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

# Estimation and Detection Theory

Year 2019-20

February 16, 2020

Universidad Carlos III de Madrid

# Contents

# Chapter 1
# Statistical Estimation Theory

## 1.1 Introduction to the Estimation Problem

The contents of this lesson cover an introduction to the estimation problems. So, during this session, we will present some basic concepts of estimation design using statistical information. Important concepts, such as *a priori* and *a posteriori* probabilities, observations, cost functions, or likelihoods will be illustrated through a series of simple examples.

### 1.1.1 Example 1: Bayesian estimation without observations

**Problem 1.1** A food delivery company wants to develop a new service to estimate the time that will elapse from the reception of an order to its delivery to the customer's home. To do this, the total service or delivery time, $S$, is modelled as a random variable given by the sum of two independent r.v.s:

$$S = T_1 + T_2,$$

where $T_1$ models the time (in minutes) needed to prepare the order and $T_2$ is the shipping time (in minutes). Analyzing these times the company has characterized r.v.s $T_1$ and $T_2$ with the following probability distributions:

$$p_{T_1}(t_1) = 0.5\exp\left[-0.5(t_1 - 10)\right] \qquad t_1 > 10$$
$$p_{T_2}(t_2) = \frac{0.2}{r}\exp\left[-\frac{0.2}{r}(t_2 - 5)\right] \qquad t_2 > 5$$

where r is the distance (in kilometers) from the company store to the customer's home. To estimate the value of $S$, let's start solving the following questions:

a) Knowing the probability distributions of $T_1$ and $T_2$, can we obtain the probability distribution of $S$?
b) Knowing the probability distribution of $S$, can we estimate the total delivery time?
c) Which is the optimal estimator for a given cost?

**Solution 1.1** a) Can we obtain the probability distribution of $S$?

Computing the distribution of $S$ requires applying a change of random variable in which

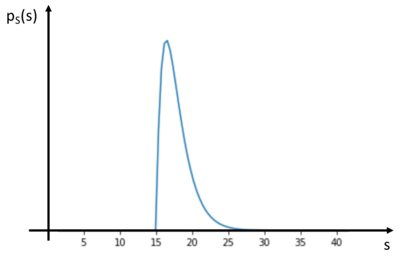


**Fig. 1.1** Representation of the probability distribution of $S$ for $r = 1$.

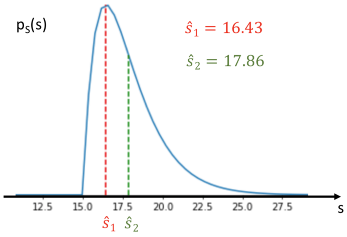


**Fig. 1.2** Some possibles estimators of $S$ analyzing $p_S(s)$.

we have to transform two random variables ($T_1$ and $T_2$) into a new variable ($S$). Since $S$ is the sum of two independent random variables, we know that their distribution will be given by the convolution of the distributions of $T_1$ and $T_2$:

$$p_S(s) = p_{T_1}(t_1 = s) * p_{T_2}(t_2 = s) = \int p_{T_1}(t_1 = s) p_{T_2}(t_2 = s - t_1) dt_1$$

after some mathematical manipulations (the complete mathematical development is left as an exercise), we can see that $p_S(s)$ is given by the following expression (see Figure 1.1):

$$p_S(s) = \left(0.5 + \frac{0.2}{r}\right)(s-15)\exp\left[-\left(0.5 + \frac{0.2}{r}\right)(s-15)\right] \quad s > 15$$

b) Can we now estimate the total delivery time?
Let's consider that we receive an order to be delivered to one kilometer of distance ($r = 1$Km), so we have that

$$p_S(s) = 0.7\,(s-15)\exp\left[-0.7\,(s-15)\right] \quad s > 15.$$

Knowing this distribution, we can know which values of $S$ are most probable and which values are completely unlikely; for instance, analyzing Figure 1.1, we can realize that it is quite likely that $S$ is between 15 and 20, whereas it is impossible that it is lower than 15, and it is almost impossible that it is greater than 30. So, in light of the distribution of $S$, we can estimate the total delivery time considering different criteria (see Figure 1.2):

- One could consider that a good estimation could be given by the most likely value of $S$, that is, by the mode of $S$:

$$\hat{s}_1 = \arg\max_s p_S(s)$$

and we can compute this value by deriving $p_S(s)$ and setting the result to zero:

$$\left.\frac{\partial p_S(s)}{\partial s}\right|_{s=\hat{s}_1} = 0.7\exp\left[-0.7\left(\hat{s}_1 - 15\right)\right] - 0.7^2\left(\hat{s}_1 - 15\right)\exp\left[-0.7\left(\hat{s}_1 - 15\right)\right] = 0$$

Now, we can cancel the term[1] $0.7\exp\left[-0.7\left(\hat{s}_1 - 15\right)\right]$ and get:

$$1 - 0.7(\hat{s}_1 - 15) = 0$$
$$\hat{s}_1 = 15 + \frac{1}{0.7} = 16.43 \text{ min.}$$

To complete this calculation, we have to check that this solution is a maximum (the derivative only guarantees returning relative extremes or saddle points). We can do this either analyzing the shape of $p_S(s)$ or checking that the second derivative is negative.

- Another possible estimation could be given by the expected value of $S$,

$$\hat{s}_2 = \mathbb{E}\{S\} = \int s p_S(s) ds = \int_{15}^{\infty} 0.7 s\left(s - 15\right)\exp\left[-0.7\left(s - 15\right)\right] ds$$

and solving this integral by parts, we have

$$\hat{s}_2 = 15 + \frac{2}{0.7} = 17.86 \text{ min.}$$

- Or we could even raise other estimators, such as the median of the distribution or the 25% or 75% percentiles.

