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

Estimation and Detection Theory

Year 2020-21

May 12, 2021

Universidad Carlos III de Madrid

Contents

1 Statistical Estimation Theory 1
1.1 Statistical Estimation Theory 2
1.1.1 General view of the estimation problem 2
1.1.2 Statistical information involved in estimation problems 3
1.1.3 Cost functions for estimation problems 3
1.2 Design of estimators 6
1.2.1 ML and MAP estimators 6
1.2.2 Bayesian design of estimators 8
1.3 Common bayesian estimators 10
1.3.1 Minimum Mean Squared Error estimator (MSE) 10
1.3.2 Minimum Mean Absolute Deviation Estimator (MAD) 11
1.4 Estimation with constrains 14
1.4.1 General principles 14
1.4.2 Linear estimation of minimum squared mean error 15
1.5 Estimation with gaussian distributions 20
1.5.1 One dimensional case 20
1.5.2 Case with multidimensional variables 23
1.5.3 Linear estimation and Gaussian estimation 24
1.6 ML estimation of probability distributions parameters 25
1.7 Problems 27

2 Linear Filtering 31
2.1 Introduction 31
2.2 The filtering problem 31
2.3 ML solution 32
2.4 Bayesian Solution 33
2.4.1 Probabilistic prediction of the filter output 34
2.5 Online calculus 34
2.5.1 Bayesian solution 35
2.5.2 ML solution 35
2.6 Wiener filter 36
2.7 Problems 36

3 Spectral Estimation . . . 39
3.1 Introduction . . . 40
3.2 Preliminaries: Spectral analysis of deterministic signals . . . 40
3.3 Non-parametric methods in spectral estimation . . . 44
3.3.1 The periodogram . . . 44
3.3.2 The Blackman-Tukey estimator . . . 48
3.3.3 Estimators based on the averaged periodogram . . . 49
3.4 Parametric methods in spectral estimation . . . 50
3.4.1 Rational models for parametric spectral estimation . . . 50
3.4.2 The auto-correlation function of ARMA processes . . . 51
3.4.3 AR processes . . . 54
3.4.4 MA processes . . . 56

4 Statistical Detection Theory . . . 59
4.1 Some introductory examples . . . 60
4.1.1 Example 1: Binary detection with no observations . . . 60
4.1.2 Example 2: Binary decision with observations . . . 63
4.1.3 Example 3: Working the solution from the likelihoods . . . 66
4.2 Introduction to Detection Theory . . . 69
4.2.1 Hypotheses-based problems . . . 69
4.2.2 Statistical information involved in detection problems . . . 70
4.3 Analysis of the detection performance . . . 71
4.3.1 Probability of error . . . 72
4.3.2 Average (or mean) cost . . . 73
4.4 Detector design . . . 75
4.4.1 Maximum likelihood and maximum *a posteriori* detectors . . . 75
4.4.2 Minimum expected cost detector . . . 78
4.4.3 The Gaussian case . . . 80

5 Sequential Detection . . . 87
5.1 Some introductory examples . . . 88
5.1.1 Example 1: Sequential detection with no gathering cost . . . 88
5.1.2 Example 2: Sequential detection with gathering cost . . . 94
5.2 Sequential test . . . 95
5.3 Sequential probability ratio test . . . 98

A Transformations of random variables . . . 103
A.1 Change of Random Variable . . . 103
A.1.1 Some usual r.v. changes . . . 106

Chapter 1
Statistical Estimation Theory

1.1 Statistical Estimation Theory

Once we have faced some of the main concepts involved in estimation problems, we are ready to formalize the problem for a general case.

1.1.1 General view of the estimation problem

The design of an estimator consists of constructing a real function that, from the value of certain observation variables, provides predictions about an objective variable (or vector).

For a general case, we will denote that random variable to be estimated as S which can take any real value. As indicated in Fig. 1.1, we assume also that we have access to an observation vector $\mathbf{x}$ that can be considered as the realization of a random variable $\mathbf{X}$ lying in observation space $\mathscr{X}$. Note also that for the estimation task of s or S from $\mathbf{X}$ to make sense, there must be some statistical relationship between them.

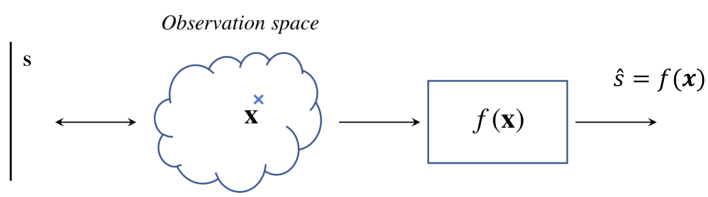


Fig. 1.1 Diagram block of estimation problems.

The estimation module implements a real output function, $\hat{S} = f(\mathbf{X})$, $f(\cdot)$ being the estimation function. It is common to refer to this function simply as *estimator*, and to its output as *estimation*. A fundamental characteristic of the estimator is the deterministic character of the $f(\cdot)$ function, that is, for a given value $\mathbf{x}$ the estimator will always provide the same output. Note that, even though $f(\cdot)$ is deterministic, its output can be modeled as a random variable if we consider the input to the function is random vector $\mathbf{X}$.

Since the estimator is expected to make a certain error in each application, a certain cost (or, alternatively, a profit) will be entailed. An optimum design of our estimator must take into account this cost during the design minimizing (or maximizing) its mean value.

We consider two different kinds of problems involving estimation problems:

- Analysis of estimators: Here, an estimator is given, and our purpose is to analyze its performance with respect to certain performance measure (cost function).
- Design of estimators: The goal is to build a function $f(\mathbf{x})$ to optimize a given objective.

1.1.2 Statistical information involved in estimation problems

Before approaching the design of the estimators themselves, we collect in this subsection the different probability functions that statistically characterize the existing relationship between observations and the variable to be estimated:

- First, the **likelihood** of the variable S is given by $p_{\mathbf{X}|S}(\mathbf{x}|s)$, and statistically characterizes the generation of observations for each specific value of the variable to be estimated.
- **Posterior distribution** of S: $p_{S|\mathbf{X}}(s|\mathbf{x})$. It indicates which S values are more or less likely to concentrate for each particular value in the observation vector.
- **Marginal or a priori** distribution of S: $p_S(s)$
- **Joint distribution** of $\mathbf{X}$ and S: $p_{\mathbf{X},S}(\mathbf{x},s) = p_{\mathbf{X}|S}(\mathbf{x}|s)p_S(s)$. It provides the most complete statistical modeling of the joint behavior of $\mathbf{X}$ and S.

It is important to note that the information available for estimator design may be different in each specific situation. A typical situation, because it is related to the physical process of generating the observations, is the one in which likelihood and the marginal distribution of S are available. Note that from them the calculation of the joint distribution is immediate and the posterior distribution $p_{S|\mathbf{X}}(s|\mathbf{x})$ can be calculated by means of Bayes' Theorem. Remember that Bayes' Theorem allows us to obtain the posterior distribution from the *a priori* distribution of S and its likelihood:

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

1.1.3 Cost functions for estimation problems

The evaluation and design of an estimator requires some objective criteria. In our case, we will consider that this criterion can materialize in the form of some function whose value we seek to maximize or minimize.

Given that the cost function is associated with a penalty whose origin is in the discrepancy between the actual and the estimated value of S, it is common to accept that $c(s,\hat{s}) \geq 0$, verifying equality when $s = \hat{s}$. Alternatively, a profit function can be defined whose average value is to be maximized. In addition, it is frequent that the cost function does not depend on the specific values of s and $\hat{s}$, but on the estimation error that is defined as the difference between the two, $e = s - \hat{s}$, in which case we have $c(s,\hat{s}) = c(s-\hat{s}) = c(e)$.

As an example, some frequently used cost functions are:

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

Accepting that this function[1] is $c(S,\hat{S})$, the evaluation of an estimator is carried out by evaluating the mean value of this cost and the estimator design criterion usually involves the minimization of its mean value; i.e, this cost is used in a statistical sense, evaluating/minimizing its mean value, which is equivalent to evaluating/minimizing the average cost that would be obtained by performing an infinitely large number of experiments.

In general, the mean cost of an estimator is given by

$$\mathbb{E}\{c(S,\hat{S})\} = \int_{\mathbf{x}}\int_{s} c(s,\hat{s})p_{S,\mathbf{X}}(s,\mathbf{x})dsd\mathbf{x} \tag{1.2}$$

where it should be noted that $\hat{s}$ is generally a function of $\mathbf{x}$.

Example 1.1 (Evaluation of estimators 1)

Suppose that the joint distribution of S and X is given by

$$p_{S,X}(s,x) = \begin{bmatrix} \frac{1}{x}, & 0 < s < x \text{ and } 0 < x < 1 \\ 0, & \text{otherwise} \end{bmatrix} \tag{1.3}$$

Let's consider two 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,

$$\begin{aligned} \mathbb{E}\{(S-wX)^2\} &= \int_0^1\int_0^x (s-wx)^2 p_{S,X}(s,x)dsdx \\ &= \int_0^1\int_0^x (s-wx)^2\frac{1}{x}dsdx \\ &= \int_0^1 \left(\frac{1}{3}-w+w^2\right)x^2dx \\ &= \frac{1}{3}\left(\frac{1}{3}-w+w^2\right) \end{aligned} \tag{1.4}$$

Taking $w = 1/2$ results in

$$\mathbb{E}\{(S-\hat{S}_1)^2\} = \mathbb{E}\{(S-\frac{1}{2}X)^2\} = \frac{1}{3}\left(\frac{1}{3}-\frac{1}{2}+\frac{1}{4}\right) = \frac{1}{36} \tag{1.5}$$

Alternatively, by taking $w = 1$ we get

$$\mathbb{E}\{(S-\hat{S}_2)^2\} = \mathbb{E}\{(S-X)^2\} = \frac{1}{3}\left(\frac{1}{3}-1+1\right) = \frac{1}{9} \tag{1.6}$$

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

[1] Note that the cost function is denoted with a c minuscule 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., $C = c(S,\hat{S})$.

Example 1.2 (Evaluation of estimators 2) Consider that X is a noisy observation of S, so that

$$X = S + R \tag{1.7}$$

where S is a random variable of mean 0 and variance 1, and R is a random Gaussian variable, independent of S, of mean 0 variance v. Considering the estimator $\hat{S} = X$, obtain the associated mean quadratic cost and mean absolute error.

The mean quadratic cost is given by:

$$\mathbb{E}\{(S-\hat{S})^2\} = \mathbb{E}\{(S-X)^2\} = \mathbb{E}\{R^2\} = v \tag{1.8}$$

And the mean absolute error

$$\begin{aligned}\mathbb{E}\{|S-\hat{S}|\} = \mathbb{E}\{|R|\} &= \int_{-\infty}^{\infty} |r| \frac{1}{\sqrt{2\pi v}} \exp\left(-\frac{r^2}{2v}\right) dr \\ &= 2\int_{0}^{\infty} r \frac{1}{\sqrt{2\pi v}} \exp\left(-\frac{r^2}{2v}\right) dr = \sqrt{\frac{2v}{\pi}}\end{aligned} \tag{1.9}$$

1.2 Design of estimators

1.2.1 ML and MAP estimators

We define the maximum likelihood estimator (ML) as

$$\hat{s}_{\mathrm{ML}} = \arg\max_s p_{\mathbf{X}|S}(\mathbf{x}|s) = \arg\max_s \ln(p_{\mathbf{X}|S}(\mathbf{x}|s)) \tag{1.10}$$

It is important to emphasize that the maximization of $p_{\mathbf{X}|S}(\mathbf{x}|s)$ has to be done with respect to the value of s, which is not the variable for which this probability function is defined.

On the other hand, we define the maximum a posterior estimator (MAP) as

$$\hat{s}_{\mathrm{MAP}} = \arg\max_s p_{S|\mathbf{X}}(s|\mathbf{x}) = \arg\max_s \ln(p_{S|\mathbf{X}}(s|\mathbf{x})) \tag{1.11}$$

in this case, maximization is performed on the same variable of the distribution that is being maximized.

Note that both (1.10) and (1.11) alternatively include the use of the logarithm function. This is done by practical reasons, since for the maximization of either the likelihood or the posterior of S it may be useful to introduce an auxiliary function that simplifies the analytical form of the function and, since the logarithm function is defined for every positive value of its argument and is strictly increasing, it implies that if $p_{S|\mathbf{X}}(s_1|\mathbf{x}) > p_{S|\mathbf{X}}(s_2|\mathbf{x})$, then also $\ln p_{S|\mathbf{X}}(s_1|\mathbf{x}) > \ln p_{S|\mathbf{X}}(s_2|\mathbf{x})$). So, the introduction of the logarithm function will be useful when the likelihood or the a posterior present products or exponentials, as it will transform products into sums and it will cancel exponentials. In this way, the maximization process can be simplified considerably.

If we compare both estimators, the ML estimator uses as statistical the likelihood of S (a distribution which models the generation of the observations), whereas the MAP estimator uses the posterior distribution of S (characterizes the behaviour of S for any observed value), so the MAP estimator has a more complete information of the variable to be estimated. Nevertheless, the ML estimation does not require the definition of probability densities on the variable to be estimated (a prior or posterior distribution of the S are not needed). Therefore, the use of the ML estimator is often used (or it is more appropriated) when such information is not available.

The ML estimator coincides with the MAP when S has a uniform distribution in a range of values and, therefore, the application of the ML estimator in the absence of information about the a prior distribution of S is equivalent to assuming uniformity in S and applying the MAP estimator. To check this equivalence, one need only consider the relationship between the likelihood and the posterior distribution of S by means of the Bayes Theorem,

$$\hat{s}_{\mathrm{MAP}} = \arg\max_s p_{S|\mathbf{X}}(s|\mathbf{x}) = \arg\max_s \frac{p_{\mathbf{X}|S}(\mathbf{x}|s)p_S(s)}{p_{\mathbf{X}}(\mathbf{x})}$$

Since $p_{\mathbf{X}}(\mathbf{x})$ does not depend on s and we are assuming that $p_S(s)$ is constant, we get:

$$\hat{s}_{\mathrm{MAP}} = \arg\max_s p_{\mathbf{X}|S}(\mathbf{x}|s) = \hat{s}_{\mathrm{ML}}$$

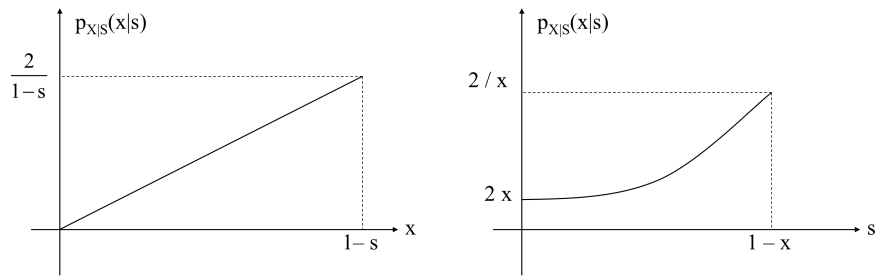


Fig. 1.2 Representation of the likelihood distribution of the exercise 1.3 as a function of x and s.

That is, the value of s that maximizes the posterior has to coincide with the one that maximizes likelihood.

