Statistical Signal Processing
5CTA0

Simona Turco and Franz Lampel

Q1 2022/2023

Contents

Part I

Random variables and random signals

1

Introduction

Compared to *deterministic* signals, which can be described univocally by a mathematical formula or a well-defined rule, *stochastic* signals cannot be predicted with certainty at any given time. The uncertainty may come from several different sources; it may be due to measurement noise, interference, or artifacts; or sometimes a mathematical description may exist, but it is so complex that is not practically useful. Although the difference between deterministic and stochastic signals may be subtle at times, due to measurement or numeric noise, almost any real-world or synthetic signal can be considered a random signal. The question may then arise: how do we process such signals? This part of the course provides the basic tools necessary to answer this question.

The topics covered in this part are:

- **Probability and random variables**: The term "stochastic" comes from the Greek word for chance. To understand stochastic signals, a review of probability theory is necessary.
- **Random processes and random signals**: When coping with uncertainty, a deterministic signal can better be treated as a random signal. Here we discuss the statistical tools to characterize random signals.
- **Rational signal models**: Can we generate signals with desired statistical properties? Rational signal models provide an efficient representation for suitable types of random signals.

2

Probability and random variables

2.1 Introduction

Most of the fundamental tools in signal processing, e.g., the rules of convolution, the Z- and Fourier transforms, and digital filter designs, are based on the framework of deterministic signal processing. However, when uncertainty is involved, deterministic approaches are not suitable. Uncertainty is almost always present in the world around us. A very simple example where uncertainty plays an important role is the weather forecast. The weather forecast makes a prediction of the weather in the days ahead. However, this prediction does not guarantee that the predicted event actually takes place. Probability theory provides the mathematical tools to reason when uncertainty is present.

Before discussing the fundamental concepts of probability theory, a set of important terms used in probability theory are provided:

- An **experiment** is a procedure that can be repeated infinitely many times with an underlying model that defines which outcomes will be likely to be observed;
- An **observation** (or a **trial**) is one realization of the experiment;
- The **outcome** of the experiment is any possible observation of the experiment;
- The **sample space**, denoted by $\mathcal{S}$, is the set of all possible outcomes;
- An **event** is a set of outcomes of an experiment, which can be the sample space or a subset of the sample space;
- Two events are called *disjoint* or *mutually exclusive* if these sets of outcomes do not have common outcomes;
- If an event is an empty set of outcomes it is a **null event**, which is denoted by $\emptyset$;
- The **event space** is a set of disjoint events together forming the sample space.

In order to get some intuition about the practical meaning of these definitions, we turn to the following example.

Example 2.1

Suppose we are flipping two coins and observing which sides of the coins land on top. In this case, the experiment is the flipping of the two coins, where the top side of both coins is observed. The underlying model behind the experiment is based on the fact that we assume fair coins, meaning that the probability of a coin landing heads is equal to the probability of a coin landing tails.

An observation or trial is flipping both coins just once, with the outcome characterized by the top sides of the coins. Let us write this corresponding outcome using two letters indicating the top sides of both coins respectively, where we use h to indicate heads and t to indicate tails. A possible outcome of the trial is for example ht. On the contrary, h or hth are not possible outcomes of this experiment, which involves flipping two coins.

The sample space for this particular experiment is defined as

$$\mathcal{S} = \{hh, ht, th, tt\}, \tag{2.1}$$

which is written in the set notation.

Let us define an event A, which is the set of all possible outcomes where the first coin is heads, and an event B, which is the set of all possible outcomes where the second coin is tails. Both events can be written in the set notation as $A = \{hh, ht\}$ and $B = \{ht, tt\}$.

The events A and B are not disjoint, because both share the outcome ht. An example of a null set for this experiment is the event when a coin lands with a blank side facing upwards. This face is not defined in our experiment and therefore this outcome cannot be observed, leading to an empty set.

A possible event space is the set given by the events A and B, where A resembles the event that the first coin lands head and where B resembles the event that the first coin lands tails. Both events do not contain similar outcomes, i.e. $A = \{hh, ht\}$ and $B = \{tt, th\}$, but together they form the entire sample space.

2.2 Sets of outcomes

In Example 2.1, we already saw that we could write our events as sets of outcomes. A set can be regarded as a group or collection of elements. A set is denoted with curly brackets $\{\cdot\}$ which enclose all elements in that particular set. These sets can also be visually represented in the form of Venn diagrams, as shown in Figure 2.1. The sample space $\mathcal{S}$ containing all possible outcomes is represented by a square. The event A represents a set of possible outcomes and is a subset of the sample space.