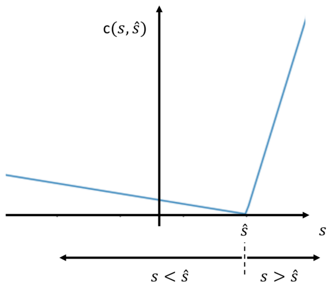
Finally, it is important to bear in mind that regardless of the estimator we use, we probably have an estimation error (it is practically impossible for the estimated value to coincide with the actual one) and the error of each estimator will indicate us which estimator is more adequate. In fact, in this unit we will pursue as a design criterion the minimization of the mean value of a cost criterion that establishes how we should penalize different kinds of errors.

c) How can we find the optimal estimator for a given cost? For the design of the estimator, the food delivery company wants to minimize the following cost function (see Figure 1.3):

$$c(\hat{s}, s) = \begin{cases} 0.005|s - \hat{s}| & \text{if} \quad \hat{s} > s \\ 0.1|s - \hat{s}| & \text{if} \quad \hat{s} < s \end{cases}$$

As any cost function, this cost function indicates how we have to penalize the fact that the estimated value differs from the actual value of $S$. However, unlike typical cost functions,

[1] Note that by cancelling this term we are skipping the solution $\hat{s}_1 \rightarrow \infty$, but analyzing the shape of $p_S(s)$ we can check that this root is a minimum.



$$c(\hat{s},s) = \begin{cases} 0.005|s-\hat{s}| & \text{if} \quad \hat{s} > s \\ 0.1|s-\hat{s}| & \text{if} \quad \hat{s} < s \end{cases}$$

**Fig. 1.3** Asymmetric cost function to be minimized during the estimator design.

this is an asymmetric cost function (see Figure 1.3), that is, it applies different penalties in case of overestimating the values of $S$ ($\hat{s} > s$) or underestimating them ($\hat{s} < s$). In this way, this cost function is indicating that if the order arrives before the time we have estimated it will penalize less than in case the customer has to be waiting longer than expected, i.e., if our order takes longer to arrive than we have estimated. In this case, using the mean or median of $S$ is not the most appropriate estimator, and we should choose a value higher than the expected one. I.e., since subestimations of the actual time of delivery are highly penalized, we should try to be conservative and produce estimators that are only rarely exceeded. So, it is possible that a value around the 70%-80% percentile of the $p_S(s)$ distribution is close to the estimator we are looking for.

Reviewing the expression of the cost function to be minimized, we can see that the cost value depends both on the estimator $\hat{s}$ and on the random variable to be estimated $S$. So, as the cost function is a function of a random variable, it is itself another random variable. For this reason, we are going to denote it as $C = c(\hat{s},S)$.

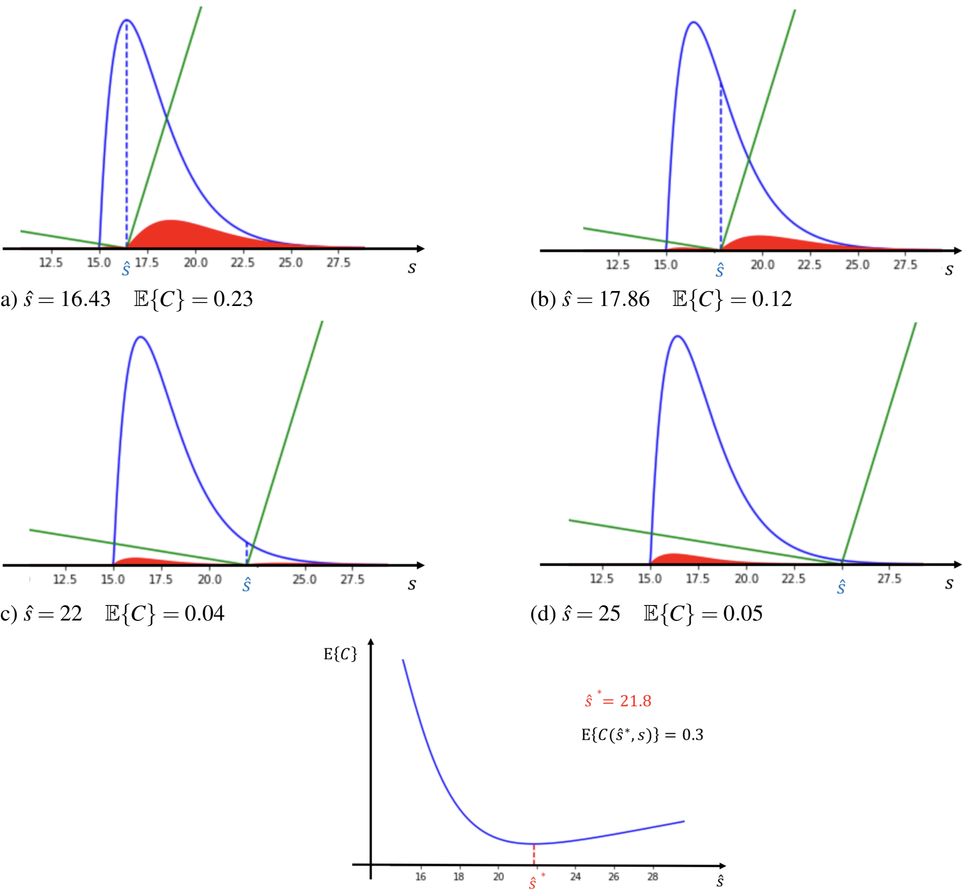
When we wanted to find the value of the estimator that minimizes the cost $C$, we would have to find the value of the estimator which minimizes the expected cost or the mean cost. So, the optimum estimator would be given by:

$$\hat{s}^* = \arg\min_{s} \ \mathbb{E}\{C\} = \arg\min_{\hat{s}} \ \mathbb{E}\{c(\hat{s},S)\}$$

where the mean cost is computed as:

$$\mathbb{E}\{c(\hat{s},S)\} = \int c(\hat{s},s)p_s(s)ds$$

For each possible value of the estimator, we will get a different mean cost, and we will have to select the estimator value that provides the minimum mean cost. Fig 1.4 shows the average cost for different values of the estimator for the given asymmetric cost function. In fact, Subfigures 1.4(a)-(d) show how the mean cost is computed for different values of the estimator; note that the mean cost is computed as the area resulting from multiplying the distribution $p_S(s)$ by the cost function $c(\hat{s},S)$, and, as different values of the estimator, $\hat{s}$, will place the cost function in different positions, this process will result in different mean costs. If we directly compute the mean cost for any possible value of $\hat{s}$ we would obtain a curve similar to that shown in Subfigure 1.4(e) and $\hat{s}^*$ would be the



(a) $\hat{s} = 16.43 \quad \mathbb{E}\{C\} = 0.23$

(b) $\hat{s} = 17.86 \quad \mathbb{E}\{C\} = 0.12$

(c) $\hat{s} = 22 \quad \mathbb{E}\{C\} = 0.04$

(d) $\hat{s} = 25 \quad \mathbb{E}\{C\} = 0.05$

(e) Evolution of the mean cost with the value of $\hat{s}$

**Fig. 1.4** Process of minimization of the mean cost for different values of the estimator.

value of $\hat{s}$ which minimizes it. In this case, we can see that the optimum estimator would be $\hat{s}^* = 21.8$ min and it generates a mean cost of 0.3.

### 1.1.2 Example 2: Bayesian estimation with observations

**Problem 1.2** To obtain a more accurate estimation of the delivery time, the company has improved the food preparation process, so that it is able to know the exact time it will take to prepare the order $T_1 = t_1$.

When we want to estimate the value of $S$ without observations, the only distribution which provides information about the value of $S$ is $p_S(s)$; however, when we have additional information such as knowledge of the value of $t_1$ (observation), including this information in our estimation problem makes the estimation of $S$ easier (more accurate). Adding this knowledge (observation) to the estimation task implies using the posterior distribution of $S$, $p_{S|t_1}(s|t_1)$, instead of using $p_S(s)$.

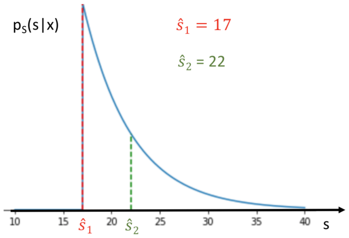


**Fig. 1.5** Some possible estimators of $S$ given that $t_1 = 12$.

To solve the estimation problem in this new scenario, let's try to answer the following questions:

a) Can we obtain the probability distribution of $S$ given the value $T_1 = t_1$?
b) Can we estimate the total delivery time for a given value $T_1 = t_1$?
c) Given a cost function to be optimized, which is the optimal estimator for a given value $T_1 = t_1$?

**Solution 1.2** a) Can we obtain the probability distribution of $S$ given the value $T_1 = t_1$?
The calculation of the posterior distribution of $S$ can be done by applying the following r.v change[2]:

$$S = t_1 + T_2$$

so, $p_{S|t_1}(s|t_1)$ can be obtained by shifting the distribution of $p_{T_2}(t_2)$ to the position of $t_1$, i.e.,

$$p_{S|t_1}(s|t_1) = p_{T_2}(t_2 = s - t_1) = \frac{0.2}{r}\exp\left[-\frac{0.2}{r}(s - t_1 - 5)\right] \qquad s > t_1 + 5$$

b) Can we estimate the total delivery time for a given value $T_1 = t_1$?
To answer this question, let's consider that a customer is calling the food company to place an order from a distance of one kilometer ($r = 1$Km), and in this moment and for this order the preparation time is $t_1 = 12$ minutes. So we have:

$$p_{S|t_1}(s|t_1) = p_{T_2}(t_2 = s - t_1) = 0.2\exp[-0.2(s - 17)] \qquad s > 17$$

and, examining this distribution (see Figure 1.2), we can consider different estimators such as the maximum of the distribution, which is $\hat{s}_1 = 17$ min., or the expected value of $S$ given $t_1 = 12$, which can be computed (by solving the integral by parts) as:

$$\hat{s}_2 = \mathbb{E}\{S|t_1 = 12\} = \int s p_{S|t_1}(s|t_1)ds = 17 + \frac{1}{0.2} = 22\text{min}.$$

For any value of the observation, the posterior distribution of $S$ will change (in this case, the $p_{S|t_1}(s|t_1)$ will be shifted) and the value of the estimator will depend on the observation value ($t_1$). If we want to obtain a general expression for these estimators (for

[2] Note that as we are calculating the distribution of $S$ given $T_1$, the value of $T_1$ is known ($T_1 = t_1$).

any value of $t_1$), we can directly compute both the maximum and the mean by using the expression of the posterior for any value of $t_1$ (we are still considering $r = 1$):

$$p_{S|t_1}(s|t_1) = 0.2\exp\left[-0.2(s - t_1 - 5)\right] \qquad s > t_1 + 5$$

For example:

- If we consider that the mode of $p_{S|t_1}(s|t_1)$ could be an adequate estimator, the estimator will be:

  $$\hat{s}_1 = \arg\max_{s} p_{S|t_1}(s|t_1)$$

  In this case, as $p_{S|t_1}(s|t_1)$ is a decreasing function for $s > t_1 + 5$, its maximum is

  $$\hat{s}_1 = t_1 + 5.$$

- We can also consider that the expected value of $S$ given $t_1$ is a good estimator. In this case (computing the integral by parts):

  $$\hat{s}_2 = \mathbb{E}\{S|t_1\} = \int s p_{S|t_1}(s|t_1) ds = 5 + t_1 + \frac{1}{0.2} = 10 + t_1$$

However, in order to decide which estimator is best, we need, as before, to define which cost function we want to minimize.

c) Given a cost function to be optimized, which is the optimal estimator for a given value $T_1 = t_1$?
Again, when we are designing an estimator we may want to design it in such a way that it minimizes the mean value of a given cost function. Now, as we are now working with observations, our goal will be finding the optimal estimator for any observed value (for any given value of $T_1 = t_1$).
Thus, we can now find the optimum estimator by

$$\hat{s}^* = \arg\min_{s}\ \mathbb{E}\{c(\hat{s}, S)|t_1\}$$

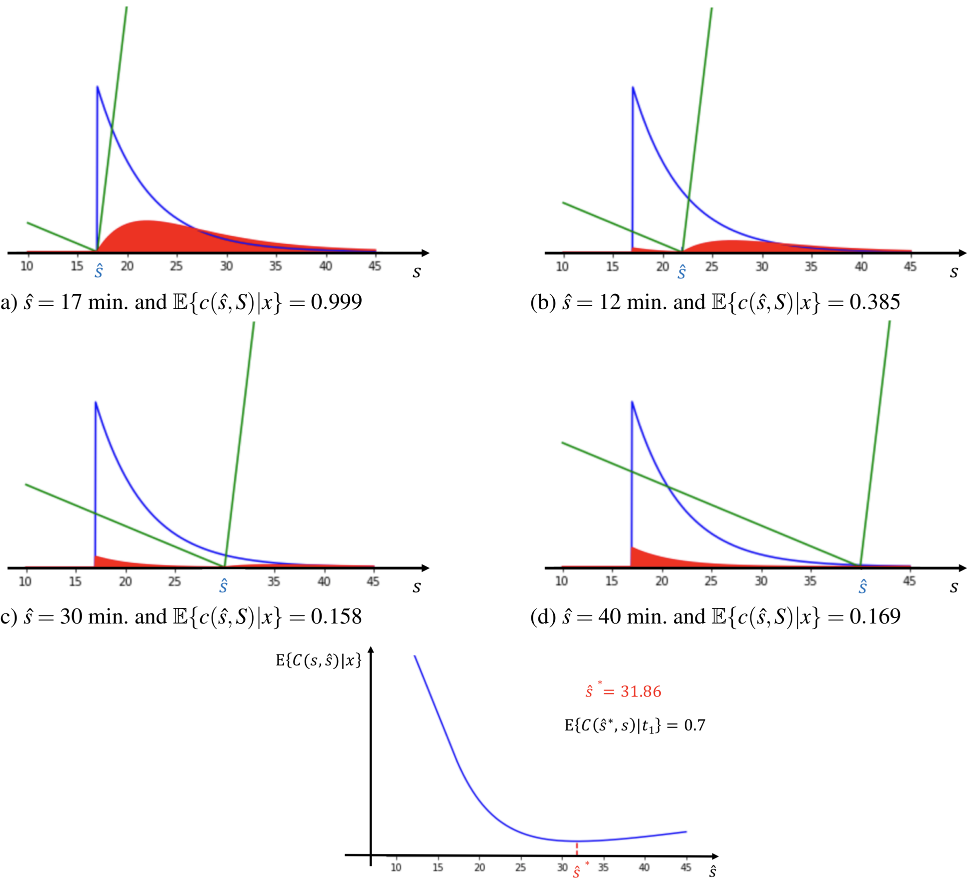
where

$$\mathbb{E}\{c(\hat{s}, S)|t_1\} = \int c(\hat{s}, s) p_{S|t_1}(s|t_1) ds$$

Once again, each possible value of the estimator will provide a different mean cost for a given value of $t_1$ and our goal will be to select the estimator that minimizes said mean cost. Summarizing our example problem:

- An order is placed to be shipped to a distance of one kilometer ($r = 1$Km);
- For this order the preparation time is $t_1 = 12$ minutes;
- We want to minimize the asymmetric cost function used in Problem 1.1.1(c);

Fig 1.6 shows the procedure of minimizing the mean cost given $t_1 = 12$. Subfigures 1.6(a)-(d) plot the conditional mean costs for different values of the estimator and Subfigure 1.6(e) illustrates the mean cost as a function of $\hat{s}$. Analyzing these figures, we can check that the optimum estimator value is $\hat{s}^* = 31.86$ min and it generates a mean cost (given that $t_1 = 12$) of 0.7.