Example 1.3 (Estimation ML)

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

$$p_{X|S}(x|s) = \frac{2x}{(1-s)^2}, \;\; 0 < x < 1-s, \;\; 0 < s < 1 \tag{1.12}$$

Given the available statistical information, it is decided to construct the ML estimator of S. For this purpose, the previous likelihood should be maximized. Such likelihood is a probability density function of X as represented in Figure 1.2(a), where it is verified that the integral of this function with respect to x is unitary. However, to carry out the maximization that allows to find $\hat{s}_{ML}$ it is more useful to represent this likelihood as a function of s (Fig.1.2(b))[2]. From this graphic representation it is evident 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,

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

Example 1.4 (Estimation MAP) Considering that

$$p(s|x) = \frac{1}{x^2} s \exp\left(-\frac{s}{x}\right), \qquad x \geq 0, s \geq 0 \tag{1.13}$$

The MAP estimator can be computed maximizing

[2] 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.

$$\ln(p(s|x)) = -2\ln(x) + \ln(s) - \frac{s}{x}, \qquad x \geq 0, s \geq 0 \tag{1.14}$$

Since $l(p(s|x))$ 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

$$\frac{\partial}{\partial s} \ln p(s|x)\bigg|_{s=\hat{s}_{\text{MAP}}} = \frac{1}{\hat{s}_{\text{MAP}}} - \frac{1}{x} = 0, \qquad x \geq 0, s \geq 0 \tag{1.15}$$

Thus,

$$\hat{s}_{\text{MAP}} = x \tag{1.16}$$

1.2.2 Bayesian design of estimators

It is worth asking, for a given cost and distribution, which is the best possible estimator. We can find this out by taking into account that, generally speaking, the average cost can be expressed as

$$\begin{aligned}\mathbb{E}\{c(S,\hat{S})\} &= \int_{\mathbf{x}} \int_{s} c(s,\hat{s}) p_{S|\mathbf{X}}(s|\mathbf{x}) ds \; p_{\mathbf{X}}(\mathbf{x}) \mathbf{dx} = \\ &= \int_{\mathbf{x}} \mathbb{E}\{c(S,\hat{s})|\mathbf{X}=\mathbf{x}\} p_{\mathbf{X}}(\mathbf{x}) \mathbf{dx}.\end{aligned} \tag{1.17}$$

The last line of this equation shows that a strategy for minimizing the overall estimation cost consists of minimizing the mean cost for each possible value of the observation vector, $\mathbb{E}\{c(S,\hat{s})|\mathbf{X}=\mathbf{x}\}$, which we will refer to as mean posterior cost or mean cost given $\mathbf{X}$. Therefore, both strategies (minimization of the expected cost for all S and $\mathbf{X}$, or conditioned to the value of $\mathbf{X}$) are in principle equivalent in order to obtain the optimal estimator associated with a given cost function.

The Bayesian Estimator associated with a cost function is defined as that which minimizes (1.17), that is:

$$\hat{s}^* = \arg\min_{\hat{s}} \; \mathbb{E}\{c(S,\hat{s})|\mathbf{X}=\mathbf{x}\} \tag{1.18}$$

where $\hat{s}^*$ is the Bayesian Estimator. According to our previous discussion, the Bayesian Estimator also minimizes the expected cost in a global sense, i.e., for all S and $\mathbf{X}$. Note, however, that for your design the expression (1.18) is more useful than the direct minimization of the overall cost.

$$\mathbb{E}\{c(S,\hat{S})\} = \int_{\mathbf{x}} \mathbb{E}\{c(S,\hat{s})|\mathbf{X}=\mathbf{x}\} p_{\mathbf{X}}(\mathbf{x}) \mathbf{dx} \tag{1.19}$$

since calculating the integral in $\mathbf{x}$ would require knowing beforehand the relationship between $\hat{s}$ and $\mathbf{x}$, which is precisely the objective of the estimator design problem.

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

$$p_{S|X}(s|x) = \frac{p_{S,X}(s,x)}{p_X(x)}. \tag{1.20}$$

Knowing that

$$p_X(x) = \int_0^1 p_{S,X}(s,x)ds = \int_0^x \frac{1}{x}ds = 1, \tag{1.21}$$

we obtain

$$p_{S|X}(s|x) = \left[\begin{array}{ll} \frac{1}{x}, & 0 < s < x < 1 \\ 0, & \text{otherwise} \end{array}\right. \tag{1.22}$$

The mean cost given the observation will be given by

$$\begin{aligned} \mathbb{E}\{c(S,\hat{s})|X = x\} &= \mathbb{E}\{(S-\hat{s})^2|X = x\} \\ &= \int_0^1 (s-\hat{s})^2 p_{S|X}(s|x)ds \\ &= \frac{1}{x}\int_0^x (s-\hat{s})^2 ds \\ &= \frac{1}{x}\left(\frac{(x-\hat{s})^3}{3} + \frac{\hat{s}^3}{3}\right) \\ &= \frac{1}{3}x^2 - \hat{s}x + \hat{s}^2. \end{aligned} \tag{1.23}$$

As a function of $\hat{s}$, the mean cost conditioned to the observation is a polynomial of second degree, whose minimum can be calculated immediately by derivation. Being

$$\frac{d}{d\hat{s}}\mathbb{E}\{c(S,\hat{s})|X = x\} = -x + 2\hat{s}, \tag{1.24}$$

the lowest mean quadratic error estimator will be

$$\hat{s}^* = \frac{1}{2}x, \tag{1.25}$$

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 (1.18) 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 for the design of the Bayesian Optimal Estimator. As mentioned above, this distribution is often calculated from the likelihood of S and its a prior 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 estimator of minimum mean squared error (MSE) is the one associated with the cost function $c(e) = e^2 = (s - \hat{s})^2$, and therefore is characterized by

$$\hat{s}_{\text{MSE}} = \arg\min_{\hat{s}} \mathbb{E}\{c(S, \hat{s}) | \mathbf{X} = \mathbf{x}\} = \tag{1.26}$$

$$= \arg\min_{\hat{s}} \int_s (s - \hat{s})^2 p_{S|\mathbf{X}}(s|\mathbf{x}) ds \tag{1.27}$$

Figure 1.3 illustrates the design problem with the minimum mean squared error estimator. The mean posterior cost can be obtained by integrating in s the function resulting from the product of the cost function and the posterior probability density of S. The argument for minimization is $\hat{s}$, which allows to move the graph corresponding to the cost function (represented with discontinuous stroke) so that the result of that integral is minimal.

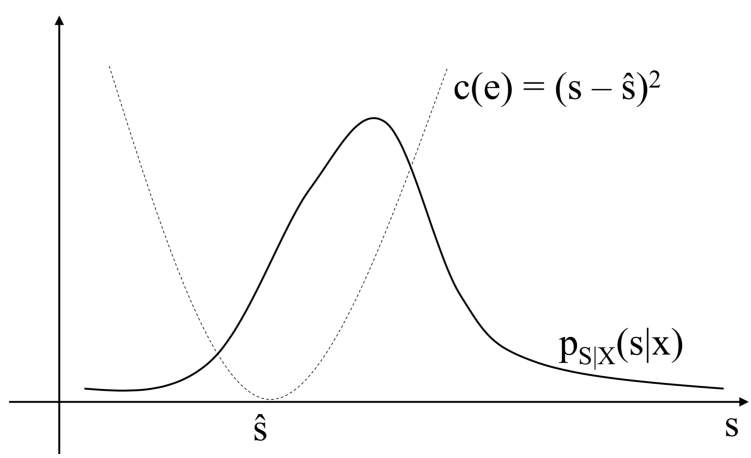


Fig. 1.3 Graphical representation of the process of calculating the posterior mean for a generic value $\hat{s}$.

The value of $\hat{s}_{\text{MSE}}$ can be analytically obtained by taking the derivative of the posterior mean cost and equaling the result to 0. The calculation of the derivative does not pose any difficulty since the derivative and the integral can be commutated (it is integrated with respect to s and is derived with respect to $\hat{s}$):

$$\left.\frac{d\mathbb{E}\{(S-\hat{s})^2|\mathbf{X}=\mathbf{x}\}}{d\hat{s}}\right|_{\hat{s}=\hat{s}_{\text{MSE}}} = -2\int_s (s-\hat{s}_{\text{MSE}})p_{S|\mathbf{X}}(s|\mathbf{x})ds = 0 \tag{1.28}$$

Bearing in mind that the integral in (1.28) should be cancelled, and using the fact that $\int p_{S|\mathbf{X}}(s|\mathbf{x})ds = 1$, it is easy to demonstrate that the minimum mean squared error estimator of S is given by

$$\boxed{\hat{s}_{\text{MSE}} = \int s\; p_{S|\mathbf{X}}(s|x)ds = \mathbb{E}\{S|\mathbf{X}=\mathbf{x}\} \qquad (1.29)}$$

In other words, the minimum mean squared error estimator of S is the mean of S given $\mathbf{X}$ or the posterior mean of S , i.e., the expected value of $p_{S|\mathbf{X}}(s|\mathbf{x})$.

Example 1.6 (Straightforward calculation of the MSE estimator) According to (1.29), minimum mean squared error estimator obtained in 1.1 can alternatively be obtained as follows

$$\hat{s}_{\text{MSE}} = \int_0^1 s p_{S|X}(s|x)ds = \int_0^x \frac{s}{x}ds = \frac{1}{2}x \tag{1.30}$$

which coincides with (1.25).

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

$$\begin{aligned}\hat{s}_{\text{MAD}} &= \arg\min_{\hat{s}}\; \mathbb{E}\{|S-\hat{s}|\;|\mathbf{X}=\mathbf{x}\} = \\ &= \arg\min_{\hat{s}} \int_s |s-\hat{s}|\; p_{S|\mathbf{X}}(s|\mathbf{x})ds\end{aligned} \tag{1.31}$$

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:

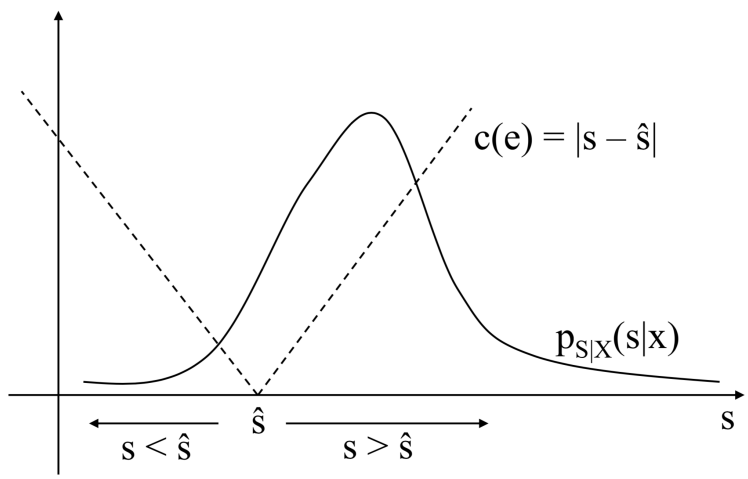


Fig. 1.4 Graphical representation of the process of calculating the posterior mean absolute error for a generic value $\hat{s}$.

$$\begin{aligned}\mathbb{E}\{|S-\hat{s}|\ |\mathbf{X}=\mathbf{x}\} &= \int_{-\infty}^{\hat{s}} (\hat{s}-s)\ p_{S|\mathbf{X}}(s|\mathbf{x})ds + \int_{\hat{s}}^{\infty} (s-\hat{s})\ p_{S|\mathbf{X}}(s|\mathbf{x})ds \\ &= \hat{s}\left[\int_{-\infty}^{\hat{s}} p_{S|\mathbf{X}}(s|\mathbf{x})ds - \int_{\hat{s}}^{\infty} p_{S|\mathbf{X}}(s|\mathbf{x})ds\right] + \\ &\quad + \int_{\hat{s}}^{\infty} s\ p_{S|\mathbf{X}}(s|\mathbf{x})ds - \int_{-\infty}^{\hat{s}} s\ p_{S|\mathbf{X}}(s|\mathbf{x})ds\end{aligned} \tag{1.32}$$

The fundamental theorem of calculus[3] allows us to obtain the derivative of the posterior mean cost as

$$\frac{d\mathbb{E}\{|S-\hat{s}|\ |\mathbf{X}=\mathbf{x}\}}{d\hat{s}} = 2F_{S|\mathbf{X}}(\hat{s}|\mathbf{x}) - 1 \tag{1.33}$$

where $F_{S|\mathbf{X}}(s|\mathbf{x})$ is the posterior distribution function of S given $\mathbf{X}$. Since $\hat{s}_{\text{MAD}}$ represents the minimum of the mean cost, the previous derivative must be cancelled for the estimator, verifying that $F_{S|\mathbf{X}}(\hat{s}_{\text{MAD}}|\mathbf{x}) = 1/2$. In other words, the absolute minimum error estimator is given by the median of $p_{S|\mathbf{X}}(s|\mathbf{x})$:

$$\boxed{\hat{s}_{\text{MAD}} = \text{median}\{S|\mathbf{X}=\mathbf{x}\}} \tag{1.34}$$

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 mean absolute error estimator will verify that

$$P\{S > \hat{s}_{\text{MAD}}\} = P\{S < \hat{s}_{\text{MAD}}\}$$

Example 1.7 (Design of a Minimum Mean Absolute Deviation Estimator)

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

In the scenario of the example 1.1, the a posterior distribution of S given X is uniform between 0 and x, the median of which is $x/2$. So,

$$\hat{s}_{\text{MAD}} = \frac{1}{2}x \tag{1.35}$$

Note that, in this case, the MAD estimator matches the MSE obtained at (1.25). This is a consequence of the symmetry of the a posterior distribution. In general, both estimators do not have to coincide.

1.4 Estimation with constrains

1.4.1 General principles

Sometimes it may be useful to impose a certain parametric shape on the estimator, $\hat{S} = f_{\mathbf{w}}(\mathbf{X})$, where $\mathbf{w}$ is a vector containing all the parameters of the function. For example, in a case with two observations $\mathbf{X} = [X_1, X_2]^T$, it might be a design requirement to restrict the estimator search to the family of quadratic estimators of the form $\hat{S} = w_0 + w_1 X_1^2 + w_2 X_2^2$. In these cases, the estimator design task is to find the optimal parameter vector $\mathbf{w}^*$ which provides a minimum average cost subject to the constraint imposed in the estimator architecture:

$$\begin{aligned}\mathbf{w}^* &= \arg\min_{\mathbf{w}} \mathbb{E}\{c(S,\hat{S})\} = \arg\min_{\mathbf{w}} \mathbb{E}\{c(S, f_{\mathbf{w}}(\mathbf{X}))\} \\ &= \arg\min_{\mathbf{w}} \int_{\mathbf{x}} \int_{s} c(s, f_{\mathbf{w}}(\mathbf{x})) p_{S,\mathbf{X}}(s,\mathbf{x}) ds d\mathbf{x} \end{aligned} \tag{1.36}$$

It can easily be understood that the imposition of constraints on the analytical form of the estimator results in incurring a higher average cost than would be obtained using the Bayesian estimator associated with the same cost function[4]. However, there may be practical reasons that make the use of the former preferable, for example for simplicity in the design or application of the estimator. An example of this can be found in the Section 1.4.2, dedicated to the study of linear estimators with minimum mean squared error.

Example 1.8 (Calculating an Estimator with Constrains)

Continuing the example 1.5, we want to calculate the minimum quadratic mean error estimator that has the form $\hat{s} = wx^2$. Starting from the mean cost given the observation calculated in (1.23), the expression of the global average cost can be obtained as

$$\begin{aligned}\mathbb{E}\{c(S,\hat{S})\} &= \int_{\mathbf{x}} \mathbb{E}\{c(S,\hat{s})|X = x\}\, p_{\mathbf{X}}(\mathbf{x}) \mathbf{dx} \\ &= \int_{\mathbf{x}} \left(\frac{1}{3}x^2 - \hat{s}x + \hat{s}^2\right) p_{\mathbf{X}}(\mathbf{x}) \mathbf{dx}\end{aligned} \tag{1.37}$$

Forcing $\hat{s} = wx^2$ and taking into account that $p_{\mathbf{X}}(\mathbf{x}) = 1$ for $0 < x < 1$, you get the global average cost as a function of w.

$$\mathbb{E}\{c(S, w\mathbf{X}^2)\} = \int_{\mathbf{x}} \left(\frac{1}{3}x^2 - wx^3 + w^2x^4\right) d\mathbf{x} \tag{1.38}$$

$$= \frac{1}{9} - \frac{1}{4}w + \frac{1}{5}w^2 \tag{1.39}$$

[4] The only exception to this rule is precisely the case where the constraints imposed allow the optimal estimator to be obtained or, in other words, when the Bayesian estimator presents an analytical form compatible with the constraints imposed.

The w^* value that optimizes (1.39) can be calculated by deriving respect to w and zeroing the expression obtained:

$$\frac{d}{d\hat{w}}\mathbb{E}\{c(S, w\mathbf{X}^2)\}\bigg|_{w=w^*} = -\frac{1}{4} + \frac{2}{5}w^* = 0, \tag{1.40}$$

$$w^* = \frac{5}{8}, \tag{1.41}$$

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

1.4.2 Linear estimation of minimum squared mean error

In this section we will focus on the study of random variable estimators that obtain their output as a linear combination of the values of the observations, using the minimization of the mean squared error as design criterion. Therefore, we will exclusively consider estimators that calculate their output as

$$\hat{S} = w_0 + w_1 X_1 + \cdots + w_N X_N \tag{1.42}$$

where N denotes the number of available observable variables, $\{X_i\}_{i=1}^N$, and $\{w_i\}_{i=0}^N$ are the weights that characterize the estimator. In this context, it is common to refer to the term independent of the above expression, w_0, as a bias term. For analytical simplicity, it is more convenient to enter the following matrix notation:

$$\hat{S} = w_0 + \mathbf{w}^T\mathbf{X} = \mathbf{w}_\mathrm{e}^T\mathbf{X}_\mathrm{e} \tag{1.43}$$

where $\mathbf{w} = [w_1, \ldots, w_N]^T$ and $\mathbf{X} = [X_1, \ldots, X_N]^T$ are the (column) vectors of parameters and observations, respectively, and $\mathbf{w}_\mathrm{e} = [w_0, \mathbf{w}^T]^T$ and $\mathbf{X}_\mathrm{e} = [1, \mathbf{X}^T]^T$ are extended versions of these vectors.