Figure 2.1 introduces several set operators. The complement of a set A is denoted by A^C and is a set containing all outcomes of the sample space excluding the outcomes in A. The intersection operator $\cap$ denotes the intersection between two sets. In $A \cap B$ the intersection contains all outcomes that are both in sets A and B. The union operator $\cup$ denotes the union between two sets. In $A \cup B$ the union contains all outcomes that are in A, B, and both A and B. A subset, which contains a part of a larger set, is denoted as $B \subset A$ (read B is a subset of A). Lastly, as defined previously, two events are disjoint if these sets of outcomes do not have common outcomes.

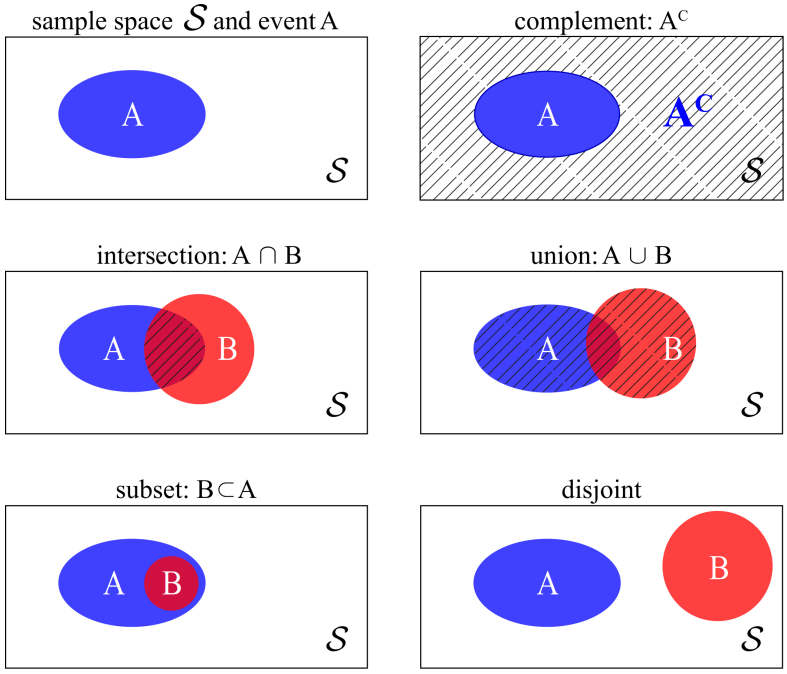


Figure 2.1: Venn diagrams showing the relations between different sets. The shaded areas represent the results of the different operations.

Example 2.2

Ricardo's offers customers two kinds of pizza crust, Roman (R) and Neapolitan (N). All pizzas have cheese but not all pizzas have tomato sauce. Roman pizzas can have tomato sauce or they can be white (W); Neapolitan pizzas always have tomato sauce. It is possible to order a Roman pizza with mushrooms (M) added. A Neapolitan pizza can contain mushrooms or onions (O) or both, in addition to the tomato sauce and cheese. Draw a Venn diagram that shows the relationship among the ingredients N, M, O, T, and W in the menu of Ricardo's pizzeria.

Solution.

At Ricardo's, the pizza crust is either Roman (R) or Neapolitan (N). To draw the Venn diagram as shown below, we make the following observations:

- The set R, N is a partition so we can draw the Venn diagram with this partition;
- Only Roman Pizza's can be white. Hence $W \subset R$;
- Only a Neapolitan pizza can have onions. Hence $O \subset N$;
- Both Neapolitan and Roman pizzas can have mushrooms so that event M straddles the R, N partition;
- The Neapolitan pizza can have both mushrooms and onions so $M \cap O$ cannot be empty;

- The problem statement does not preclude putting mushrooms on a white Roman pizza. Hence the intersection $W \cap M$ should not be empty.

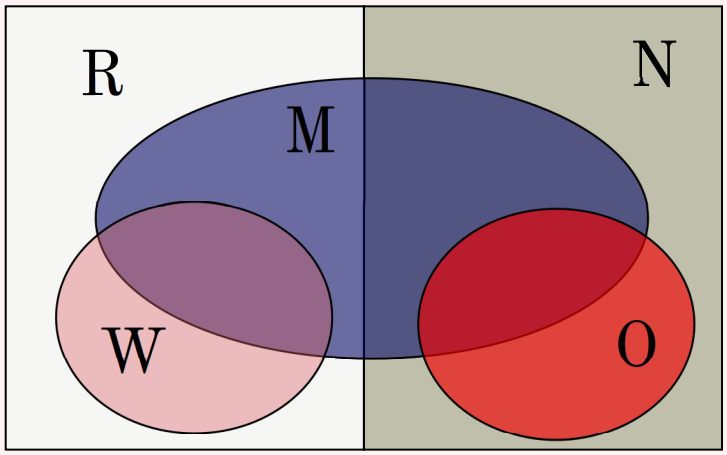


Figure 2.2: Solution to the exercise in the form of a Venn diagram.