(a) $\hat{s} = 17$ min. and $\mathbb{E}\{c(\hat{s},S)|x\} = 0.999$

(b) $\hat{s} = 12$ min. and $\mathbb{E}\{c(\hat{s},S)|x\} = 0.385$

(c) $\hat{s} = 30$ min. and $\mathbb{E}\{c(\hat{s},S)|x\} = 0.158$

(d) $\hat{s} = 40$ min. and $\mathbb{E}\{c(\hat{s},S)|x\} = 0.169$

(e) Evolution of the mean cost given $t_1$ with the value of $\hat{s}$

**Fig. 1.6** Process of minimization of the mean cost given $t_1$ for different values of the estimator.

Finally, it is important to note that the involved mean cost is for a value of $t_1 = 12$. If we wanted to compare this cost with that incurred by the estimator designed without observations, we would have to compute the optimum estimator and its mean cost (given the observation) for any value of $t_1$ and average these costs taking into account the probability of each $t_1$ value.

## 1.2 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.

### 1.2.1 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. 1.7, 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.

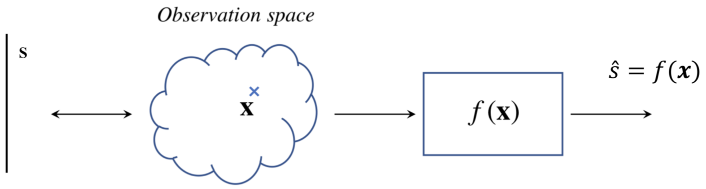


**Fig. 1.7** Diagram block of estimation problems.

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.

### 1.2.2 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:

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

### 1.2.3 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[3] 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

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

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

$$p_{S,X}(s,x) = \left[\begin{array}{ll} \frac{1}{x}, & 0 < s < x \text{ and } 0 < x < 1 \\ 0, & \text{otherwise} \end{array}\right. \tag{1.3}$$

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 $w$,

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

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 $\hat{S}_2$.

[3] 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

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

## 1.3 Design of estimators

### 1.3.1 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:

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

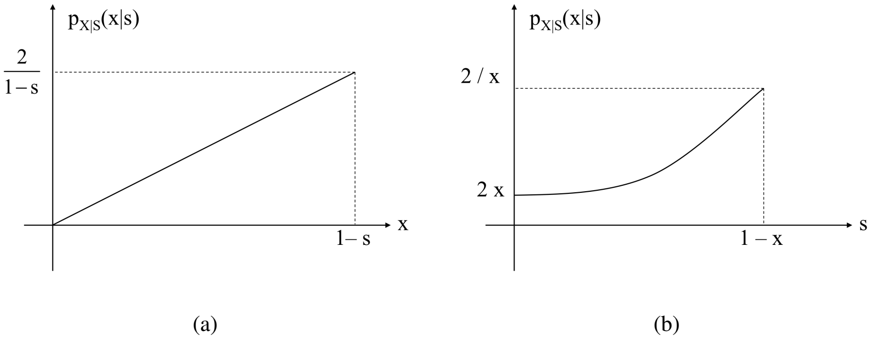


**Fig. 1.8** Representation of the likelihood distribution of the exercise 1.3 as a function of $x$ and $s$.

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 1.8(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.1.8(b))[4]. 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

[4] 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

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

Thus,

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

### 1.3.2 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

$$p_{S|X}(s|x) = \begin{bmatrix} \frac{1}{x}, & 0 < s < x < 1 \\ 0, & \text{otherwise} \end{bmatrix} \tag{1.22}$$

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.

## 1.4 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.

### 1.4.1 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 1.9 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.

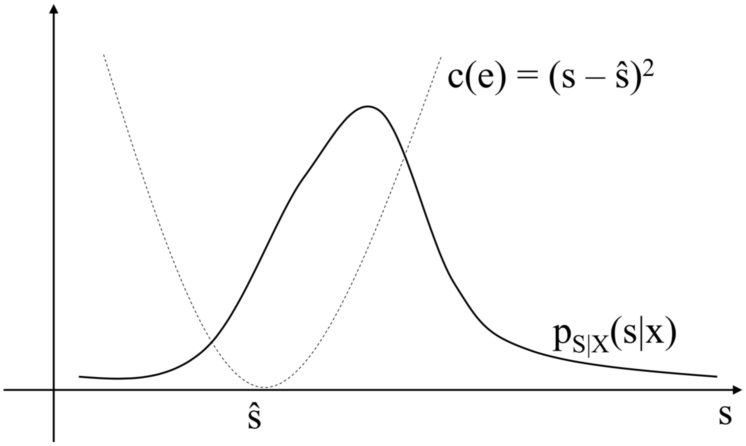


**Fig. 1.9** Graphical representation of the process of calculating the posterior mean for a generic value $\hat{s}$.

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

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

which coincides with (1.25).

### 1.4.2 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. 1.10). This representation also suggests the convenience of splitting the integral into two parts corresponding to the two slopes of the cost function:

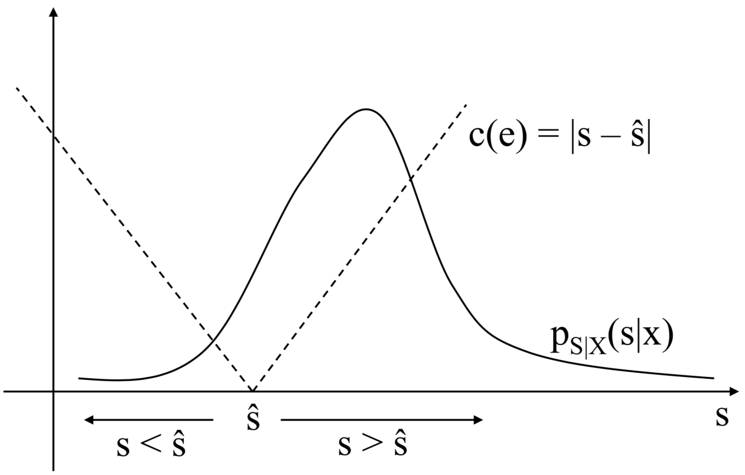


**Fig. 1.10** Graphical representation of the process of calculating the posterior mean absolute error for a generic value $\hat{s}$.