It can be understood that, by imposing a restriction on the analytical form implemented by the estimator, linear estimators will generally obtain lower performance than the optimal Bayesian estimator. However, the interest of linear estimators is justified by their simplicity and ease of design. As we shall see, for the calculation of the linear estimator of minimum squared mean error, it will be sufficient to know the first and second order statistical moments (means and covariances) associated with the observable variables and the variable to be estimated.

1.4.2.1 Minimization of the mean squared error.

As we have already mentioned, 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 average value of this cost function:

$$\mathbf{w}_{\mathrm{e}}^{*} = \arg\min_{\mathbf{w}_{\mathrm{e}}} \mathbb{E}\{(S-\hat{S})^2\} = \arg\min_{\mathbf{w}_{\mathrm{e}}} \mathbb{E}\{(S-\mathbf{w}_{\mathrm{e}}^T\mathbf{X}_{\mathrm{e}})^2\} \tag{1.44}$$

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

$$\hat{S}_{\mathrm{LMSE}} = \mathbf{w}_{\mathrm{e}}^{*T}\mathbf{X}_{\mathrm{e}}$$

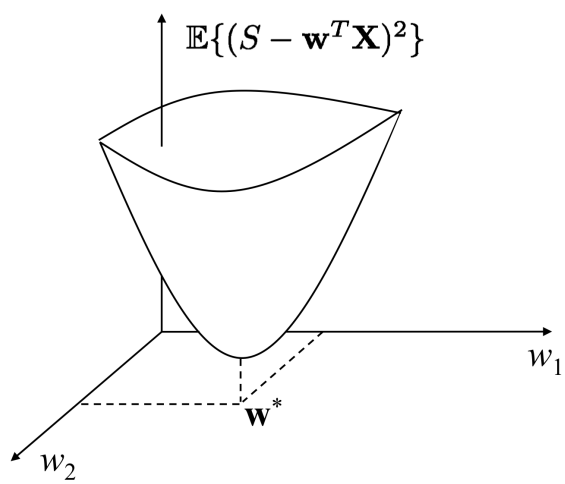


Fig. 1.5 Surface of the mean squared error of the linear estimator as a function of the estimator weights.

Figure 1.5 represents 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. In addition, since the average cost is not negative, it is guaranteed that the function is convex, and its minimum can be located by equaling $\mathbf{0}$ the gradient of the average cost with respect to the weight vector[5]:

$$\begin{aligned}\nabla_{\mathbf{w}_{\mathrm{e}}}\mathbb{E}\{(S-\hat{S})^2\}\big|_{\mathbf{w}_{\mathrm{e}}=\mathbf{w}_{\mathrm{e}}^*} &= -2\mathbb{E}\{(S-\mathbf{w}_{\mathrm{e}}^T\mathbf{X}_{\mathrm{e}})\mathbf{X}_{\mathrm{e}}\}\big|_{\mathbf{w}_{\mathrm{e}}=\mathbf{w}_{\mathrm{e}}^*} = \\ &= -2\mathbb{E}\{(S-\mathbf{w}_{\mathrm{e}}^{*T}\mathbf{X}_{\mathrm{e}})\mathbf{X}_{\mathrm{e}}\} = \mathbf{0}\end{aligned} \tag{1.45}$$

The second line of the above expression defines the conditions to be met by the optimal weight vector. Note that this equation is actually a system of $N+1$ equations (as many as dimensions have $\mathbf{X}_{\mathrm{e}}$) with $N+1$ unknowns (the components of $\mathbf{w}_{\mathrm{e}^*}$).

In order to find the optimal weight vector, it is convenient to rewrite the last line of (1.45) as follows

$$\mathbb{E}\{S\mathbf{X}_{\mathrm{e}}\} = \mathbb{E}\{\mathbf{X}_{\mathrm{e}}(\mathbf{X}_{\mathrm{e}}^T\mathbf{w}_{\mathrm{e}}^*)\} \tag{1.46}$$

[5] 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 $\mathbf{w}$: $\nabla_{\mathbf{w}}f(\mathbf{w}) = \left[\frac{\partial f}{\partial w_1},\ldots\frac{\partial f}{\partial w_N}\right]^T$.

Defining the cross-correlation vector

$$\mathbf{r}_{S\mathbf{X}_e} = \mathbb{E}\{S\mathbf{X}_e\} \tag{1.47}$$

and the correlation matrix

$$\mathbf{R}_{\mathbf{X}_e} = \mathbb{E}\{\mathbf{X}_e\mathbf{X}_e^T\} \tag{1.48}$$

(which is a symmetrical matrix) ec. (1.46) can be written as

$$\mathbf{r}_{S\mathbf{X}_e} = \mathbf{R}_{\mathbf{X}_e}\mathbf{w}_{\mathrm{e}}^* \tag{1.49}$$

Thus, the searched weight vector is:

$$\boxed{\mathbf{w}_{\mathrm{e}}^* = \mathbf{R}_{\mathbf{X}_e}^{-1}\mathbf{r}_{S\mathbf{X}_e}} \tag{1.50}$$

1.4.2.2 Properties of the optimal linear estimator

Equation (1.49) solves the problem of calculating the weights of the estimator $\hat{S}_{\mathrm{LMSE}}$. But it is interesting to return to the vector equation (1.45) to analyze some of its properties. Note that the term in parentheses in this equation is the estimation error

$$E^* = S - \mathbf{w}_{\mathrm{e}}^{*T}\mathbf{X}_{\mathrm{e}} \tag{1.51}$$

so we can rewrite (1.45) as

$$\mathbb{E}\{E^*\mathbf{X}_{\mathrm{e}})\} = \mathbf{0} \tag{1.52}$$

Taking, on the one hand, the first component of this equation (taking into account that $X_{\mathrm{e},1} = 1$, and the rest on the other hand, two fundamental properties of the lowest quadratic mean error linear estimator are obtained:

Property 1: The error has zero mean:

$$\mathbb{E}\{E^*\} = \mathbf{0} \tag{1.53}$$

When an estimator has this property it is said that it is **unbiased**.

Property 2 (Orthogonality Principle): the error is statistically orthogonal to the observations:

$$\mathbb{E}\{E^*\mathbf{X}\} = \mathbf{0} \tag{1.54}$$

1.4.2.3 Alternative expression of the estimator

Expanding Ecs. (1.53) and (1.54), we can obtain the following explicit formulas for the coefficients w_0^* and $\mathbf{w}^*$ of the estimator:

$$w_0^* = m_S - \mathbf{w}^{*T}\mathbf{m_x} \tag{1.55}$$

$$\mathbf{w}^* = \mathbf{V}_\mathbf{X}^{-1}\,\mathbf{v}_{S,\mathbf{X}} \tag{1.56}$$

It can be observed that the role of the bias term w_0 is to compensate for differences between the means of the variable to be estimated and the observations. Therefore, when all the variables involved have null means, $w_0^* = 0$. In contrast to the paper of w_0, we can affirm that the weight vector $\mathbf{w}$ minimizes the mean quadratic error of the fluctuations of S around its mean, exploiting for it the existing statistical relation between S and $\mathbf{X}$.

We will dedicate this section to obtaining the expressions (1.55) and (1.56). The first is a direct consequence of (1.53) that can be developed as

$$m_S - \mathbf{w}^{*T}\mathbf{m_x} - w_0^* = 0 \tag{1.57}$$

solving for w_0^*, we obtain (1.55).

We will now search for an expression for $\mathbf{w}^*$. From (1.54) results

$$\mathbb{E}\{(S - \mathbf{w}^{*T}\mathbf{X} - w_0^*)\mathbf{X}\} = \mathbf{0} \tag{1.58}$$

which can be rewritten as

$$\begin{aligned}\mathbb{E}\{S\mathbf{X}\} &= \mathbb{E}\{(\mathbf{w}^{*T}\mathbf{X} + w_0^*)\mathbf{X}\}\\ &= \mathbb{E}\{\mathbf{X}(\mathbf{X}^T\mathbf{w}^*)\} + w_0^*\mathbb{E}\{\mathbf{X}\}\\ &= \mathbb{E}\{\mathbf{X}\mathbf{X}^T\}\mathbf{w}^* + w_0^*\mathbf{m_X}\end{aligned} \tag{1.59}$$

Now using the expressions that relate the correlation and covariance of two variables:

$$\mathbb{E}\{S\mathbf{X}\} = \mathbf{v}_{S,\mathbf{X}} + m_S\mathbf{m_X} \tag{1.60}$$

$$\mathbb{E}\{\mathbf{X}\mathbf{X}^T\} = \mathbf{V_X} + \mathbf{m_X}\mathbf{m_X}^T \tag{1.61}$$

eq. (1.59) becomes

$$\begin{aligned}\mathbf{v}_{S,\mathbf{X}} &= \mathbf{V_X}\mathbf{w}^* + \mathbf{m_X}\mathbf{m_X}^T\mathbf{w}^* + w_0^*\mathbf{m_X} - m_S\mathbf{m_X}\\ &= \mathbf{V_X}\mathbf{w}^* + \mathbf{m_X}(w_0^* + \mathbf{m_X}^T\mathbf{w}^* - m_S)\\ &= \mathbf{V_X}\mathbf{w}^*\end{aligned} \tag{1.62}$$

where, in the last equality, we have applied (1.55). So, solving for $\mathbf{w}^*$, we get (1.56).

1.4.2.4 Minimum squared mean error

Here we will calculate the mean squared error associated with the linear estimator of minimum mean squared error, $\hat{S}_{\mathrm{LMSE}}$. As commented at the beginning of this section, the mean squared error obtained will, in general, be higher than the minimum mean squared error

of the Bayesian estimator ($\hat{S}_{\text{MMSE}}$), except when this last estimator has precisely a linear structure (in this case, it would be the same).

To calculate the mean squared error we only have to develop the expression of the mean quadratic error, particularizing it for $\hat{S}_{\text{LMSE}}$ and leaving the result in function of the mathematical expectations of the involved random variables:

$$\begin{aligned}\mathbb{E}\{(S-\hat{S}_{\text{LMSE}})^2\} &= \mathbb{E}\{E^*(S-w_0^*-\mathbf{w}^{*T}\mathbf{X})\}\\ &= \mathbb{E}\{E^*S\}-w_0^*\mathbb{E}\{E^*\}-\mathbf{w}^{*T}\mathbb{E}\{\mathbf{X}E^*\}\\ &= \mathbb{E}\{E^*S\} \end{aligned} \tag{1.63}$$

where, in the last equality, we have applied the two properties of the minimum quadratic mean error estimator obtained in (1.53) and (1.54). Operating again the error term, E^*, results in

$$\begin{aligned}\mathbb{E}\{(S-\hat{S}_{\text{LMSE}})^2\} &= \mathbb{E}\{S(S-w_0^*-\mathbf{w}^{*T}\mathbf{X})\}\\ &= \mathbb{E}\{S^2\}-w_0^*m_S-\mathbf{w}^{*T}(\mathbf{v}_{S\mathbf{X}}+m_S\mathbf{m}_{\mathbf{X}})\}\\ &= \mathbb{E}\{S^2\}-m_S(w_0^*+\mathbf{w}^{*T}\mathbf{m}_{\mathbf{X}})-\mathbf{w}^{*T}\mathbf{v}_{S\mathbf{X}}\\ &= v_S-\mathbf{w}^{*T}\mathbf{v}_{S\mathbf{X}} \end{aligned} \tag{1.64}$$

Exercise 1.1 (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

$$\begin{array}{lll}\mathbb{E}\{S\}=1/2 & \mathbb{E}\{X_1\}=1 & \mathbb{E}\{X_2\}=0\\ \mathbb{E}\{S^2\}=4 & \mathbb{E}\{X_1^2\}=3/2 & \mathbb{E}\{X_2^2\}=2\\ \mathbb{E}\{SX_1\}=1 & \mathbb{E}\{SX_2\}=2 & \mathbb{E}\{X_1X_2\}=1/2\end{array}$$

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 Estimation with gaussian distributions

In this section we will analyze the case of random variable estimation when the combined distribution of all the variables involved (variable to be estimated and observation variables) is a multidimensional Gaussian. This case is of special interest given the frequency with which these distributions usually appear in problems in the field of telecommunications and in other scenarios. In this case, it can be shown that all marginal distributions and all conditional distributions are also Gaussian. Specifically, given that $p_{S|\mathbf{X}}(s|\mathbf{x})$ is Gaussian, it can be understood that the fashion, the mean and the median of the distribution coincide, so $\hat{s}_{\text{MSE}} = \hat{s}_{\text{MAD}} = \hat{s}_{\text{MAP}}$ will be verified. Therefore, during this section we will focus our discussion on the calculation of the minimum quadratic mean error estimator.

Besides, we will demonstrate that the 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 mean squared error estimators.

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:

$$p_{S,X}(s,x) \sim G\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} v_S & \rho \\ \rho & v_X \end{bmatrix}\right) \tag{1.65}$$

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 of S can be obtained as:

$$\begin{aligned} p_{S|X}(s|x) &= \frac{p_{S,X}(s,x)}{p_X(x)} \\ &= \frac{\dfrac{1}{2\pi\sqrt{v_X v_S - \rho^2}} \exp\left[-\dfrac{1}{2(v_X v_S - \rho^2)} \begin{bmatrix} s \\ x \end{bmatrix}^T \begin{bmatrix} v_X & -\rho \\ -\rho & v_S \end{bmatrix} \begin{bmatrix} s \\ x \end{bmatrix}\right]}{\dfrac{1}{\sqrt{2\pi v_X}} \exp\left[-\dfrac{x^2}{2v_X}\right]} \end{aligned} \tag{1.66}$$

where it has been necessary to calculate the inverse of the covariance matrix of S and X, which is easy since the matrix has dimensions of 2×2.

Our goal for obtaining $\hat{s}_{\text{MSE}}$ is to calculate the mean of that distribution. However, a direct calculation by integrating your product with s is quite complicated. However, given the joint Gaussian character of S and X, we know that the posterior distribution of S must necessarily be Gaussian, defined by its (unknown) parameters of mean and variance $m_{S|X}$ and $v_{S|X}$, respectively, allowing the above expression to be rewritten as:

$$\frac{1}{\sqrt{2\pi v_{S|X}}}\exp\left[-\frac{(s-m_{S|X})^2}{2v_{S|X}}\right] = \frac{\dfrac{1}{2\pi\sqrt{v_X v_S-\rho^2}}\exp\left[-\dfrac{1}{2(v_X v_S-\rho^2)}\begin{bmatrix}s\\x\end{bmatrix}^T\begin{bmatrix}v_X & -\rho\\-\rho & v_S\end{bmatrix}\begin{bmatrix}s\\x\end{bmatrix}\right]}{\dfrac{1}{\sqrt{2\pi v_X}}\exp\left[-\dfrac{x^2}{2v_X}\right]} \tag{1.67}$$

It is possible to break this equality down into two others associated with factors external to the exponentials and their arguments:

$$\frac{1}{\sqrt{2\pi v_{S|X}}} = \frac{\sqrt{2\pi v_X}}{2\pi\sqrt{v_X v_S-\rho^2}} \tag{1.68}$$

$$\frac{(s-m_{S|X})^2}{v_{S|X}} = \frac{1}{v_X v_S-\rho^2}\begin{bmatrix}s\\x\end{bmatrix}^T\begin{bmatrix}v_X & -\rho\\-\rho & v_S\end{bmatrix}\begin{bmatrix}s\\x\end{bmatrix} - \frac{x^2}{v_X} \tag{1.69}$$

By operating the matrix terms, the second of these equals can be more simply rewritten as

$$\frac{(s-m_{S|X})^2}{v_{S|X}} = \frac{v_X s^2+v_S x^2-2\rho xs}{v_X v_S-\rho^2} - \frac{x^2}{v_X} \tag{1.70}$$

Note that (1.70) assumes an equality between two polynomials in s (and in x). Therefore, the coefficients of the independent, linear and quadratic terms in s (i.e., which do not depend on s, or which multiply to s and s^2) that appear on both sides of the equality must match. Therefore, and taking into account that $m_{S|X}$ does not depend on s, the following three equality must be verified:

$$\frac{m_{S|X}^2}{v_{S|X}} = \frac{v_S x^2}{v_X v_S-\rho^2} - \frac{x^2}{v_X} \tag{1.71}$$

$$\frac{s\, m_{S|X}}{v_{S|X}} = \frac{\rho xs}{v_X v_S-\rho^2} \tag{1.72}$$

$$\frac{s^2}{v_{S|X}} = \frac{v_X s^2}{v_X v_S-\rho^2} \tag{1.73}$$

For the calculation of the posterior mean, it is convenient solving (1.72) for $m_{S|X}$ as

$$m_{S|X} = \frac{v_{S|X}\rho x}{v_X v_S-\rho^2} \tag{1.74}$$

Finally, the value of the posterior variance can easily be extracted from (1.68) or (1.73) as

$$v_{S|X} = \frac{v_X v_S-\rho^2}{v_X} \tag{1.75}$$

Replacing this value in (1.74) gives the expression that determines the minimum quadratic mean error estimator.

$$\hat{s}_{\mathrm{MSE}} = m_{S|X} = \frac{\rho}{v_X} x \tag{1.76}$$

As can be seen, the estimator obtained is linear.

