizing, at time $n = 0$, the sequential test must[2]

$$\begin{array}{l} \text{decide } D = 1 \text{ for } p \leq p_L, \\ \text{decide } D = 0 \text{ for } p \geq p_U, \\ \text{take another sample for } p_L < p < p_U. \end{array} \tag{5.4}$$

The question that remains to be answered is: What do we have to do if we have decided to take another sample? To answer this question, we must note that at $n = 1$, we have already taken the sample, i.e., the gathering cost has been already spent. Moreover, the possible decisions at $n = 1$ are exactly those of $n = 0$: decide between $H = 0$, $H = 1$, or take another sample. That is, the effect of having already taken one sample does not modify the test as there are still an infinite number of available samples. However, there is one important difference. The value of $x[1]$ provides some (partial) knowledge about the hypothesis. Then, conditioned on having observed $x[1]$, we should repeat the sequential test in (5.4), but instead of using the prior probability $P(H = 0) = p$, we must use the posterior probability $P(H = 0|X[1] = x[1]) = p_1$, i.e.,

$$\begin{array}{l} \text{decide } D = 1 \text{ for } p_1 \leq p_L, \\ \text{decide } D = 0 \text{ for } p_1 \geq p_U, \\ \text{take another sample for } p_L < p < p_U. \end{array}$$

Similarly, at a generic time n, the sequential test is

$$\begin{array}{l} \text{decide } D = 1 \text{ for } p_n \leq p_L, \\ \text{decide } D = 0 \text{ for } p_n \geq p_U, \\ \text{take another sample for } p_L < p < p_U, \end{array} \tag{5.5}$$

where the posterior probability is now

$$p_n = P(H = 0|X[1] = x[1], \ldots, X[n] = x[n]),$$

with $p_0 = P(H = 0) = p$. To conclude the derivation of the sequential test, we must find an explicit expression for p_n. First, using Bayes's theorem we can rewrite p_n as

$$\begin{aligned} p_n &= P(H = 0|X[1] = x[1], \ldots, X[n] = x[n]) \\ &= \frac{P(H = 0, X[1] = x[1], \ldots, X[n] = x[n])}{P(X[1] = x[1], \ldots, X[n] = x[n])} \\ &= \frac{P(X[1] = x[1], \ldots, X[n] = x[n]|H = 0)P_H(0)}{P(X[1] = x[1], \ldots, X[n] = x[n])}. \end{aligned}$$

Applying now the law of total probability to the denominator, p_n becomes

[2] Although this sequential test was obtained for a particular example, it is a general result since $\overline{C}_{1,G}(p)$ is a concave function of p in $[0,1]$ for any likelihood and any other costs c_{DH} and c_G. Nevertheless, the derived values of p_L and p_U would be different.

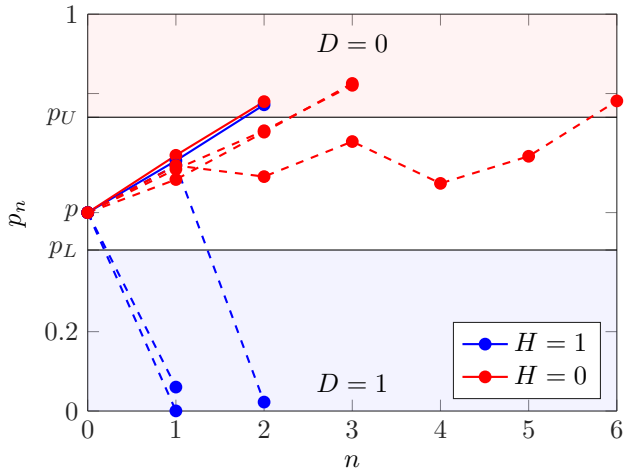


Fig. 5.7 Realizations of the sequential test

$$p_n = \frac{P(X[1]=x[1],\ldots,X[n]=x[n]|H=0)P(H=0)}{\sum_{h=0}^{1} P(X[1]=x[1],\ldots,X[n]=x[n]|H=h)P_H(h)}$$
$$= \frac{p_{\mathbf{X}_n|H}(\mathbf{x}_n|0)p}{p_{\mathbf{X}_n|H}(\mathbf{x}_n|0)p + p_{\mathbf{X}_n|H}(\mathbf{x}_n|1)(1-p)}, \tag{5.6}$$

which is a function of both likelihoods, $P_{\mathbf{X}_n|H}(\mathbf{x}_n|0)$ and $P_{\mathbf{X}_n|H}(\mathbf{x}_n|1)$, and p. Finally, taking into account the i.i.d. assumption, we can simplify (5.6) as

$$p_n = \frac{p}{p+(1-p)\dfrac{p_{\mathbf{X}_n|H}(\mathbf{x}_n|1)}{p_{\mathbf{X}_n|H}(\mathbf{x}_n|0)}} = \frac{p}{p+(1-p)\displaystyle\prod_{i=1}^{n}\frac{p_{X[i]|H}(x[i]|1)}{p_{X[i]|H}(x[i]|0)}}, \tag{5.7}$$

where, for the sake of consistency, we define $\prod_{i=1}^{0} = 1$.