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

The fundamental theorem of calculus[5] 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 $\hat{s}_{\mathrm{MAD}}$ represents the minimum of the mean cost, the previous derivative must be cancelled for the estimator, verifying that $F_{S|\mathbf{X}}(\hat{s}_{\mathrm{MAD}}|\mathbf{x}) = 1/2$. In other words, the absolute minimum error estimator is given by the median of $p_{S|\mathbf{X}}(s|\mathbf{x})$:

$$
\boxed{\hat{s}_{\mathrm{MAD}} = \mathrm{median}\{S|\mathbf{X}=\mathbf{x}\} \qquad (1.34)}
$$

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

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

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

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

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,

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

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.

## 1.5 Estimation with constrains

### 1.5.1 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:

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

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[6]. 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 1.5.2, 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

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

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 $w$.

$$\begin{aligned}\mathbb{E}\{c(S, w\mathbf{X}^2)\} &= \int_{\mathbf{x}} \left(\frac{1}{3}x^2 - wx^3 + w^2x^4\right) d\mathbf{x} \qquad (1.38) \\ &= \frac{1}{9} - \frac{1}{4}w + \frac{1}{5}w^2 \qquad (1.39)\end{aligned}$$

[6] 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: $\hat{s} = \frac{5}{8}x^2$.

### 1.5.2 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.

#### 1.5.2.1 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:

$$\mathbf{w}_e^* = \arg\min_{\mathbf{w}_e} \mathbb{E}\{(S-\hat{S})^2\} = \arg\min_{\mathbf{w}_e} \mathbb{E}\{(S-\mathbf{w}_e^T\mathbf{X}_e)^2\} \qquad (1.44)$$

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

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

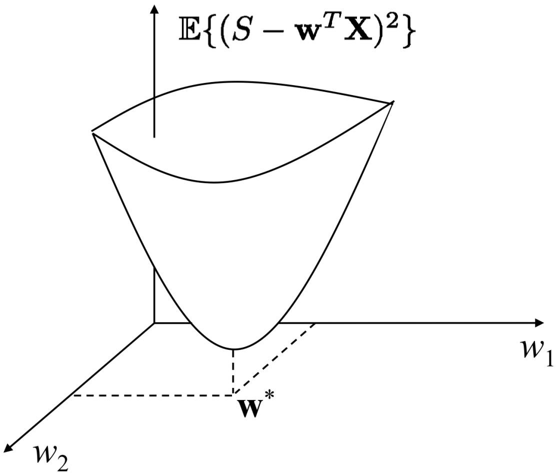


**Fig. 1.11** Surface of the mean squared error of the linear estimator as a function of the estimator weights.

Figure 1.11 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[7]:

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

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 $\mathbf{X}_e$) with $N+1$ unknowns (the components of $\mathbf{w}_{e^*}$).

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

$$\mathbb{E}\{S\mathbf{X}_e\} = \mathbb{E}\{\mathbf{X}_e(\mathbf{X}_e^T\mathbf{w}_e^*)\} \qquad (1.46)$$

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

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:

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

#### 1.5.2.2 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:

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

#### 1.5.2.3 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:

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

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

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).

### 1.5.2.4 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 ($\hat{S}_{\mathrm{MMSE}}$), 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 $\hat{S}_{\mathrm{LMSE}}$ and leaving the result in function of the mathematical expectations of the involved random variables:

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

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

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

**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: $[X_1,X_2]=[3,1]$.

## 1.6 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.

### 1.6.1 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 $\rho$ 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 $2 \times 2$.

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:

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

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.

$$\boxed{\hat{s}_{\text{MSE}} = m_{S|X} = \frac{\rho}{v_X} x \qquad (1.76)}$$

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

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

*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 (1.12) represents the situation described for a case with $v_S < v_R$.

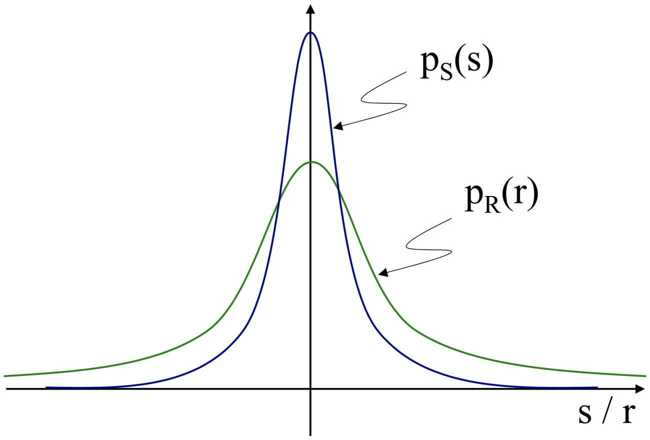


**Fig. 1.12** Estimation of Gaussian random variable $S$ contaminated by Gaussian noise $R$.

According to (1.76), for the resolution of the problem we must find the variance of $X$ and the covariance between $S$ and $X$ ($\rho$). 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:

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

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

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

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 $\hat{s}_{\mathrm{MSE}} \to x$, 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, $\hat{s}_{\mathrm{MSE}} \to 0$.

### 1.6.2 Case with multidimensional variables

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

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

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

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

Since the MMSE estimator of **S** given **X** is precisely the posterior mean, we can write

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

This estimator expression is simplified when **S** and **X** have zero means, resulting in

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

### 1.6.3 Linear estimation and Gaussian estimation

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

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

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.