Exercise 1.2 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

$$\hat{s}_{\mathrm{MSE}} = m_S + \frac{\rho}{v_X}(x - m_X) \tag{1.77}$$

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 obtained as the sum of the signal to be estimated and a noise component independent of the signal: $X = S + R$. Both the signal and the noise present Gaussian distributions of zero means and variances v_S and v_R, respectively.

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

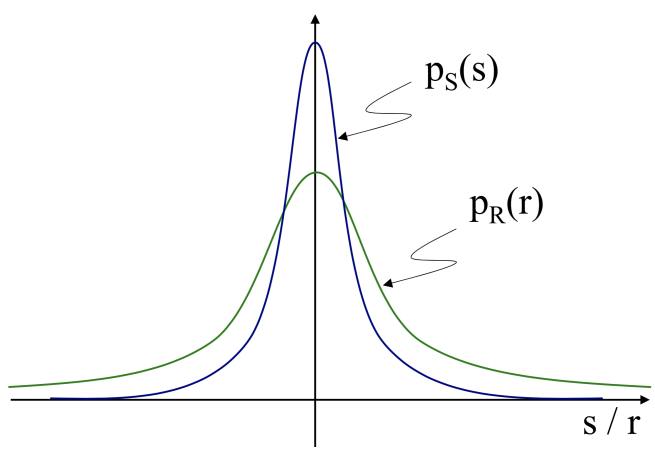


Fig. 1.6 Estimation of Gaussian random variable S contaminated by Gaussian noise R.

According to (1.76), 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 we can proceed as follows:

$$\rho = \mathbb{E}\{(X - m_X)(S - m_S)\} = \mathbb{E}\{X\,S\} = \mathbb{E}\{(S + R)S\} = \mathbb{E}\{S^2\} + \mathbb{E}\{S\,R\} = v_S \tag{1.78}$$

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

Replacing these results in (1.76) we get

$$\hat{s}_{\text{MSE}} = \frac{v_S}{v_S + v_R} x \tag{1.79}$$

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), $v_S \gg v_R$) you have to $\hat{s}_{\text{MSE}} \to x$, 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 low ($v_S \ll v_R$), 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.

$$p_{\mathbf{S},\mathbf{X}}(\mathbf{s},\mathbf{x}) \sim G\left(\begin{bmatrix} \mathbf{m_S} \\ \mathbf{m_X} \end{bmatrix}, \begin{bmatrix} \mathbf{V_S} & \mathbf{V_{SX}} \\ \mathbf{V_{SX}}^T & \mathbf{V_X} \end{bmatrix} \right) \tag{1.80}$$

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,

$$\mathbf{V_S} = \mathbb{E}\{(\mathbf{S} - \mathbf{m_S})(\mathbf{S} - \mathbf{m_S})^T\} \tag{1.81}$$

$$\mathbf{V_X} = \mathbb{E}\{(\mathbf{X} - \mathbf{m_X})(\mathbf{X} - \mathbf{m_X})^T\} \tag{1.82}$$

$$\mathbf{V_{SX}} = \mathbb{E}\{(\mathbf{S} - \mathbf{m_S})(\mathbf{X} - \mathbf{m_X})^T\} \tag{1.83}$$

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

$$\mathbf{m_{S|X}} = \mathbf{m_S} + \mathbf{V_{SX}}\mathbf{V_X}^{-1}(\mathbf{x} - \mathbf{m_X}) \tag{1.84}$$

and covariance

$$\mathbf{V_{S|X}} = \mathbf{V_S} - \mathbf{V_{SX}}\mathbf{V_X}^{-1}\mathbf{V_{SX}}^T \tag{1.85}$$

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

$$\boxed{\hat{\mathbf{s}}_{\text{MSE}} = \mathbf{m_S} + \mathbf{V_{SX}}\mathbf{V_X}^{-1}(\mathbf{x} - \mathbf{m_X}) \qquad (1.86)}$$

This estimator expression is simplified when $\mathbf{S}$ and $\mathbf{X}$ have zero means, resulting in

$$\hat{\mathbf{s}}_{\mathrm{MSE}} = \mathbf{m}_{\mathbf{S}|\mathbf{X}} = \mathbf{V}_{\mathbf{SX}}\mathbf{V}_{\mathbf{X}}{}^{-1}\mathbf{x} \tag{1.87}$$

1.5.3 Linear estimation and Gaussian estimation

Regrouping the terms of (1.86), we can express $\hat{\mathbf{s}}_{\mathrm{MSE}}$ as:

$$\boxed{\hat{\mathbf{s}}_{\mathrm{MSE}} = (\mathbf{m}_{\mathbf{S}} - \mathbf{V}_{\mathbf{SX}}\mathbf{V}_{\mathbf{X}}{}^{-1}\mathbf{m}_{\mathbf{X}}) + \mathbf{V}_{\mathbf{SX}}\mathbf{V}_{\mathbf{X}}{}^{-1}\mathbf{x}} \tag{1.88}$$

and identifying these terms with the coefficients of a linear estimator, we get

$$\mathbf{w}^T = \mathbf{V}_{\mathbf{SX}}\mathbf{V}_{\mathbf{X}}{}^{-1} \tag{1.89}$$

$$w_0 = \mathbf{m}_{\mathbf{S}} - \mathbf{w}^T\mathbf{m}_{\mathbf{X}} \tag{1.90}$$

These expressions coincides with the alternatives solution of the linear estimation of mean squared error (equations 1.55 and 1.56).This is not surprising: since the unrestricted MSE estimator in the Gaussian case is linear, the best linear estimator must match the one obtained for the Gaussian case.

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 they are not treated them 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 a maximum likelihood criteria.

Note that in order to use some Bayesian estimator, it would be necessary to know the posterior and without having information on the prior of these parameters we cannot know the posterior. Therefore, the only estimator we can apply in this scenario is the maximum likelihood estimator.

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

It is known that the weight of individuals of a family of mollusks follows a Gaussian distribution, whose mean and variance is to be estimated. It is available for the estimation of the weights of l individuals taken independently, $\{X^{(k)}\}_{k=1}^{l}$.

The likelihood of the mean and the variance, in this case, consists simply of the probability distribution of the observations, which is given by:

$$p_X(x) = p_{X|m,v}(x|m,v) = \frac{1}{\sqrt{2\pi v}} \exp\left[-\frac{(x-m)^2}{2v}\right] \tag{1.91}$$

for each observation. 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:

$$\begin{aligned} p_{\{X^{(k)}\}|m,v}(\{x^{(k)}\}|m,v) &= \prod_{k=1}^{l} p_{X|m,v}(x^{(k)}|m,v) \\ &= \frac{1}{(2\pi v)^{l/2}} \prod_{k=1}^{l} \exp\left[-\frac{(x^{(k)}-m)^2}{2v}\right] \end{aligned} \tag{1.92}$$

The maximum likelihood estimators of m and v will be the values of those parameters that make the above expression maximum. The analytical form of (1.92) suggests the use of the logarithm function to simplify the maximization process:

$$L = \ln\left[p_{\{X^{(k)}\}|m,v}(\{x^{(k)}\}|m,v)\right] = -\frac{l}{2}\ln(2\pi v) - \frac{1}{2v}\sum_{k=1}^{l}(x^{(k)}-m)^2 \tag{1.93}$$

To obtain the maximum likelihood estimators we will proceed to derive (1.93) with respect to m and v, and to equal the result with respect to 0. Thus, the system of equations to solve is

$$\begin{aligned}
\left.\frac{d\,L}{d\,m}\right|_{\substack{m=\hat{m}_{\mathrm{ML}}\\ v=\hat{v}_{\mathrm{ML}}}} &= -\frac{1}{v}\sum_{k=1}^{l}(x^{(k)}-m)\Bigg|_{\substack{m=\hat{m}_{\mathrm{ML}}\\ v=\hat{v}_{\mathrm{ML}}}} = 0 \\
\left.\frac{d\,L}{d\,v}\right|_{\substack{m=\hat{m}_{\mathrm{ML}}\\ v=\hat{v}_{\mathrm{ML}}}} &= -\frac{l}{2v}+\frac{1}{2v^2}\sum_{k=1}^{l}(x^{(k)}-m)^2\Bigg|_{\substack{m=\hat{m}_{\mathrm{ML}}\\ v=\hat{v}_{\mathrm{ML}}}} = 0
\end{aligned} \tag{1.94}$$

The first of these equations allows to obtain the estimator of the mean in a simple way as the sample average of the observations, i.e.,

$$\hat{m}_{\mathrm{ML}} = \frac{1}{l}\sum_{k=1}^{l} x^{(k)} \tag{1.95}$$

On the other hand, we can solve the second equation of the system for the ML estimator of the variance, obtaining

$$\hat{v}_{\mathrm{ML}} = \frac{1}{l}\sum_{k=1}^{l}(x^{(k)}-\hat{m}_{\mathrm{ML}})^2 \tag{1.96}$$

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.,

$$\hat{M}_{\mathrm{ML}} = \frac{1}{l}\sum_{k=1}^{l} X^{(k)} \tag{1.97}$$

$$\hat{V}_{\mathrm{ML}} = \frac{1}{l}\sum_{k=1}^{l}[X^{(k)}-\hat{M}_{\mathrm{ML}}]^2 \tag{1.98}$$

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 givne by the following posterior distribution:

$$p_{S|X}(s|x) = x \exp(-xs), \;\; s > 0 \tag{1.99}$$

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 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 $\hat{S}_{\text{MMSE}}$.

1.5 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 to estimate r by maximum likelihood given the 64 pixel values of the transformed image $\{y^{(k)}\}_{k=1}^{64}$, without knowing the original image.

1.6 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.

1.7 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 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 on-line 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)

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

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)

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

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

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

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:

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

Chapter 2
Linear Filtering

2.1 Introduction

A common problem in estimation is that of wanting to determine the coefficients of a linear filter with M coefficients from the mere observation of its inputs and outputs. This task, as well as related ones, is known by the generic name of "linear filtering". In this block we will show how the techniques described in block B1 can be used to design ML, MAP, MAD and MMSE estimators of the coefficients of said filter, as well as of future filter outputs if 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 $u[n]$. The result is added a certain Gaussian noise $\varepsilon[n]$, which is i.i.d. zero-mean stochastic process with variance σ_ε^2, giving rise to an observation $x[n]$. That is, the corresponding entries are

$$x[n] = u[n] * s[n] + \varepsilon[n] \tag{2.1}$$

$$= u[n]s[0] + u[n-1]s[1] + \ldots + u[n-M+1]s[M-1] + \varepsilon[n]. \tag{2.2}$$

Joining the nonzero coefficients in vector $\mathbf{s} = [s[0], s[1], \ldots, s[M-1]]^\top$ and compacting every M-length sequence of consecutive input values into vectors $\mathbf{u}[n] = [u[n], u[n-1], \ldots, u[n-M+1]]^\top$, we can write

$$x[n] = \mathbf{u}[n]^\top \mathbf{s} + \varepsilon[n]. \tag{2.3}$$

The filtering problem consists in estimating the filter coefficients $\mathbf{s}$ from a set of observed inputs and outputs, as well as estimating the output x_* corresponding to a new input $\mathbf{u}_*$.

If we have the signals $u[n]$ and $x[n]$ in the range $0 \leq n \leq N-1$ and assuming that both signals are null for $n < 0$, we will have a total of N input-output pairs, $\{\mathbf{u}[n], x[n]\}_{n=0}^{N-1}$. We can group these input-output couples in the $\mathbf{x}$ and $\mathbf{U}$ matrices:

$$\mathbf{x} = \begin{bmatrix} x[0] \\ x[1] \\ \vdots \\ x[N-1] \end{bmatrix}_{N\times 1}, \tag{2.4}$$

$$\begin{aligned} \mathbf{U} &= [\mathbf{u}[0] \;\; \mathbf{u}[1] \;\; \dots \;\; \mathbf{u}[M-1] \;\; \dots \;\; \mathbf{u}[N-1]] \\ &= \begin{bmatrix} u[0] & u[1] & \dots & u[M-1] & \dots & u[N-1] \\ 0 & u[0] & \dots & u[M-2] & \dots & u[N-2] \\ \vdots & \vdots & \ddots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & u[0] & \dots & u[N-M] \end{bmatrix}_{M\times N}, \end{aligned} \tag{2.5}$$

Also, defining the noise vector

$$\boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon[0] \\ \varepsilon[1] \\ \vdots \\ \varepsilon[N-1] \end{bmatrix}_{N\times 1}, \tag{2.6}$$

$$\tag{2.7}$$

we can write the signal model (2.3) as

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

This matrix form will be useful to obtain compact expressions in the following sections.

Note: Along the subsequent derivations, signal $u[n]$ signal and therefore matrix $\mathbf{U}$ matrix are considered as observed and deterministic values, to which all probabilistic expressions are implicitly conditioned.

2.3 ML solution

The problem statement itself provides us the likelihood of the $\mathbf{s}$ filter coefficients given the n-th observation: The problem statement povides

$$p(x[n]|\mathbf{s}) = \mathcal{N}(x[n]|\mathbf{u}[n]^\top \mathbf{s}, \sigma_\varepsilon^2), \tag{2.9}$$

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

Given a set of observations, we simply take the product of the previous likelihoods, since the noise terms are independent

$$p(\mathbf{x}|\mathbf{s}) = \prod_{n=0}^{N-1} \mathcal{N}(x[n]|\mathbf{u}[n]^\top \mathbf{s}, \sigma_\varepsilon^2) = \mathcal{N}(\mathbf{x}|\mathbf{U}^\top \mathbf{s}, \sigma_\varepsilon^2 \mathbf{I}). \tag{2.10}$$

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

$$\begin{aligned}
\hat{\mathbf{s}}_{\mathrm{ML}} &= \underset{\mathbf{s}}{\operatorname{argmax}}\ p(\mathbf{x}|\mathbf{s}) = \underset{\mathbf{s}}{\operatorname{argmax}}\ \log p(\mathbf{x}|\mathbf{s}) \\
&= \underset{\mathbf{s}}{\operatorname{argmin}}\ \frac{1}{2}(\mathbf{x}-\mathbf{U}^\top\mathbf{s})^\top(\sigma_\varepsilon^2\mathbf{I})^{-1}(\mathbf{x}-\mathbf{U}^\top\mathbf{s}) + \frac{1}{2}\log|\sigma_\varepsilon^2\mathbf{I}| + \frac{N}{2}\log(2\pi) \\
&= \underset{\mathbf{s}}{\operatorname{argmin}}\ ||\mathbf{x}-\mathbf{U}^\top\mathbf{s}||^2 \qquad (2.11) \\
&= (\mathbf{U}\mathbf{U}^\top)^{-1}\mathbf{U}\mathbf{x}. \qquad (2.12)
\end{aligned}$$

The last step is simply the least squares solution seen in the regression chapter. This minimum can be easily obtained by taking the gradient with respect to $\mathbf{s}$, equalizing to zero and clearing.

2.4 Bayesian Solution

To obtain a Bayesian estimator of $\mathbf{s}$ we need to know its a priori probability $p(\mathbf{s})$. Although this is generally unknown, it is sensible to use

$$p(\mathbf{s}) = \mathcal{N}(\mathbf{s}|\mathbf{0}, \sigma_s^2\mathbf{I}), \qquad (2.13)$$

since it considers acceptable any set of real coefficients, and assumes that these have a null mean and a dispersion set by σ_s^2. It is also possible to set $\sigma_s^2 \to \infty$ to achieve a uniform distribution. In any case, the use of this distribution a priori allows to obtain the distribution a posteriori analytically.