Finally, and continuing with our example, where

$$\frac{p_{X[i]|H}(x[i]|1)}{p_{X[i]|H}(x[i]|0)} = \frac{1}{2}\exp\left(\frac{3}{8}x^2[i]\right),\ i=1,2,\ldots,$$

Figure 5.7 shows several realizations of the sequential test in (5.5) for this example, when $p=0.5$. Some of these realizations were obtained for $x[n]$ generated under $H=0$ and other realizations for $x[n]$ generated under $H=1$. In this figure, we can see that as soon as p_n is above p_U or below p_L, the detector stops the acquisition of more observations. In the former case, the decision is $D=0$, whereas in the latter, it is $D=1$.

5.3 Sequential probability ratio test

Using (5.7), the sequential test in (5.5) is

$$\begin{aligned} &\text{decide } D=1 \text{ for } \phi_n \geq \tfrac{p(1-p_L)}{p_L(1-p)}, \\ &\text{decide } D=0 \text{ for } \phi_n \leq \tfrac{p(1-p_U)}{p_U(1-p)}, \\ &\text{take another sample for } \tfrac{p(1-p_U)}{p_U(1-p)} < \phi_n < \tfrac{p(1-p_L)}{p_L(1-p)}, \end{aligned} \tag{5.8}$$

where

$$\phi_n = \prod_{i=1}^{n} \frac{p_{X[i]|H}(x[i]|1)}{p_{X[i]|H}(x[i]|0)}.$$

Then, the sequential test boils down to a likelihood ratio test, and it is therefore named the sequential probability ratio test (SPRT). Note that the SPRT can be computed recursively when a new observation comes in, i.e.,

$$\phi_n = \prod_{i=1}^{n} \frac{P_{X[i]|H}(x[i]|1)}{P_{X[i]|H}(x[i]|0)} = \left(\prod_{i=1}^{n-1} \frac{P_{X[i]|H}(x[i]|1)}{P_{X[i]|H}(x[i]|0)} \right) \frac{P_{X[n]|H}(x[n]|1)}{P_{X[n]|H}(x[n]|0)} = \phi_{n-1} \cdot \frac{P_{X[n]|H}(x[n]|1)}{P_{X[n]|H}(x[n]|0)},$$

with $\phi_0 = 1$.

Actually, (5.8) is just one example of the SPRT for a particular choice of the thresholds (those that optimize the expected cost). The most general SPRT is given by

$$\begin{aligned} &\text{decide } D=1 \text{ for } \phi_n \geq \eta_U, \\ &\text{decide } D=0 \text{ for } \phi_n \leq \eta_L, \\ &\text{take another sample for } \eta_L < \phi_n < \eta_U, \end{aligned} \tag{5.9}$$

where the thresholds should satisfy

$$0 < \eta_L < 1 < \eta_U < \infty.$$

Intuitively, for larger values of η_U, it is more unlikely to decide $D=1$. This implies that it will take longer to decide $D=1$, while at the same time, it will be more unlikely to decide $D=1$ when $H=0$. Similarly, for smaller values of η_L, it is more unlikely to decide $D=0$ and, therefore, it will take longer to decide $D=0$, while at the same time, it will be more unlikely to decide $D=0$ when $H=1$. These suggests that there are three metrics at play: the probability of false alarm ($P_{\text{FA}} = P(D=1|H=0)$), the probability of missing ($P_{\text{M}} = P(D=0|H=1)$), and the sample size N, which is defined as

$$N = \min_n \{n \mid \phi_n \geq \eta_U \text{ or } \phi_n \leq \eta_L\}.$$

Thus, the most established objective is to design the SPRT, i.e., the thresholds η_L and η_U in (5.9), that minimizes N while guaranteeing that $P_{\text{FA}} \leq \alpha$ and $P_{\text{M}} \leq \beta$, where α and β are the target values of probability of false alarm and probability of missing, respectively. In the following, we will derive P_{FA} and P_{M} as a function of the thresholds.