## 1.7 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

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

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.,

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

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

## 1.8 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 8$x$8 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, $\alpha$, 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, $A_{\text{min}}$, 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 1.3**

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)

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

(note the expression above is not the joint pdf of $Z$ and $S$)

$$p_{\mathbf{X}}(\mathbf{x}) = \int p_{\mathbf{X},S}(\mathbf{x},s)\,ds = \int_0^\infty s^K\ \exp[-s(z+1)]\,ds$$

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 $\sigma_\varepsilon^2$, giving rise to an observation $x[n]$. That is, the corresponding entries are

$$\begin{aligned} x[n] &= u[n] * s[n] + \varepsilon[n] && (2.1)\\ &= u[n]s[0] + u[n-1]s[1] + \ldots + u[n-M+1]s[M-1] + \varepsilon[n]. && (2.2) \end{aligned}$$

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

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

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:

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

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

which will allow 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

$$p(x[n]|\mathbf{s}) = \mathcal{N}(x[n]|\mathbf{u}[n]^\top\mathbf{s}, \sigma_\varepsilon^2), \tag{2.6}$$

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

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

$$p(\mathbf{x}|\mathbf{s}) = \prod_{n=0}^{N-1} \mathcal{N}(x[n]|\mathbf{u}[n]^\top\mathbf{s}, \sigma_\varepsilon^2) = \mathcal{N}(\mathbf{x}|\mathbf{U}^\top\mathbf{s}, \sigma_\varepsilon^2\mathbf{I}). \tag{2.7}$$

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

$$\begin{aligned} \hat{\mathbf{s}}_{\mathrm{ML}} &= \underset{\mathbf{s}}{\operatorname{argmax}}\ p(\mathbf{x}|\mathbf{s}) = \underset{\mathbf{s}}{\operatorname{argmax}}\ \log p(\mathbf{x}|\mathbf{s}) \\ &= \underset{\mathbf{s}}{\operatorname{argmin}}\ \frac{1}{2}(\mathbf{x}-\mathbf{U}^\top\mathbf{s})^\top(\sigma_\varepsilon^2\mathbf{I})^{-1}(\mathbf{x}-\mathbf{U}^\top\mathbf{s}) + \frac{1}{2}\log|\sigma_\varepsilon^2\mathbf{I}| + \frac{N}{2}\log(2\pi) \\ &= \underset{\mathbf{s}}{\operatorname{argmin}}\ ||\mathbf{x}-\mathbf{U}^\top\mathbf{s}||^2 && (2.8) \\ &= (\mathbf{U}\mathbf{U}^\top)^{-1}\mathbf{U}\mathbf{x}. && (2.9) \end{aligned}$$

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 **s**, equalizing to zero and clearing.

## 2.4 Bayesian Solution

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

$$p(\mathbf{s}) = \mathscr{N}(\mathbf{s}|\mathbf{0}, \sigma_s^2\mathbf{I}), \tag{2.10}$$

since it considers acceptable any set of real coefficients, and assumes that these have a null mean and a dispersion set by $\sigma_s^2$. 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 **s** and **x**. A simple way to do this is to observe that

$$\begin{bmatrix}\mathbf{s}\\ \mathbf{x}\end{bmatrix} = \begin{bmatrix}\mathbf{I}\\ \mathbf{U}^\top\end{bmatrix}\mathbf{s} + \begin{bmatrix}\mathbf{0}\\ \varepsilon\end{bmatrix} \text{ with } \varepsilon = [\varepsilon[0], \ldots, \varepsilon[N-1]]^\top, \tag{2.11}$$

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 $\sigma_\varepsilon^2$ and, thus, it is jointly Gaussian. The computation of the mean and the variance of $[\mathbf{s}^\top\ \mathbf{x}^\top]^\top$ is straightforward:

$$\begin{bmatrix}\mathbf{s}\\ \mathbf{x}\end{bmatrix} = \mathscr{N}\left(\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix}, \begin{bmatrix}\sigma_s^2\mathbf{I} & \sigma_s^2\mathbf{U}\\ \sigma_s^2\mathbf{U}^\top & \sigma_s^2\mathbf{U}^\top\mathbf{U} + \sigma_\varepsilon^2\mathbf{I}\end{bmatrix}\right) \tag{2.12}$$

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

$$p(\mathbf{s}|\mathbf{x}) = \mathscr{N}(\mathbf{s} \mid \sigma_s^2\mathbf{U}(\sigma_s^2\mathbf{U}^\top\mathbf{U} + \sigma_\varepsilon^2\mathbf{I})^{-1}\mathbf{x},\ \ \sigma_s^2\mathbf{I} - \sigma_s^2\mathbf{U}(\sigma_s^2\mathbf{U}^\top\mathbf{U} + \sigma_\varepsilon^2\mathbf{I})^{-1}\mathbf{U}^\top\sigma_s^2), \tag{2.13}$$

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

$$p(\mathbf{s}|\mathbf{x}) = \mathscr{N}(\mathbf{s} \mid \mathbf{P}\mathbf{U}\mathbf{x},\ \ \sigma_\varepsilon^2\mathbf{P}), \tag{2.14}$$

where

$$\mathbf{P} = (\mathbf{U}\mathbf{U}^\top + \tfrac{\sigma_\varepsilon^2}{\sigma_s^2}\mathbf{I})^{-1} \tag{2.15}$$

. Thus the MMSE and MAP estimates of **s** are:

$$\hat{\mathbf{s}}_{\text{MMSE}} = \hat{\mathbf{s}}_{\text{MAP}} = \hat{\mathbf{s}}_{\text{MAD}} = \mathbf{P}\mathbf{U}\mathbf{x} \tag{2.16}$$

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 (2.9).