Given the likelihood, $p(\mathbf{x}|\mathbf{s})$, and the a priori distribution $p(\mathbf{s})$, we can obtain the posterior distribution $p(\mathbf{s}|\mathbf{x})$. To do this, we could directly apply Bayes' theorem and simplify the quotient as much as possible, but this is a very tedious process. Instead, we will get the result in two steps.

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

$$\begin{bmatrix}\mathbf{s}\\ \mathbf{x}\end{bmatrix} = \begin{bmatrix}\mathbf{I}\\ \mathbf{U}^\top\end{bmatrix}\mathbf{s} + \begin{bmatrix}\mathbf{0}\\ \boldsymbol{\varepsilon}\end{bmatrix} \qquad (2.14)$$

that is, vector $[\mathbf{s}^\top\ \mathbf{x}^\top]^\top$ is a linear combination of r.v. with Gaussian pdf plus an indpendent white Gaussian noise with variance σ_ε^2 and, thus, it is jointly Gaussian. The computation of the mean and the variance of $[\mathbf{s}^\top\ \mathbf{x}^\top]^\top$ is straightforward:

$$\begin{bmatrix}\mathbf{s}\\ \mathbf{x}\end{bmatrix} = \mathcal{N}\left(\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix}, \begin{bmatrix}\sigma_s^2\mathbf{I} & \sigma_s^2\mathbf{U}\\ \sigma_s^2\mathbf{U}^\top & \sigma_s^2\mathbf{U}^\top\mathbf{U}+\sigma_\varepsilon^2\mathbf{I}\end{bmatrix}\right) \qquad (2.15)$$

and using the Gaussian conditioning formula in he previous chapter, we get

$$p(\mathbf{s}|\mathbf{x}) = \mathcal{N}(\mathbf{s} \mid \sigma_s^2\mathbf{U}(\sigma_s^2\mathbf{U}^\top\mathbf{U}+\sigma_\varepsilon^2\mathbf{I})^{-1}\mathbf{x},\ \ \sigma_s^2\mathbf{I}-\sigma_s^2\mathbf{U}(\sigma_s^2\mathbf{U}^\top\mathbf{U}+\sigma_\varepsilon^2\mathbf{I})^{-1}\mathbf{U}^\top\sigma_s^2), \qquad (2.16)$$

Using the matrix investion lemma and some algebra, this can be show to be equivalent to the followinglo expression, whichis computationally more efficient for $M < N$:

$$p(\mathbf{s}|\mathbf{x}) = \mathcal{N}(\mathbf{s} \mid \mathbf{PUx},\ \ \sigma_\varepsilon^2\mathbf{P}), \tag{2.17}$$

where

$$\mathbf{P} = (\mathbf{UU}^\top + \tfrac{\sigma_\varepsilon^2}{\sigma_s^2}\mathbf{I})^{-1} \tag{2.18}$$

. Thus the MMSE and MAP estimates of $\mathbf{s}$ are:

$$\hat{\mathbf{s}}_{\text{MMSE}} = \hat{\mathbf{s}}_{\text{MAP}} = \hat{\mathbf{s}}_{\text{MAD}} = \mathbf{PUx} \tag{2.19}$$

Note that taking $\sigma_s^2 \to \infty$ (which can be interpreted as assuming an infinitely uniform prior) the MAP solution becomes equvalent to the ML in (2.12).

2.4.1 Probabilistic prediction of the filter output

Once we have resolved several estimators of filter $\mathbf{s}$ filter, we now begin to consider the problem of estimating a new output x_* corresponding to a new entry $\mathbf{u}_*$. Continuing with the Bayesian perspective, we will obtain the fdp a posteriori of the variable to be estimated, x_*, in view of the outputs already observed, $\mathbf{x}$. That is, we want to calculate $p(x_*|\mathbf{x})$.

First, it should be noted that $\mathbf{x}, x_*$ and $\mathbf{s}$ are jointly Gaussian. This follows from Eq. (2.15), which can be extended to any arbitrary number of outputs, including x_*. This necessarily implies that $\mathbf{x}$ and x_* are jointly Gaussian (when marginalizing $\mathbf{s}$) and finally that $p(x_*|\mathbf{x})$ must be Gaussian. Since

$$x_* = \mathbf{u}_*^\top \mathbf{s} + \varepsilon_* \tag{2.20}$$

is a linear transformation of $\mathbf{s}$ with independent white noise, we can easily compute the mean, $\mathbb{E}[x_*|\mathbf{x}]$, and the variance, $\mathbb{V}[x_*|\mathbf{x}]$, of this Gaussian posterior distribution using $p(\mathbf{s}|\mathbf{x})$, obtaining

$$p(x_*|\mathbf{x}) = \mathcal{N}(x_* \mid \mathbf{u}_*^\top \mathbf{PUx},\ \ \sigma_\varepsilon^2 + \sigma_\varepsilon^2 \mathbf{u}_*^\top \mathbf{Pu}_*) \tag{2.21}$$

. that inmediatly provides the following estimators for x_*:

$$\hat{x}_{*\text{MMSE}} = \hat{x}_{*\text{MAP}} = \hat{x}_{*\text{MAD}} = \mathbf{u}_*^\top \mathbf{PUx} = \mathbf{u}_*^\top \hat{\mathbf{s}}_{\text{MMSE}}. \tag{2.22}$$

We observe, thus, that in order to obtaine the estimators, for the new out x_*, we only need to know the new input, $\mathbf{u}_*$, and the estimator $\mathbf{s}_{\text{MMSE}}$.

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. (2.12) or (2.19) requires inverting an $M \times M$ matrix. This has a cost $\mathcal{O}(M^3)$, 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

$\hat{\mathbf{s}}_{\text{MMSE}}$ 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

$$\mathbf{P}_N = (\mathbf{U}\mathbf{U}^\top + \tfrac{\sigma_\varepsilon^2}{\sigma_s^2}\mathbf{I})^{-1}, \tag{2.23}$$

$$\mathbf{r}_N = \mathbf{U}\mathbf{x} \tag{2.24}$$

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

$$\hat{\mathbf{s}}_{\text{ML}} = \underset{\mathbf{s}}{\operatorname{argmax}}\ p(\mathbf{x}|\mathbf{s}) = \underset{\mathbf{s}}{\operatorname{argmin}}\ ||\mathbf{x} - \mathbf{U}^\top\mathbf{s}||^2 \tag{2.25}$$

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

Notice that

$$||\mathbf{x} - \mathbf{U}^\top\mathbf{s}||^2 = \sum_{n=0}^{N-1}(x[n] - \mathbf{u}[n]^\top\mathbf{s})^2, \tag{2.26}$$

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

$$\hat{\mathbf{s}}_{N+1} = \hat{\mathbf{s}}_N + \mu \left(x[n] - \mathbf{u}[n]^\top \hat{\mathbf{s}}_N \right) \mathbf{u}[n], \tag{2.27}$$

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

2.6 Wiener filter

The Wiener filter $\mathbf{s}_{\text{Wiener}}$ is the filter that minimizes the expected square error between a desired output $x[n]$ and the output produced when used to filter the input $u[n]$. In this section, both $x[n]$ and $u[n]$ are considered null half signals and $u[n]$ is treated as a stochastic process and not as a deterministic signal, as has been done up to now.

This problem can be posed as a linear estimation problem of minimum mean square error (MMSE), so the formulation of the previous chapter can be used to give rise to the following solution:

$$\mathbf{s}_{\text{Wiener}} = \mathbf{R}_{uu}^{-1} \mathbf{r}_{ux}, \tag{2.28}$$

where $\mathbf{R}_{uu}$ is the autocorrelation matrix of the input signal $u[n]$ and $\mathbf{r}_{ux}$ is the cross-correlation vector between $u[n]$ and $x[n]$. Unfortunately, these two quantities are generally unknown, so in most cases, the Wiener filter cannot be calculated. However, it is common to use the above expression using sample estimates for the correlation matrix $\hat{\mathbf{R}}_{uu} = \frac{1}{N}\mathbf{U}\mathbf{U}^\top$ and the cross-correlation vector $\hat{\mathbf{r}}_{ux} = \frac{1}{N}\mathbf{U}\mathbf{x}$. The result is an approximation to the Wiener filter $\hat{\mathbf{s}}_{\text{Wiener}} = \hat{\mathbf{R}}_{uu}^{-1}\hat{\mathbf{r}}_{ux}$ that minimizes the sample quadratic error (often called "least-squares estimate") and which matches the ML solution, that is $\hat{\mathbf{s}}_{\text{Wiener}} = \hat{\mathbf{s}}_{\text{ML}}$.

As the number of samples available for the estimation of the $\mathbf{R}_{uu}$ and $\mathbf{r}_{ux}$ statistics increases, these estimates become more precise, so that $\hat{\mathbf{s}}_{\text{Wiener}}$ and therefore $\hat{\mathbf{s}}_{\text{ML}}$ match asymptotically with the exact Wiener filter.

2.7 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 $\mathbf{s}$? (Wiener filter based on approximate statistics).
b) Use the obtained filter to predict $x[7]$, $\hat{x}_{\text{ML}}$.
c) Calculate the MMSE, MAP and MAD estimates of $\mathbf{s}$ assuming that the a priori pdf of its components is $s_i \sim \mathcal{N}(0,1)$.
d) Get the MMSE estimate of $x[7]$, $\hat{x}_{\text{MMSE}}$.

e) Calculate the expected square error in prediction b). (That is, the hope of $(\hat{x}_{\mathrm{ML}} - x[7])^2$ in view of the available data).
f) Calculate the expected square error in prediction d). (That is, the hope of $(\hat{x}_{\mathrm{MMSE}} - x[6])^2$ in view of the available data)

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 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.

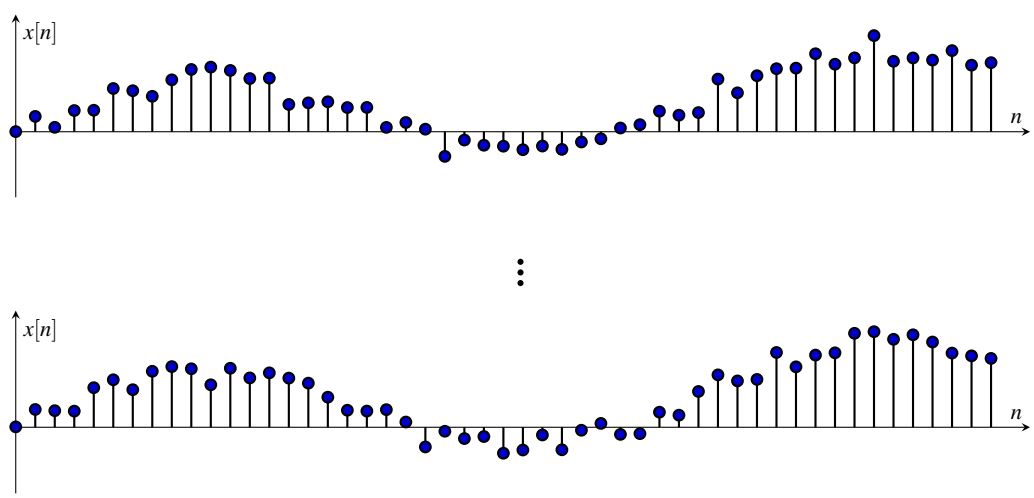


Fig. 3.1 Realizations of a discrete stochastic process

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 $4\pi/N$.

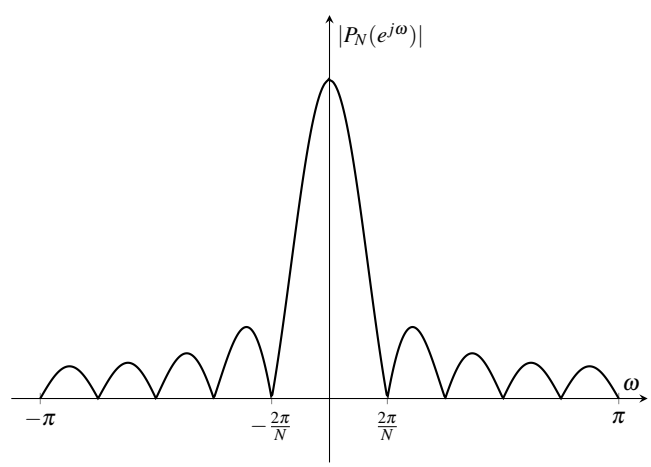


Fig. 3.2 Module of the Fourier transform of the rectangular window

[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

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

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}.

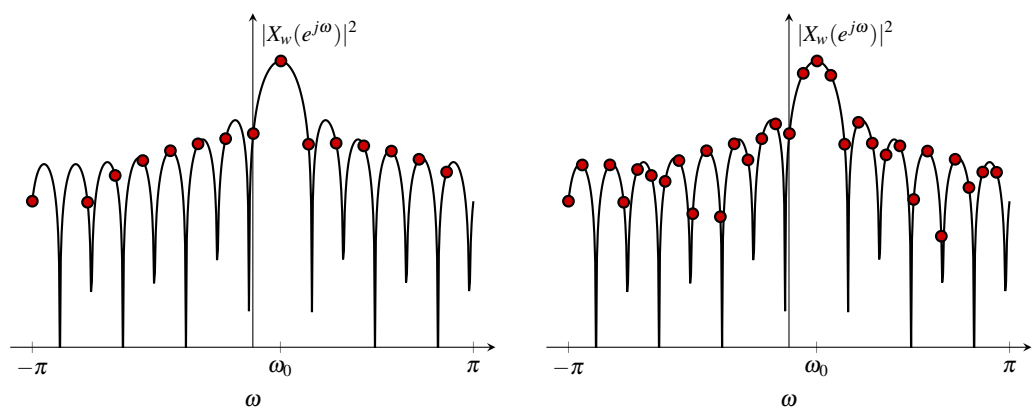


Fig. 3.3 Fourier transform (in logarithmic scale) of a windowed complex exponential

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.

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

$$\begin{aligned}|X_w(e^{j\omega})|^2 = &\left|P_N\left(e^{j(\omega-\omega_0)}\right)\right|^2 + \left|P_N\left(e^{j(\omega-\omega_1)}\right)\right|^2 \\ &+ e^{j(\omega_0-\omega_1)(N-1)/2} P_N\left(e^{j(\omega-\omega_0)}\right) P_N\left(e^{j(\omega-\omega_1)}\right) \\ &+ e^{-j(\omega_0-\omega_1)(N-1)/2} P_N\left(e^{j(\omega-\omega_0)}\right) P_N\left(e^{j(\omega-\omega_1)}\right).\end{aligned}$$

Taking into account Euler's formula, $|X_w(e^{j\omega})|^2$ can be simplified as

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

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

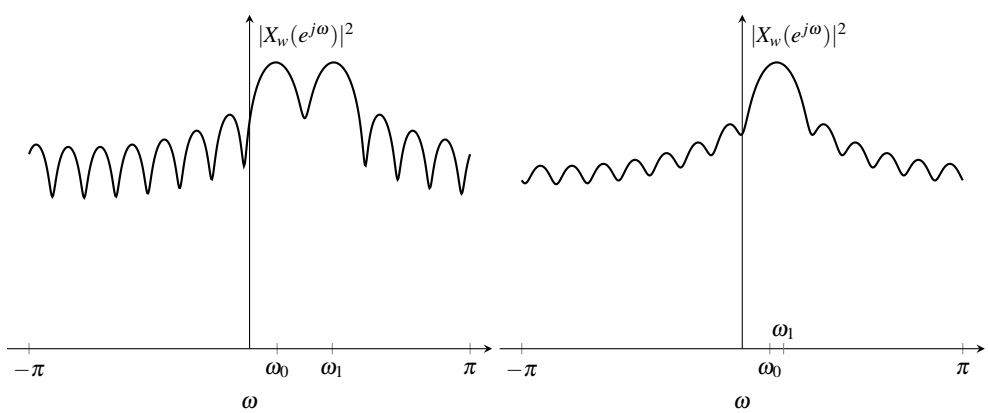


Fig. 3.4 Fourier transform (in logarithmic scale) of the sum of two complex exponentials

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.

3.3.1 The periodogram

Despite the title section, we will start with an estimator known as correlogram, which is based on a first definition of the PSD.

Definition 3.1 Given a WSS process $x[n]$, the power spectral density is defined as

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

where

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

is the auto-correlation function of the process $x[n]$.