Let us start by the probability of false alarm, which is defined as

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

where $\mathbf{x}_\infty \in \mathbb{R}^\infty$ and

$$\mathscr{X}_1 = \{\mathbf{x}_\infty \in \mathbb{R}^\infty | \phi_N \geq \eta_U\}.$$

To continue, we must note that the set $\mathscr{X}_1$ can be decomposed as

$$\mathscr{X}_1 = \bigcup_{n=1}^{\infty} \mathscr{X}_{1,n},$$

with

$$\mathscr{X}_{1,n} = \{\mathbf{x}_\infty \in \mathbb{R}^\infty | N = n \text{ and } \phi_n \geq \eta_U\}.$$

That is, we decide $D=1$ when we decide $D=1$ at $n=1$, or at $n=2$, or at $n=3$, ... Moreover, taking into account that if we decide $D=1$ at n, we could not decide it at a different time instant m, with $n \neq m$, the sets $\mathscr{X}_{1,n}$ and $\mathscr{X}_{1,m}$ are mutually exclusive, i.e., they do not overlap, allowing us to write

$$\begin{aligned} P_{\text{FA}} &= \int_{\mathscr{X}_1} p_{\mathbf{X}_\infty|H}(\mathbf{x}_\infty|0)d\mathbf{x}_\infty = \int_{\cup_{n=1}^{\infty}\mathscr{X}_{1,n}} p_{\mathbf{X}_\infty|H}(\mathbf{x}_\infty|0)d\mathbf{x}_\infty = \sum_{n=1}^{\infty}\int_{\mathscr{X}_{1,n}} p_{\mathbf{X}_n|H}(\mathbf{x}_n|0)d\mathbf{x}_n \\ &= \sum_{n=1}^{\infty}\int_{\mathscr{X}_{1,n}} \prod_{i=1}^{n} p_{X[i]|H}(x[i]|0)d\mathbf{x}_n. \end{aligned}$$

Using now that

$$\phi_n \geq \eta_U \Rightarrow \prod_{i=1}^{n} p_{X[i]|H}(x[i]|0) \leq \eta_U^{-1} \prod_{i=1}^{n} p_{X[i]|H}(x[i]|1)$$

for $\mathbf{x}_n \in \mathscr{X}_{1,n}$, we have

$$P_{\text{FA}} \leq \eta_U^{-1} \sum_{n=1}^{\infty}\int_{\mathscr{X}_{1,n}} \prod_{i=1}^{n} p_{X[i]|H}(x[i]|1)d\mathbf{x}_n.$$

Since the probability of detection is

$$\begin{aligned} P_{\text{D}} &= \int_{\mathscr{X}_1} p_{\mathbf{X}_\infty|H}(\mathbf{x}_\infty|1)d\mathbf{x}_\infty = \sum_{n=1}^{\infty}\int_{\mathscr{X}_{1,n}} \prod_{i=1}^{n} p_{X[i]|H}(x[i]|1)d\mathbf{x}_n \\ &= 1 - P_{\text{M}}, \end{aligned}$$

we get

$$P_{\text{FA}} \leq \eta_U^{-1}(1 - P_{\text{M}}). \tag{5.10}$$

To compute the probability of missing, it is possible to follow a similar approach. Define

$$\mathscr{X}_0 = \{\mathbf{x}_\infty \in \mathbb{R}^\infty | \phi_N \leq \eta_L\} = \bigcup_{n=1}^{\infty} \mathscr{X}_{0,n},$$

where

$$\mathscr{X}_{0,n} = \{\mathbf{x}_\infty \in \mathbb{R}^\infty | N = n \text{ and } \phi_n \leq \eta_L\}.$$

Hence, the probability of missing is

$$P_{\mathrm{M}} = P(D=0|H=1) = \int_{\mathscr{X}_0} p_{\mathbf{X}_\infty|H}(\mathbf{x}_\infty|1) d\mathbf{x}_\infty = \sum_{n=1}^{\infty} \int_{\mathscr{X}_{0,n}} \prod_{i=1}^{n} p_{X[i]|H}(x[i]|1) d\mathbf{x}_n.$$

The above expression can be bounded as

$$P_{\mathrm{M}} \leq \eta_L \sum_{n=1}^{\infty} \int_{\mathscr{X}_{0,n}} \prod_{i=1}^{n} p_{X[i]|H}(x[i]|0) d\mathbf{x}_n,$$

since

$$\phi_n \leq \eta_L \Rightarrow \prod_{i=1}^{n} p_{X[i]|H}(x[i]|1) \leq \eta_L \prod_{i=1}^{n} p_{X[i]|H}(x[i]|0),$$

for $\mathbf{x}_n \in \mathscr{X}_{0,n}$. Finally, we get

$$P_{\mathrm{M}} \leq \eta_L (1 - P_{\mathrm{FA}}), \tag{5.11}$$

where we have taken into account that

$$\sum_{n=1}^{\infty} \int_{\mathscr{X}_{0,n}} \prod_{i=1}^{n} p_{X[i]|H}(x[i]|0) d\mathbf{x}_n = 1 - \sum_{n=1}^{\infty} \int_{\mathscr{X}_{1,n}} \prod_{i=1}^{n} p_{X[i]|H}(x[i]|0) d\mathbf{x}_n = 1 - P_{\mathrm{FA}}.$$