### 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. (2.12), 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

$$x_* = \mathbf{u}_*^\top \mathbf{s} + \varepsilon_* \tag{2.17}$$

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

$$p(x_*|\mathbf{x}) = \mathcal{N}(x_* \mid \mathbf{u}_*^\top \mathbf{P}\mathbf{U}\mathbf{x},\ \ \sigma_\varepsilon^2 + \sigma_\varepsilon^2 \mathbf{u}_*^\top \mathbf{P}\mathbf{u}_*) \tag{2.18}$$

. that inmediatly provides the following estimators for $x_*$:

$$\hat{x}_{*\mathrm{MMSE}} = \hat{x}_{*\mathrm{MAP}} = \hat{x}_{*\mathrm{MAD}} = \mathbf{u}_*^\top \mathbf{P}\mathbf{U}\mathbf{x} = \mathbf{u}_*^\top \hat{\mathbf{s}}_{\mathrm{MMSE}}. \tag{2.19}$$

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 $\mathbf{s}_{\mathrm{MMSE}}$.

## 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. (2.9) or (2.16) 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 $\mathcal{O}(M^3)$ that would have a "naive" method that simply recalculates everything again every time a sample arrives.

### 2.5.1 Bayesian solution

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

$$\mathbf{P}_N = (\mathbf{U}\mathbf{U}^\top + \tfrac{\sigma_\varepsilon^2}{\sigma_s^2}\mathbf{I})^{-1}, \tag{2.20}$$

$$\mathbf{r}_N = \mathbf{U}\mathbf{x} \tag{2.21}$$

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

$$\hat{\mathbf{s}}_{\mathrm{ML}} = \underset{\mathbf{s}}{\operatorname{argmax}}\ p(\mathbf{x}|\mathbf{s}) = \underset{\mathbf{s}}{\operatorname{argmin}}\ ||\mathbf{x}-\mathbf{U}^\top\mathbf{s}||^2 \tag{2.22}$$

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

Notice that

$$||\mathbf{x}-\mathbf{U}^\top\mathbf{s}||^2 = \sum_{n=0}^{N-1}(x[n]-\mathbf{u}[n]^\top\mathbf{s})^2, \tag{2.23}$$

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

$$\hat{\mathbf{s}}_{N+1} = \hat{\mathbf{s}}_N + \mu\left(x[n]-\mathbf{u}[n]^\top\hat{\mathbf{s}}_N\right)\mathbf{u}[n], \tag{2.24}$$

where $\mu$ 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:

$$\mathbf{s}_{\text{Wiener}} = \mathbf{R}_{uu}^{-1}\mathbf{r}_{ux}, \tag{2.25}$$

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

# Appendix A
# Transformations of random variables

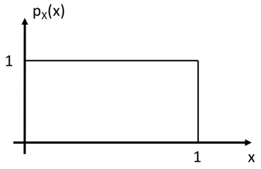
## A.1 Change of Random Variable

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 $p_Y(y)$.

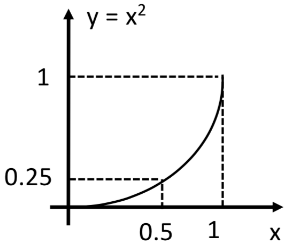
To understand how this new distribution or **change of random variable** is calculated, let's firstly solve a particular case:

- $X$ 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

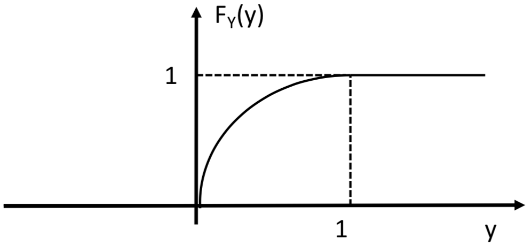
$$F_Y(y_0) = P\{Y < y_0\} = P\{X < \sqrt{y_0}\}$$

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

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

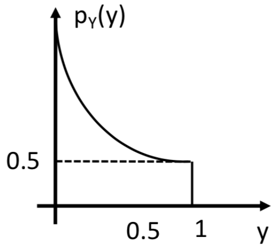
So, we have that

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



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

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

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

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

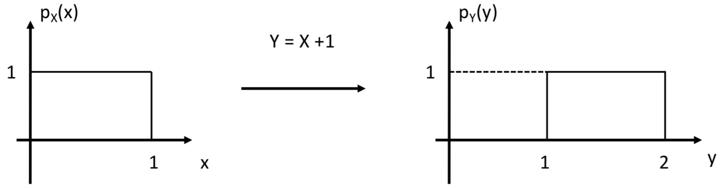
### A.1.1 Some usual r.v. changes

The demonstration of these changes is left as homework.

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

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

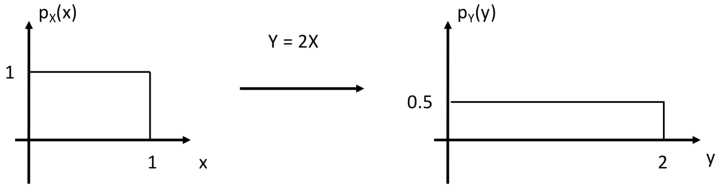
   when we are adding a constant to any r.v., we are shifting the distribution from the origin to the position of the constant



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.


p_X(x)
1
1
x
Y = 2X
p_Y(y)
0.5
2
y