Based on Definition 3.1, the first estimator of the PSD is given by

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

where $\hat{r}_x[m]$ is an estimator of the auto-correlation function. Concretely, we have two alternatives for this estimator: a biased and an unbiased estimator. The biased estimator of the auto-correlation, given the finite length $x[n], n = 0, \ldots, N-1$, is

$$\hat{r}_x^b[m] = \frac{1}{N}\sum_{n=m}^{N-1} x[n]x^*[n-m], \quad m = 0, \ldots, N-1, \tag{3.2}$$

and $\hat{r}_x^b[m] = \left[\hat{r}_x^b[-m]\right]^*, m = -N+1, \ldots, -1$, whereas the unbiased estimator is

$$\hat{r}_x^u[m] = \frac{1}{N-m}\sum_{n=m}^{N-1} x[n]x^*[n-m], \quad m = 0, \ldots, N-1, \tag{3.3}$$

and $\hat{r}_x^u[m] = [\hat{r}_x^u[-m]]^*, m = -N+1, \ldots, -1$. Consequently, the PSD estimators are

$$\hat{S}_x^b(e^{j\omega}) = \mathscr{F}(\hat{r}_x^b[m]), \tag{3.4}$$

and

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

Actually, only the estimator in (3.4), which is known as the correlogram, is a valid PSD estimator since it ensures $\hat{S}_x^b(e^{j\omega}) \geq 0$, whereas $\hat{S}_x^u(e^{j\omega}) \ngeq 0$.

The periodogram, which is a term coined by Arthur Schuster in 1898, is based on a second definition of the power spectral density. This definition states that

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

The periodogram is obtained from the above definition by simply dropping the expectation and considering a finite number of samples, i.e.,

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

where $X(e^{j\omega}) = \mathscr{F}(x[n])$.

In the following, we will shed some light on why we have started this section with the correlogram. Let us start by rewriting $\hat{r}_x^b[m]$ as

$$\hat{r}_x^b[m] = \frac{1}{N}\sum_{n=m}^{N-1} x[n]x^*[n-m] = \frac{1}{N}\sum_{n=-\infty}^{\infty} x[n]x^*[n-m] = \frac{1}{N}\left(x[n] * x^*[-n]\right),$$

and taking its Fourier transform yields

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

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

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

which is the periodogram in (3.5). That is, the periodogram and the correlogram are identical.

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.

The first question is whether the periodogram is a biased estimator of the PSD, that is,

$$E\left[\hat{S}_x^p(e^{j\omega})\right] \stackrel{?}{=} S_x(e^{j\omega}).$$

To compute the bias of the periodogram, it is easier to consider the equivalence with the correlogram, which allows us to write

$$E\left[\hat{S}_x^p(e^{j\omega})\right] = E\left[\mathscr{F}(\hat{r}_x^b[m])\right] = \mathscr{F}\left(E\left[\hat{r}_x^b[m]\right]\right),$$

where we have used that the expectation and the Fourier transform are both linear operators. Thus, the periodogram is biased if $\hat{r}_x^b[m]$ is a biased estimate of the auto-correlation function. Now, using the definition of $\hat{r}_x^b[m]$, we get[2]

$$E\left[\hat{r}_x^b[m]\right] = E\left[\frac{1}{N}\sum_{n=m}^{N-1} x[n]x^*[n-m]\right] = \frac{1}{N}\sum_{n=m}^{N-1} E\left[x[n]x^*[n-m]\right]$$
$$= \frac{1}{N}\sum_{n=m}^{N-1} r_x[m] = \frac{N-|m|}{N} r_x[m] \neq r_x[m],$$

which shows that $\hat{r}_x^b[m]$ is indeed a biased estimate of the auto-correlation function, with the exception of $m = 0$, and makes the periodogram a biased estimate of the PSD.

It is possible to obtain a closed-form expression for the bias of the periodogram by noting that

$$E\left[\hat{r}_x^b[m]\right] = \frac{N-|m|}{N} r_x[m] = w_{T,N}[m] r_x[m], \tag{3.6}$$

where the triangular, or Barlett window, is defined as

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

and is depicted in Figure 3.5. The bias given in (3.6) shows us that the larger the m the larger the bias.

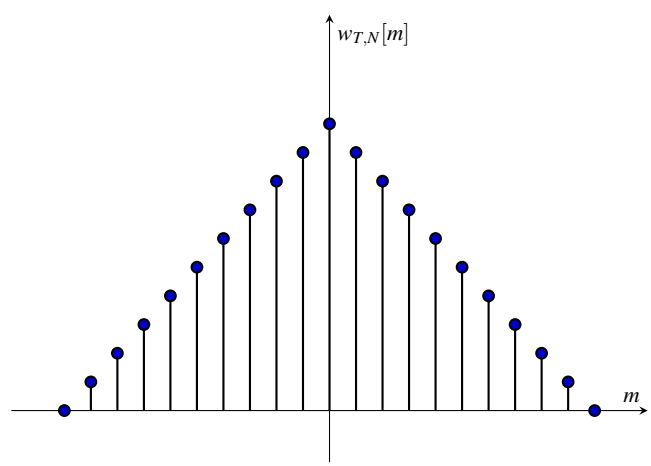


Fig. 3.5 Triangular window

Using (3.6), the bias of the periodogram becomes

$$E\left[\hat{S}_x^p(e^{j\omega})\right] = \mathscr{F}\left(w_{T,N}[m] r_x[m]\right) = \frac{1}{2\pi} W_{T,N}(e^{j\omega}) \circledast S_x(e^{j\omega}), \tag{3.7}$$

[2] It is easy to prove that $\hat{r}_x^u[m]$ is indeed an unbiased estimate of the auto-correlation function, which is left as an exercise for the student.

where

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

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.

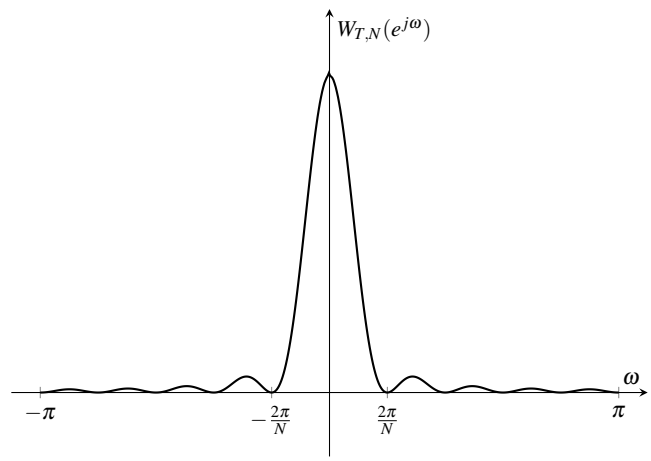


Fig. 3.6 Fourier transform of the triangular window

We have shown in (3.7) that the periodogram is biased. However, there are some cases when it is not. For instance, considering the asymptotic regime, it is unbiased:

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

Another case is when the process $x[n]$ is white noise. For this kind of processes, $r_x[m] = \sigma_x^2\delta[m]$ and, therefore

$$E\left[\hat{r}_x^b[m]\right] = \frac{N-|m|}{N} r_x[m] = \sigma_x^2\delta[m] = r_x[m].$$

Since the estimate of the auto-correlation is unbiased for white processes, it is easy to show that the periodogram is also unbiased.

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

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

where $\underset{\sim}{\propto}$ 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 $\hat{r}_x^b[m]$ 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 $\hat{r}_x^b[m]$ for m close to N. Thus, the BT estimator is

$$\hat{S}_x^{BT}(e^{j\omega}) = \mathscr{F}(w_M[m]\hat{r}_x^b[m]) = \sum_{m=-N+1}^{N-1} w_M[m]\hat{r}_x^b[m]e^{-j\omega m}, \tag{3.8}$$

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

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

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.

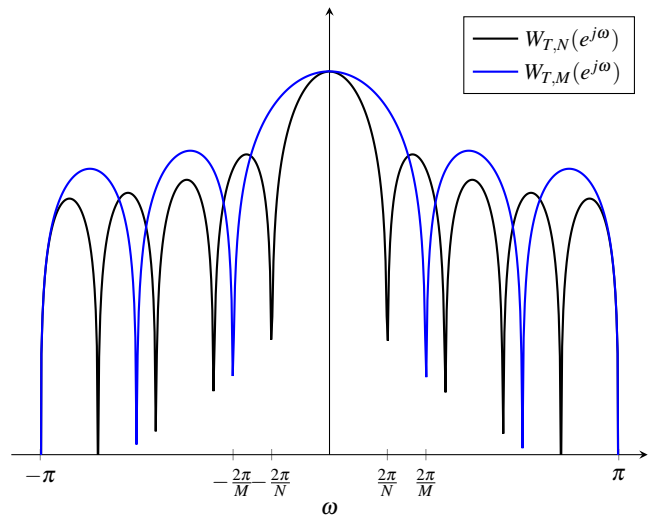


Fig. 3.7 Fourier transform (in logarithmic scale) of two triangular windows of different lengths

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

$$x_l[n] = x[(l-1)M + n],$$

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

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

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

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

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 (resolution), it is going to be smaller 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.

3.4.1 Rational models for parametric spectral estimation

These models consider a white Gaussian noise $u[n]$ with zero mean and variance σ^2 that goes through a causal and stable filter[3] $h[n]$ that has the following Fourier transform

$$H(e^{j\omega}) = \frac{B(e^{j\omega})}{A(e^{j\omega})} = \frac{\sum_{k=0}^{q} b_k e^{-j\omega k}}{1 + \sum_{k=1}^{p} a_k e^{-j\omega k}},$$

which implies that

$$x[n] = u[n] * h[n] = -\sum_{k=1}^{p} a_k x[n-k] + \sum_{k=0}^{q} b_k u[n-k]. \tag{3.9}$$

[3] A filter is said to be causal and stable if and only if all its poles are inside the unit circle.

For these models, the PSD is given by

$$S_x(e^{j\omega}) = S_u(e^{j\omega})|H(e^{j\omega})|^2 = \sigma^2 \left| \frac{\sum_{k=0}^{q} b_k e^{-j\omega k}}{1 + \sum_{k=1}^{p} a_k e^{-j\omega k}} \right|^2, \tag{3.10}$$

and we only have to estimate σ^2, $a_1, \ldots, a_p$, and $b_0, \ldots, b_q$.

According to Weierstrass theorem, for large values p and q, the PSD model in (3.10) can approximate arbitrarily close any continuous PSD. Hence, there is a strong interest in this kind of models, which are named as auto-regressive moving average (ARMA or ARMA(p,q)). There are two special cases of the ARMA model that are particularly interesting: the auto-regressive (AR or AR(p)) and the moving average (MA or MA(q)). For the former, the PSD is given by

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

whereas, for the latter, it is

$$S_x(e^{j\omega}) = \sigma^2 \left| \sum_{k=0}^{q} b_k e^{-j\omega k} \right|^2.$$

AR models are good choices if we suspect the PSD has large peaks and MA models are good choices for PSDs with large valleys.

The estimation of the model parameters is typically carried out in the time domain, for which the auto-correlation structure is required. Once the auto-correlation function is available, which will depend in general on the model parameters in a non-linear fashion, the estimation procedure consists in substituting the theoretical auto-correlation by an estimate and then solving a non-linear system of equations. The PSD estimate is obtained by substituting the estimated parameters in the corresponding model. This procedure, which is conceptually simple, is actually rather involved. However, there is an exception, which is the AR model since the dependency of auto-correlation on the parameters is linear.

3.4.2 The auto-correlation function of ARMA processes

This section computes the auto-correlation function of ARMA processes, which is defined as

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

Substituting $x[n]$ by (3.9), $r_x[m]$ becomes

$$
\begin{aligned}
r_x[m] &= E\left[\left(-\sum_{k=1}^{p} a_k x[n-k] + \sum_{k=0}^{q} b_k u[n-k]\right) x^*[n-m]\right] \\
&= -\sum_{k=1}^{p} a_k E\left[x[n-k]x^*[n-m]\right] + \sum_{k=0}^{q} b_k E\left[u[n-k]x^*[n-m]\right] \\
&= -\sum_{k=1}^{p} a_k r_x[m-k] + \sum_{k=0}^{q} b_k r_{ux}[m-k].
\end{aligned}
$$

where the cross-correlation function between $u[n]$ and $x[n]$ is

$$r_{ux}[m] = E\left[u[n]x^*[n-m]\right].$$

Now, taking into account that

$$x[n] = u[n] * h[n] = \sum_{l=-\infty}^{\infty} h[l]u[n-l],$$

the cross-correlation function becomes

$$
\begin{aligned}
r_{ux}[m] &= E\left[u[n]\sum_{l=-\infty}^{\infty} h^*[l]u^*[n-m-l]\right] \\
&= \sum_{l=-\infty}^{\infty} h^*[l]E\left[u[n]u^*[n-m-l]\right] \\
&= \sum_{l=-\infty}^{\infty} h^*[l]r_u[m+l], \\
&= \sum_{l=-\infty}^{\infty} h^*[-l]r_u[m-l], \\
&= r_u[m] * h^*[-m],
\end{aligned}
$$

where the auto-correlation of $u[n]$ is

$$r_u[m] = \mathbb{E}[u[n]u^*[n-m]] = \sigma^2\delta[m],$$

because it is a white process. Then, $r_{ux}[m]$ simplifies to

$$r_{ux}[m] = \sigma^2 h^*[-m],$$

and plugging $r_{ux}[m]$ into $r_x[m]$, the desired auto-correlation becomes

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

The second term in the right-hand side of the the above equation can be expanded as

$$\sigma^2 \sum_{k=0}^{q} b_k h^*[k-m] = \sigma^2\left(b_0 h[-m] + b_1 h[1-m] + \ldots + b_q h[q-m]\right),$$

and since the filter is causal ($h[m] = 0, \forall m < 0$), it becomes

$$\sigma^2 \sum_{k=0}^{q} b_k h^*[k-m] = \begin{cases} \sigma^2 \left(b_m h[0] + b_{m+1} h[1] + \ldots + b_q h[q-m]\right), & 0 \le m \le q, \\ 0, & m > q, \end{cases}$$
$$= \begin{cases} \sigma^2 \sum_{k=m}^{q} b_k h^*[k-m], & 0 \le m \le q, \\ 0, & m > q. \end{cases}$$

Putting all pieces together we get

$$r_x[m] = \begin{cases} -\sum_{k=1}^{p} a_k r_x[m-k] + \sigma^2 \sum_{k=m}^{q} b_k h^*[k-m], & 0 \le m \le q, \\ -\sum_{k=1}^{p} a_k r_x[m-k], & m > q, \\ r_x^*[-m], & m < 0. \end{cases} \tag{3.11}$$

Keeping in mind that $h[m]$ will depend on $a_1, \ldots, a_p$, and $b_0, \ldots, b_q$, it is easy to see in (3.11) that the relationship between the model parameters (σ^2, $a_1, \ldots, a_p$, and $b_0, \ldots, b_q$) and the auto-correlation is non-linear, which complicates tremendously the estimation of such parameters from an estimate of the auto-correlation function. This is shown in the following example for a particular ARMA model.

Example 3.3 (Auto-correlation function of an ARMA(1,1) process)

In this example, we will consider an ARMA(1,1) process, which has the following frequency response

$$H(e^{j\omega}) = \frac{1 - be^{-j\omega}}{1 - ae^{-j\omega}},$$

and the corresponding impulse response is

$$h[n] = a^n s[n] - ba^{n-1} s[n-1],$$

where

$$s[n] = \begin{cases} 1, & n \ge 0, \\ 0, & n < 0. \end{cases}$$

Now, we specialize (3.11) for the case of $p = 1, q = 1, b_0 = 1, b_1 = -b$, and $a_1 = -a$, which yields

$$\begin{aligned} r_x[0] &= a r_x[-1] + \sigma^2 \left(1 + (-b)(a-b)^*\right), \\ r_x[1] &= a r_x[0] + \sigma^2(-b), \\ r_x[2] &= a r_x[1], \\ r_x[3] &= a r_x[2], \\ &\vdots \end{aligned}$$

where we have taken into account that $h[0] = 1$ and $h[1] = a - b$. To recover the three model parameters, we need three equations, which are

$$\begin{bmatrix} r_x[0] & r_x[-1] \\ r_x[1] & r_x[0] \\ r_x[2] & r_x[1] \end{bmatrix} \begin{bmatrix} 1 \\ -a \end{bmatrix} = \begin{bmatrix} \sigma^2 \left(1 - b(a-b)^*\right) \\ \sigma^2(-b) \\ 0 \end{bmatrix}.$$

The first issue to solve the above system of equations is that we do not know $r_x[m]$, but as explained before it can be substituted by any estimator of the auto-correlation function, such as (3.2) or (3.3), yielding

$$\begin{bmatrix} \hat{r}_x[0] & \hat{r}_x^*[1] \\ \hat{r}_x[1] & \hat{r}_x[0] \\ \hat{r}_x[2] & \hat{r}_x[1] \end{bmatrix} \begin{bmatrix} 1 \\ -a \end{bmatrix} = \begin{bmatrix} \sigma^2 \left(1 - b(a-b)^*\right) \\ \sigma^2(-b) \\ 0 \end{bmatrix},$$

where we have used $\hat{r}_x[-1] = \hat{r}_x^*[1]$. The second issue is that the system of equations is non-linear, which makes it difficult to solve, even for this simple ARMA model.