2.3 Classic and frequentist definitions of probability

The concept of probability is related to randomness, or chance. Often we relate probability to what we do not know. Taking the example of flipping a coin, if we had all possible information about this experiment, such as the position of the coin, the force applied to the coin, the wind condition, the weight of the coin, the angle between the coin and our finger, the distance between the hand and the landing surface, the smoothness of the landing surface, the turbulence of the air, etc., we might be able to predict the outcome of the coin flip. However, the laws governing this experiment might be so complex and the number of variables so large, that in practice is not useful; we thus prefer to deal with the uncertainty. When we talk about a random experiment, we mean that our knowledge about the experiment is limited and therefore we cannot predict with absolute certainty its outcome. Probability theory provides us with a framework to describe and analyze random phenomena.

The **classic definition of probability** results from the works of Bernoulli and Laplace. The latter defined the probability by stating "the probability of an event is the ratio of the number of cases favorable to it, to the number of all cases possible when nothing leads us to expect that any one of these cases should occur more than any other, which renders them, for us, equally possible".

Formally, if a random experiment can result in N mutually exclusive and equally likely outcomes, and if event A results from the occurrence of N_A of these outcomes, then the probability of A is defined as

$$\Pr[A] = \frac{N_A}{N}. \tag{2.2}$$

This definition of probability is closely related to the *principle of indifference*, which states that, in absence of relevant evidence, all possible outcomes are equally likely.

In contrast, the **frequentist definition of probability**, also known as relative probability, calculates the probability based on how often an event occurs. The relative frequency of an event is given by

$$f_A = \frac{\text{number of occurrences of event } A}{\text{total number of observations}} = \frac{N(A)}{N}, \tag{2.3}$$

where $N(\cdot)$ denotes the number of occurrences of a certain event and N the total number of observations. The relative frequency can be understood as how often event A occurs relative to all observations. Theoretically, an infinitely large number of observations is needed to obtain the true probability. This leads to the frequentist definition of probability, given by

$$\Pr[A] = \lim_{N\to\infty} f_A = \lim_{N\to\infty} \frac{N(A)}{N}. \tag{2.4}$$

A low number of total observations does not give a good estimate of the true underlying probability of an event occurring.

2.4 Axioms and laws of probability

2.4.1 Probability Axioms

From the definition of probability, three important probability axioms can be determined. Axioms are statements that are regarded as true and can therefore be used to prove other statements. The probability axioms are:

- For any event A, it holds that $0 \leq \Pr[A] \leq 1$;
- It holds that $\Pr[\mathcal{S}] = 1$;
- For any countable collection of M disjoint events it holds that $\Pr[A_1 \cup A_2 \cup \ldots \cup A_M] = \Pr[A_1] + \Pr[A_2] + \ldots + \Pr[A_M]$.

Let us now discuss these axioms one by one. The first axiom states that a probability of an event is always between 0 and 1, where 0 means that there is no chance that the event will take place and 1 means that it will certainly happen; negative probabilities do not exist. Taking the frequentist definition of probability as a means to understand this, it is in fact not possible for an event to occur a negative number of times and therefore a negative probability cannot exist. Similarly, a probability larger than 1 would mean that a certain event would occur more often than all events together. Again this is not possible and therefore we are restricted to the probability bounds set by the first axiom.

The second axiom states that the probability of observing an outcome that is in the sample space $\mathcal{S}$ is always equal to 1. This axiom arises from the definition of the sample space. The sample space was defined previously as the set of all possible outcomes. Therefore we can conclude that an observation is always part of this set and thus the probability of observing an outcome that is part of the sample space equals 1.

The third axiom states that we may add the probabilities of separate events if we want to calculate the probability of the union of these events, under the constraint that the sets are all disjoint. Figure 2.1 gives an intuitive explanation of why this holds. When the union of multiple disjoint events is calculated, there is no overlap (meaning no common outcomes) between these events. Therefore the total probability does not need to be compensated for overlap and we can simply add the probabilities of the separate events.

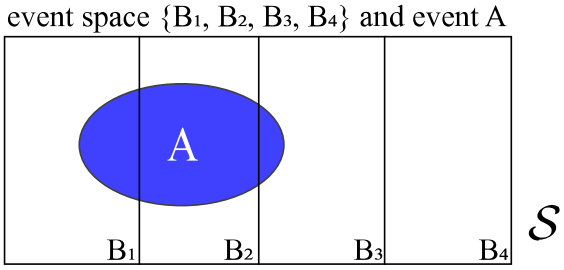


Figure 2.3: Visualization of a sample space, separated in an event space $\{B_1, B_2, B_3, B_4\}$, with an event A that can be split up in different segments.