The bounds for the probabilities of false alarm and missing in (5.10) and (5.11) allow us to obtain the values η_L and η_U that achieve the desired value for P_{FA} and P_{M}. Concretely, we only need to solve the inequalities in (5.10) and (5.11), which yields

$$\eta_L \geq \frac{P_{\mathrm{M}}}{1 - P_{\mathrm{FA}}}, \qquad \eta_U \leq \frac{1 - P_{\mathrm{M}}}{P_{\mathrm{FA}}},$$

which are, in fact, only bounds. To avoid the inequalities and derive (approximate) equalities, we need to assume that when ϕ_N crosses the boundaries η_L and η_U, the excess over the boundaries is negligible. That is,

$$\phi_N - \eta_U \to \varepsilon_1, \qquad \eta_L - \phi_N \to \varepsilon_2,$$

with ε_i arbitrarily small positive constants. Under these assumptions, which are very accurate for large N, we get

$$\eta_L \approx \frac{P_{\mathrm{M}}}{1 - P_{\mathrm{FA}}}, \qquad \eta_U \approx \frac{1 - P_{\mathrm{M}}}{P_{\mathrm{FA}}},$$

which are known as Wald's approximations. Interestingly, and contrary to what happens with the LRT, the thresholds required to achieve the desired probabilities of false alarm and missing do not depend on the likelihoods. Nevertheless, the SPRT does depend on the likelihoods, and so do the sample size, N, and the expected sample size, $\mathbb{E}\{N\}$.

We conclude the topic of sequential detection by presenting the Wald-Wolfowitz theorem. Let us denote the probabilities of false alarm and missing of the SPRT $P_{\mathrm{FA}}(\phi)$ and $P_{\mathrm{M}}(\phi)$, and its sample size $N(\phi)$. Consider an alternative sequential decision rule with probabilities of false alarm and missing $P_{\mathrm{FA}}(\psi)$ and $P_{\mathrm{M}}(\psi)$, which satisfy

$$P_{\mathrm{FA}}(\psi) \leq P_{\mathrm{FA}}(\phi), \qquad P_{\mathrm{M}}(\psi) \leq P_{\mathrm{M}}(\phi).$$

Then, the Wald-Wolfowitz theorem states that

$$\mathbb{E}\{N(\psi)\} \geq \mathbb{E}\{N(\phi)\},$$

where $N(\psi)$ is the sample size of the alternative decision rule. Hence, for a desired level of performance ($P_{\mathrm{FA}} \leq \alpha$ and $P_{\mathrm{M}} \leq \beta$), there does not exist any sequential decision rule that achieves an expected sample size smaller than that of the SPRT. Interestingly, since a fixed-sample-size detector can be seen as a (very particular) sequential decision rule, the average sample size of the SPRT can not be larger than the sample size of any fixed-sample-size detector with the same level of performance. Alternatively, the Wald-Wolfowitz theorem also states that, for a fixed expected sample size, there is no sequential rule that achieves smaller P_{FA} and P_{M} than those of the SPRT.

Appendix A
Transformations of random variables

A.1 Change of Random Variable

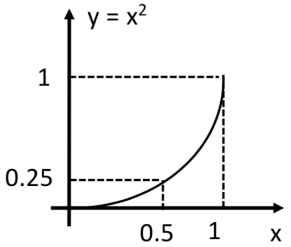
Let's consider we know the probability of a r.v. X, $p_X(x)$, and we now want to compute the probability density function of some variable $Y = f(X)$, that is, we need to calculate $p_Y(y)$.

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

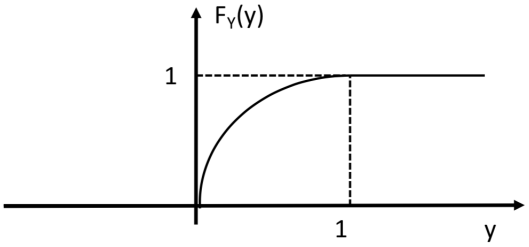
$$F_Y(y_0) = P\{Y < y_0\} = P\{X < \sqrt{y_0}\}$$

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

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

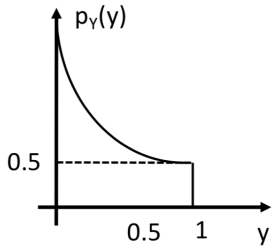
So, we have that

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



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:

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

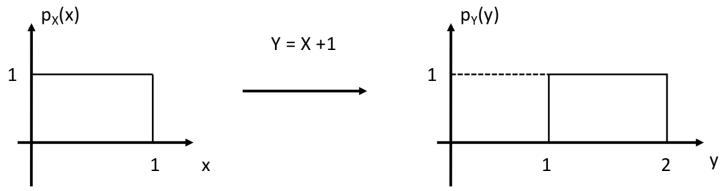
A.1.1 Some usual r.v. changes

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.