3.4.3 AR processes

In the following, we will consider AR (= ARMA(p,0)) processes, which are the most commonly used ones among the three kind of processes studied in this course. There are several reasons. The first one is that the estimation of the parameters is much simpler. Actually, it can be done by simply solving a system of equations. Moreover, from the expression of an AR model

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

where we have assumed without loss of generality that $b_0 = 1$, 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 coefficient 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 $x[N], x[N+1], \ldots$

Let's now turn our attention to the estimation of the AR model parameters. Before proceeding, we shall require the impulse response of the system, which is given by

$$h[n] = x[n]\big|_{u[n]=\delta[n]} = \delta[n] - \sum_{k=1}^{p} a_k h[n-k],$$

and since the filter is causal, we find that $h[0] = 1$, and allows us to particularize (3.11) as follows

$$r_x[m] = \begin{cases} -\sum_{k=1}^{p} a_k r_x[m-k] + \sigma^2 \overbrace{h^*[0]}^{1}, & m = 0, \\ -\sum_{k=1}^{p} a_k r_x[m-k], & m > 0, \\ r_x^*[-m], & m < 0. \end{cases} \tag{3.12}$$

Equation (3.12) shows that the auto-correlation function of the AR model does not depend on $h[n]$, which is the term that introduces non-linear relationships. Since we need to obtain $p+1$ parameters, i.e., $a_1, \ldots, a_p$ and σ^2, we need $p+1$ equations, which are

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

The last p equations depend only on $a_1, \ldots, a_p$. Writing them in matrix form, we get

$$\begin{bmatrix} r_x[0] & r_x[-1] & \cdots & r_x[-p+1] \\ r_x[1] & r_x[0] & \cdots & r_x[-p+2] \\ \vdots & \vdots & \ddots & \vdots \\ r_x[p-1] & r_x[p-2] & \cdots & r_x[0] \end{bmatrix} \begin{bmatrix} -a_1 \\ -a_2 \\ \vdots \\ -a_p \end{bmatrix} = \begin{bmatrix} r_x[1] \\ r_x[2] \\ \vdots \\ r_x[p] \end{bmatrix},$$

which are known as the Yule-Walker equations. Hence, the filter coefficients are obtained by solving a linear system of equations

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

where

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

and

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

and we have used $r_x[-m] = r_x^*[m]$. 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 $\mathscr{O}(p^3)$ to $\mathscr{O}(p^2)$. 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 replace by an estimate.

3.4.4 MA processes

The last model considered in this course is MA (= ARMA(0,q)) processes, which have the following signal model

$$x[n] = \sum_{k=0}^{q} b_k u[n-k].$$

In this case, the impulse response of the system is

$$h[n] = b_n, \quad n = 0, \ldots, q,$$

and the auto-correlation in (3.11) becomes

$$r_x[m] = \begin{cases} \sigma^2 \sum_{k=m}^{q} h[k]h^*[k-m], & 0 \le m \le q, \\ 0, & m > q, \\ r_x^*[-m], & m < 0. \end{cases}$$

or, equivalently,

$$r_x[m] = \begin{cases} \sigma^2 h[m] * h^*[-m], & 0 \le m \le q, \\ 0, & m > q, \\ r_x^*[-m], & m < 0. \end{cases}$$

As in the ARMA model, the relationship between the auto-correlation function and the model parameters is non-linear, which is again shown in the following toy example.

Example 3.4 (Auto-correlation function of a MA(1) process)
In this example, we will consider a MA(1) process, which has the impulse response

$$h[n] = h[0]\delta[n] + h[1]\delta[n-1].$$

For this $h[n]$, it is easy to find that

$$h[m] * h^*[-m] = \begin{cases} |h[0]|^2 + |h[1]|^2, & m = 0 \\ h^*[0]h[1], & m = 1 \\ h[0]h^*[1], & m = -1 \\ 0, & |m| > 1. \end{cases}$$

and the auto-correlation becomes

$$
\begin{aligned}
r_x[0] &= \sigma^2 \left(|h[0]|^2 + |h[1]|^2\right), \\
r_x[1] &= \sigma^2 h^*[0] h[1], \\
r_x[2] &= 0, \\
&\vdots
\end{aligned}
$$

Then, we have a non-linear system with 2 equations and 3 parameters, which cannot be solved. One could be tempted to add the equation corresponding to $r_x[-1]$ but it would not help as it is not independent. The actual solution is easier and is based on the fact that there is an amplitude ambiguity. That is, we will get the same output for $u[n]$ and $h[n]$ and for $u[n]/K$ and $Kh[n]$. Thus, we can set one of the parameters to one, for instance, $h[0] = 1$, which yields

$$
\begin{aligned}
r_x[0] &= \sigma^2 \left(1 + |h[1]|^2\right), \\
r_x[1] &= \sigma^2 h[1].
\end{aligned}
$$

Dividing both equation we get

$$
\frac{r_x[0]}{r_x[1]} = \frac{1 + |h[1]|^2}{h[1]},
$$

which is a second order polynomial and, as a consequence, there are two solutions for $h[1]$. Finally, for each of these solution, the estimate of the variance becomes

$$
\hat{\sigma}^2 = \frac{r_x[1]}{\hat{h}[1]}.
$$

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 $P_H(1) = 1/6$.

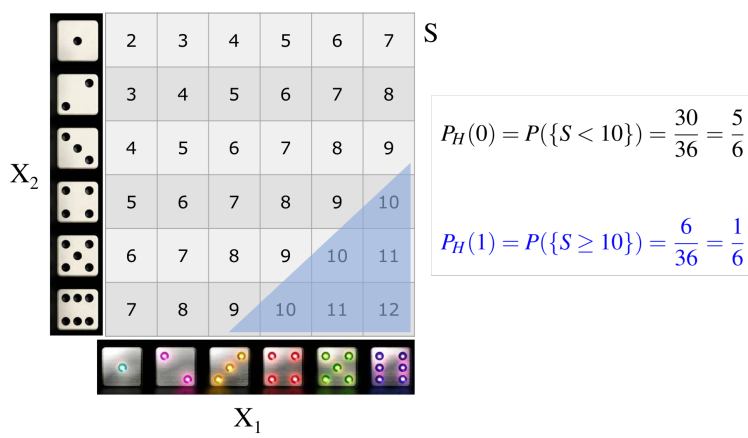


Fig. 4.1 All combinations of of X_1 and X_2 are equally probable, and therefore each of the 36 results represented in the figure have a probability of $1/36$. Counting the number of occurrences of particular values of S or H, the distribution of these variables can be calculated.

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

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

[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.

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 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.

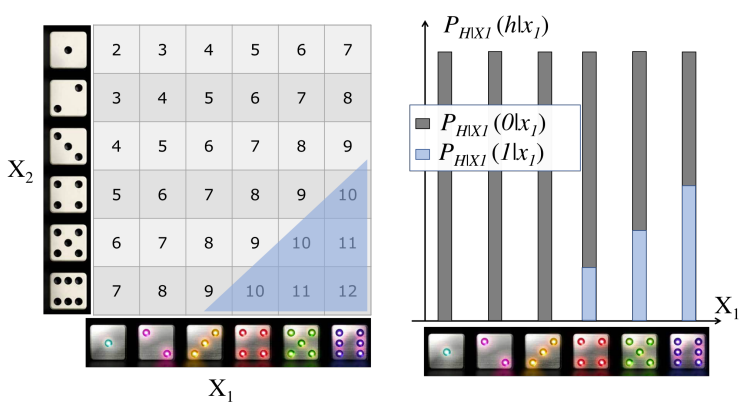


Fig. 4.2 To calculate posterior probabilities of the hypothesis, we need to count how many results in each column correspond to hypothesis 0 and how many correspond to hypothesis 1. Note that $P_{H|X_1}(0|x_1) + P_{H|X_1}(1|x_1) = 1$ for all values of X_1.

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.,

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

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

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 Detection Theory

Once we have presented some of the main concepts involved in detection problems through a series of examples, we are ready to formalize the theory for the case of two or more hypotheses.

4.2.1 Hypotheses-based problems

As we have already explained, 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.

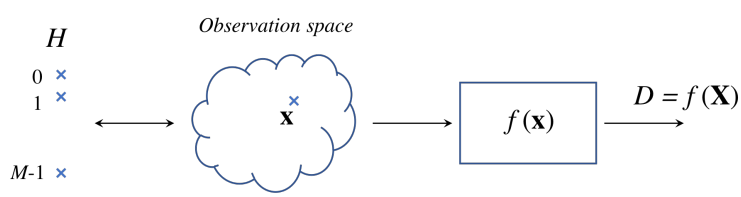


Fig. 4.3 Diagram block of hypothesis testing problems.

In this context, a detector or classifier 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. 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. Moreover, 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.

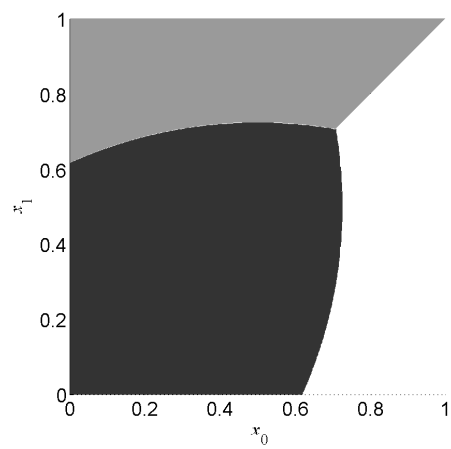


Fig. 4.4 Decision regions for the detector given in Example 4.2: $\mathscr{X}_0$ (black), $\mathscr{X}_1$ (grey), and $\mathscr{X}_2$ (white).

4.2.2 Statistical information involved in detection problems

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 Analysis of the detection performance

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):

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

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.

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}. 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.

4.3.2 Average (or mean) cost

Using the probability of error as the unique measure of the detector's performance is not a good idea, because there are many practical scenarios where not all errors are equally critical. 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,

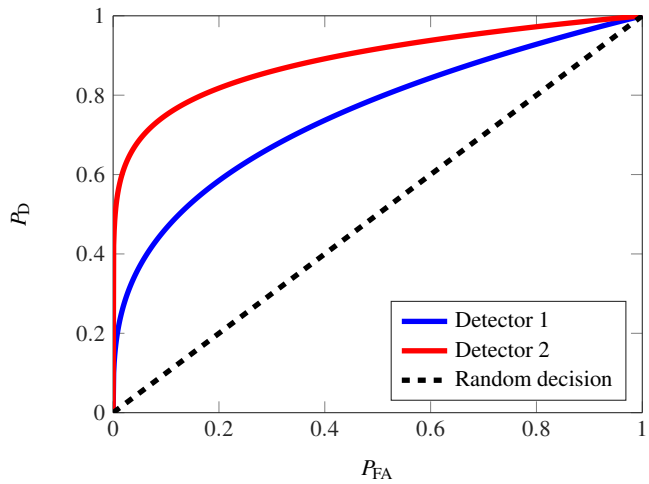


Fig. 4.5 ROC curves for two different detectors.

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 as well, and its expected value can be computed as

$$\mathbb{E}\{c_{DH}\} = \sum_{h=0}^{M-1}\sum_{d=0}^{M-1} c_{dh} P(D = d, H = h).$$

In order to compute the expected cost in terms of the probability distributions of $\mathbf{X}$ and H, we can use the relationship between D and $\mathbf{X}$, which yields

$$\begin{aligned}\mathbb{E}\{c_{DH}\} &= \sum_{h=0}^{M-1}\sum_{d=0}^{M-1} c_{dh}P(\mathbf{x}\in\mathscr{X}_d, H=h)\\ &= \sum_{h=0}^{M-1} P_H(h)\sum_{d=0}^{M-1} c_{dh}P(\mathbf{x}\in\mathscr{X}_d|H=h)\\ &= \sum_{h=0}^{M-1} P_H(h)\sum_{d=0}^{M-1} c_{dh}\int_{\mathbf{x}\in\mathscr{X}_d} p_{\mathbf{X}|H}(\mathbf{x}|h)d\mathbf{x}.\end{aligned}$$

Finally, we can also compute the expected cost conditioned on a given value of $\mathbf{x}$. 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. Consider, for instance, the computation of the mean cost for some value $\mathbf{x}$ belonging to $\mathscr{X}_d$. Thus, the expected cost is obtained as

$$\mathbb{E}\{c_{dH}|\mathbf{x}\} = \sum_{h=0}^{M-1} c_{dh}P_{H|X}(h|\mathbf{x}). \tag{4.2}$$

4.3.2.1 Binary case: Mean cost

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

$$\begin{aligned}\mathbb{E}\{c_{DH}\} &= c_{00}P(D=0,H=0)+c_{01}P(D=0,H=1)\\ &\quad + c_{10}P(D=1,H=0)+c_{11}P(D=1,H=1)\\ &= c_{00}P(D=0|H=0)P_H(0)+c_{01}P_{\mathrm{M}}P_H(1)\\ &\quad + c_{10}P_{\mathrm{FA}}P_H(0)+c_{11}P_{\mathrm{D}}P_H(1).\end{aligned}$$

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,

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

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)

$$\log\left[\frac{p_{\mathbf{X}|H}(\mathbf{x}|1)}{p_{\mathbf{X}|H}(\mathbf{x}|0)}\right] = \log p_{\mathbf{X}|H}(\mathbf{x}|1) - \log p_{\mathbf{X}|H}(\mathbf{x}|0) \underset{D=0}{\overset{D=1}{\gtrless}} 0, \tag{4.3}$$

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

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

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

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

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]

4.4.1.2 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

[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.

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

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

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

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.

4.4.2 Minimum expected cost 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 (4.2), which we reproduce here for convenience:

$$\mathbb{E}\{c_{dH}|\mathbf{x}\} = \sum_{h=0}^{M-1} c_{dh} P_{H|\mathbf{X}}(h|\mathbf{x}). \tag{4.4}$$

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.

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

$$\begin{aligned} d^\star &= \arg\min_d \sum_{h=0}^{M-1} c_{dh} P_{H|\mathbf{X}}(h|\mathbf{x}) \\ &= \arg\min_d \; c \sum_{h \neq d} P_{H|\mathbf{X}}(h|\mathbf{x}) \\ &= \arg\min_d \sum_{h \neq d} P_{H|\mathbf{X}}(h|\mathbf{x}) \\ &= \arg\min_d 1 - P_{H|\mathbf{X}}(d|\mathbf{x}) \\ &= \arg\max_d P_{H|\mathbf{X}}(d|\mathbf{x}) \\ &= d_{MAP}. \end{aligned} \tag{4.5}$$

4.4.2.1 Binary case: Minimum expected cost 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 (4.4) 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 The Gaussian case

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 (4.3) 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,

$$(\mathbf{x}-\mathbf{m}_0)^T\mathbf{V}_0^{-1}(\mathbf{x}-\mathbf{m}_0) - (\mathbf{x}-\mathbf{m}_1)^T\mathbf{V}_1^{-1}(\mathbf{x}-\mathbf{m}_1) \underset{D=0}{\overset{D=1}{\gtrless}} \mu \tag{4.6}$$

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 (4.6), 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 4.3 Figure 4.6 shows the decision boundaries for the ML detector ($\eta = 1$ in (4.6)), 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).

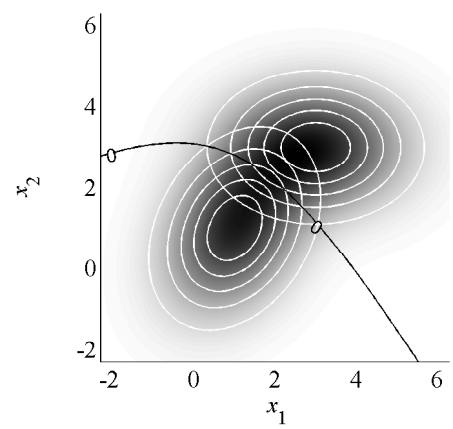


Fig. 4.6 Hyperbolic decision boundary of the ML detector and likelihoods for a Gaussian detection problem with 2D observations.

Example 4.4 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:

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

and

[3] We obtain the decision boundaries for the equality in (4.6).

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

In this case, the decision boundary is an ellipse.

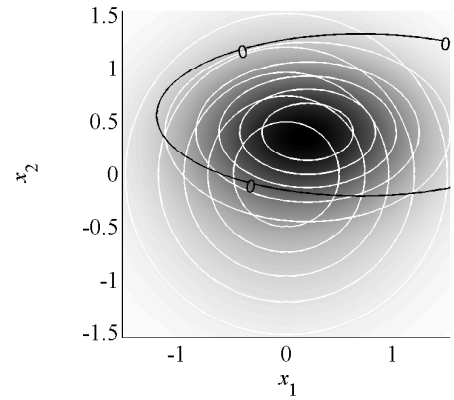


Fig. 4.7 Elliptic decision boundary of the ML detectors and likelihoods for a Gaussian detection problem with 2D observations.

4.4.3.1 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

$$(\mathbf{m}_1-\mathbf{m}_0)^T\mathbf{V}^{-1}\mathbf{x} \underset{D=0}{\overset{D=1}{\gtrless}} \tilde{\mu}, \tag{4.7}$$

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

Example 4.5 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.

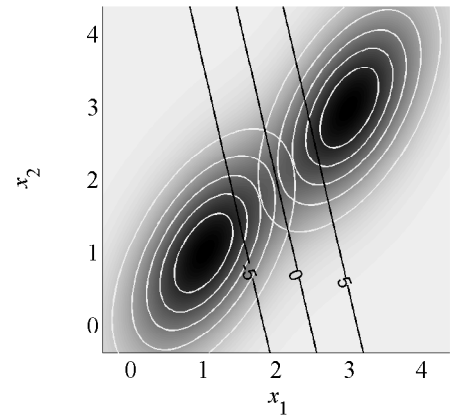


Fig. 4.8 Decision boundaries of the LLRT and likelihoods for a Gaussian detection problem with 2D observations and identical covariance matrices.

Example 4.6 (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

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

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.

4.4.3.2 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 4.7 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 $d = 1$.

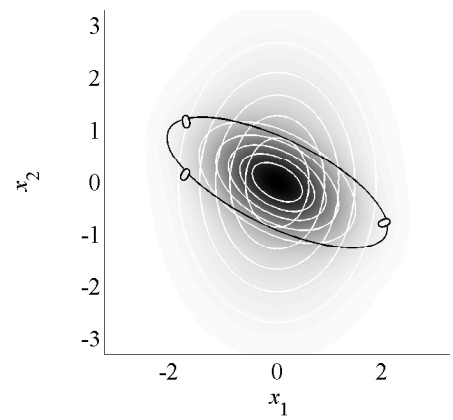


Fig. 4.9 Elliptic decision boundary for a 2D Gaussian problem with zero means.

Example 4.8 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.

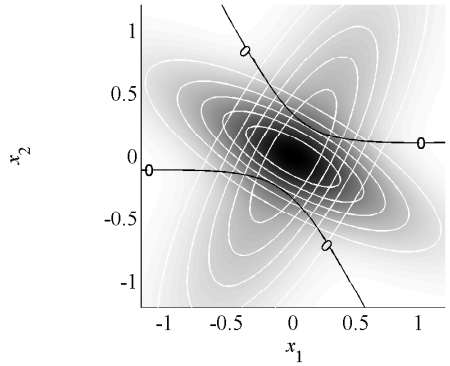


Fig. 4.10 Hyperbolic decision boundary for a 2D Gaussian problem with zero means.

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,

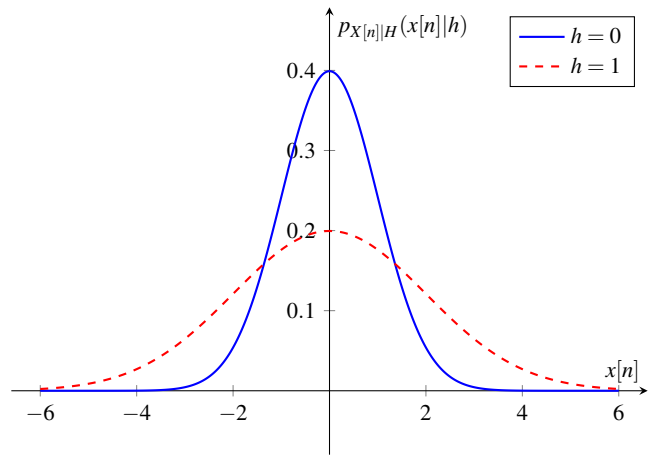


Fig. 5.1 Likelihoods considered in the introductory example

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,

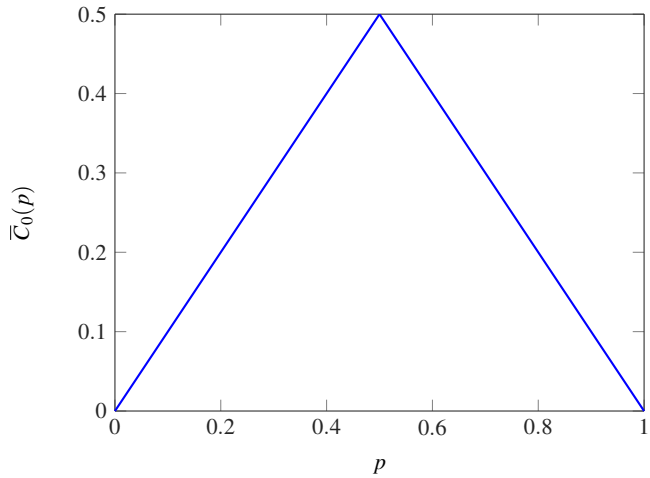


Fig. 5.2 Minimum average cost with no observations

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

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

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

where

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

and

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

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 $Z \sim \chi_I^2$. 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.

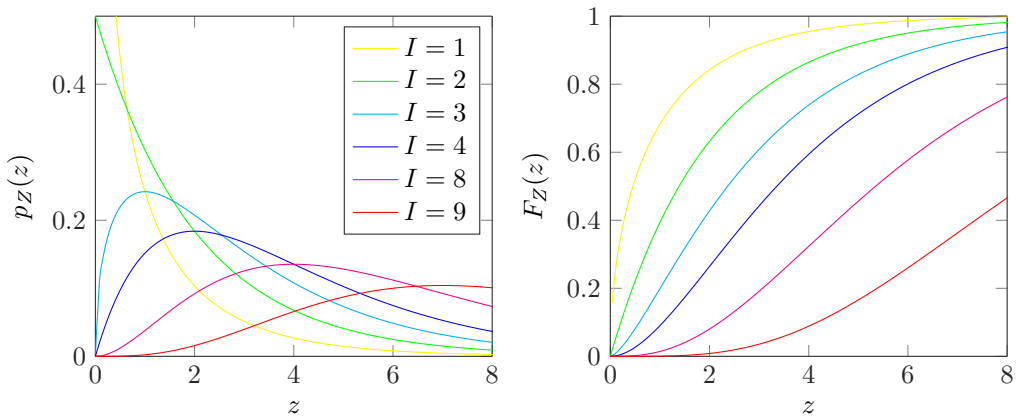


Fig. 5.3 Probability and cumulative density functions of a Chi-squared random variable with I degrees of freedom

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

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_n^2 random variable is larger than $8/3\,[\log(p/(1-p))+n\log 2]$. That is,

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

We can proceed similarly for P_{M} as follows

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

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_n^2 random variable is smaller than $2/3\,[\log(p/(1-p))+n\log 2]$, which can be expressed as

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

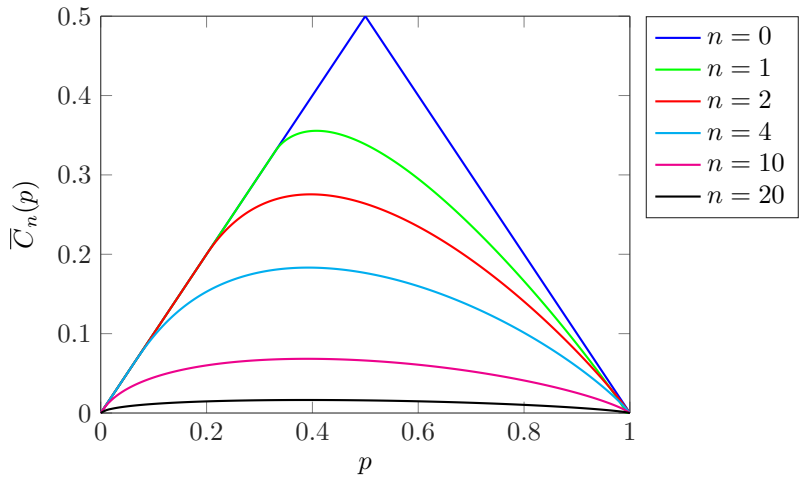


Fig. 5.4 Minimum average cost with n observations

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.

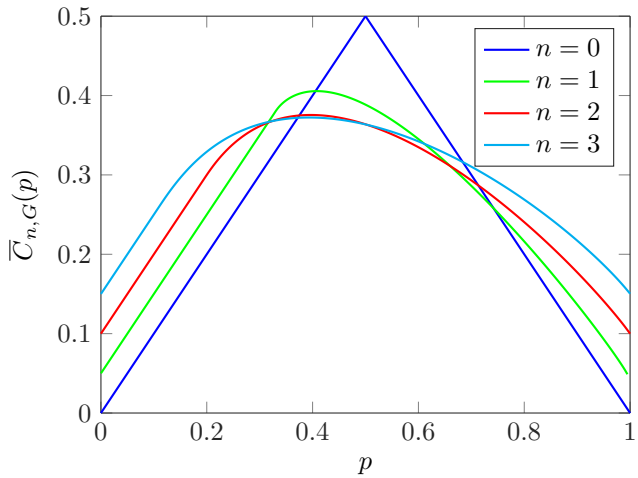


Fig. 5.5 Minimum average cost, including the gathering cost, with n observations

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, $\overline{C}_{n,G}(p) \ngeq \overline{C}_{n+1,G}(p), \forall p \in [0,1]$. 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.

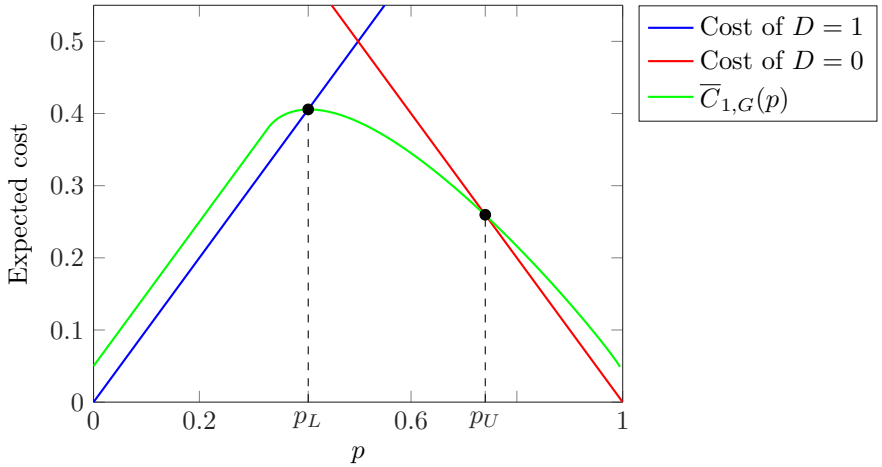


Fig. 5.6 Minimum average cost, including the gathering cost, with n observations

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]

$$\begin{aligned}&\text{decide } D = 1 \text{ for } p \le p_L,\\&\text{decide } D = 0 \text{ for } p \ge p_U,\\&\text{take another sample for } p_L < p < p_U.\end{aligned} \tag{5.4}$$

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.,

$$\begin{aligned}&\text{decide } D = 1 \text{ for } p_1 \le p_L,\\&\text{decide } D = 0 \text{ for } p_1 \ge p_U,\\&\text{take another sample for } p_L < p < p_U.\end{aligned}$$

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

$$\begin{aligned}&\text{decide } D = 1 \text{ for } p_n \le p_L,\\&\text{decide } D = 0 \text{ for } p_n \ge p_U,\\&\text{take another sample for } p_L < p < p_U,\end{aligned} \tag{5.5}$$

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.

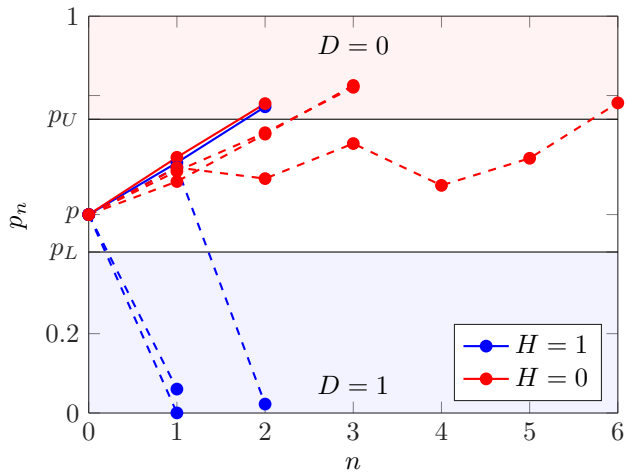


Fig. 5.7 Realizations of the sequential test

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

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 $P_{\text{FA}} \geq \alpha$ and $P_{\text{M}} \geq \beta$, 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

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

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

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

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

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

Since the probability of detection is

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

we get

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

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

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

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

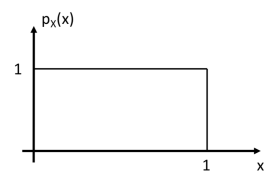
A.1 Change of Random Variable

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 $p_Y(y)$.

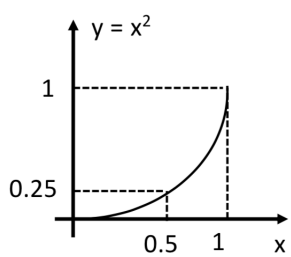
To understand how this new distribution or **change of random variable** is calculated, let's firstly solve a particular case:

- 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

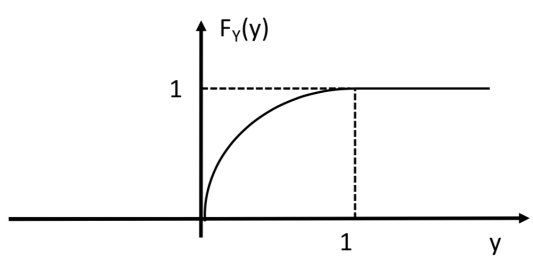
$$F_Y(y_0) = P\{Y < y_0\} = P\{X < \sqrt{y_0}\}$$

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

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

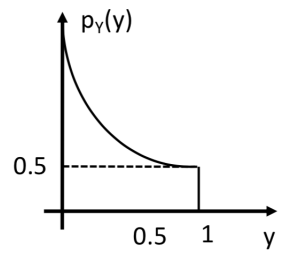
So, we have that

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



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:

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

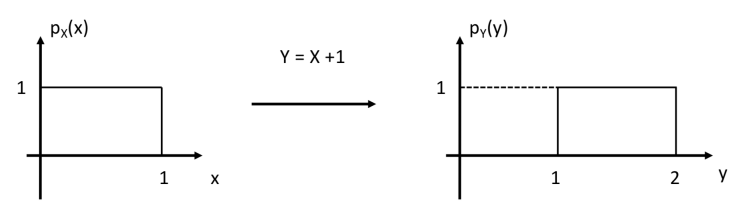
A.1.1 Some usual r.v. changes

The demonstration of these changes is left as homework.

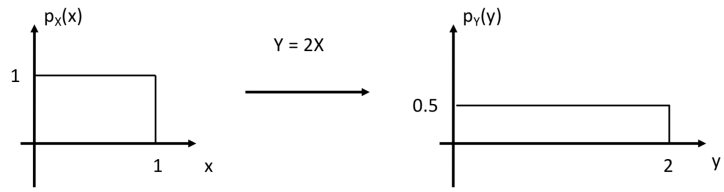
1. SHIFTING of R.V.
   $Y = X + a$, where a is a known constant. Then,

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

   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.