Consequences of the probability axioms

From the previous axioms, several consequences can be determined. These include:

1. It holds that $\Pr[\emptyset] = 0$;
2. It holds that $\Pr[A^C] = 1 - \Pr[A]$;
3. For any events A and B it holds that $\Pr[A \cup B] = \Pr[A] + \Pr[B] - \Pr[A \cap B]$;
4. If $A \subset B$ it holds that $\Pr[A] \leq \Pr[B]$;
5. For any event A and event space $\{B_1, B_2, \ldots, B_m\}$ it holds that $\Pr[A] = \sum_{i=1}^{m} \Pr[A \cap B_i] = \sum_{i=1}^{m} \Pr[AB_i]$.

In 5., $\Pr[AB_i]$ indicates the probability of both events A and B_i occurring, that is the intersection. This also often indicated as $\Pr[A, B_i]$. So $\Pr[A \cap B_i] = \Pr[AB_i] = \Pr[A, B_i]$.

The first consequence is rather straightforward. The probability of observing an outcome that is in the null event equals 0, because it reflects the chance that we observe nothing.

The second consequence can be understood again through Figure 2.1, where the definition of the complement plays an important role. The complement of an event A includes all outcomes in the sample space except for all outcomes of event A. Since axiom 2 indicates that the probability of observing any outcome equals 1, the sets A and A^C together make up the entire sample space and therefore their probabilities should add up to one.

The third consequence is a generalization of Axiom 3 and holds for all events, so not only for disjoint events. The axiom can be understood by analyzing Figure 2.1. The union of two overlapping events can be written as their sum whilst compensating for the overlapping set of outcomes, denoted by the intersection of events. Therefore, the probability of the union of events A and B can be written as the sum of the individual probabilities minus the probability of the overlapping event. Axiom 3 is a special case of this axiom, where two events are disjoint and therefore the intersection between the two events equals 0.

The fourth consequence specifies that the probability of an event A is smaller or equal to the probability of an event B when A is a subset of B. This is an immediate consequence of the definition of a subset, where the event A contains a part of the outcomes of event B. Equality only occurs if the sets are equal. The last consequence can be explained with the help of Figure 2.3. The event A can be split into multiple subsets, each in a separate region of the event space, denoted by the intersection between event A and the subset B_i. Adding all different segments of A (all intersections between A and B_i) gives the full event A, because the event space always covers the entire sample space and must therefore include the entire set of A.

If we have enough information on an experiment and its associated sample space, we can calculate the probability of an event by using the probability axioms.

Example 2.3

Let us take a look at an experiment consisting of rolling a die twice and the event of obtaining both times heads, i.e., $A = \{hh\}$. By the frequentist definition of probability, we should repeat the experiment infinitely many times to calculate this probability. How can we reach the same conclusion without having to repeat the experiment hundreds of times? First, we gather the information we have on the experiment. We know that the sample space is given by $\mathcal{S} = \{hh, ht, th, tt\}$. We also know that all the events in the sample space are disjoint and have equal probability. Thus, we can use the axiom of probability to write

$$\begin{aligned} 1 &= \Pr[\mathcal{S}] \\ &= \Pr[\{hh\}, \{ht\}, \{th\}, \{tt\}] \\ &= \Pr[\{hh\}] + \Pr[\{ht\}] + \Pr[\{th\}] + \Pr[\{tt\}] \\ &= 4\Pr[A], \end{aligned}$$

from which can conclude that

$$\Pr[A] = \frac{1}{4} = 0.25.$$

2.5 Conditional Probability

Conditional probabilities describe our knowledge about an event, given the knowledge that another event has happened. As an intuitive example, we could compare two situations. Suppose it is sunny outside and we want to know the probability that it starts raining. This probability is relatively low, whereas this probability would be a lot higher if it were cloudy. From this example, we may conclude that our knowledge of the weather at this moment, influences our prediction of rain in the near future.

2.5.1 A priori and a posteriori probability

The circumstances under which we would like to know the probability can be regarded as the observations of data. These observations provide us with insights into the circumstances and allow us to make a better estimate of the probability. The probability of an event A occurring without having made any observations is called the **a priori** probability (prior = before) and is denoted by $\Pr[A]$. The **a posteriori** probability (post = after) is the new probability after having obtained more information about the situation. This probability is denoted as $\Pr[A|B]$, which is read as "probability A given B". In the previous example, A could be regarded as the probability of rain in the near future and B as the current weather.

The conditional probability $\Pr[A|B]$ can be calculated as

$$\Pr[A|B] = \frac{\Pr[AB]}{\Pr[B]}, \tag{2.5}$$

where $\Pr[AB]$ is the probability of both events A and B occurring, which is equal to the probability of the intersection $\Pr[A \cap B]$. This equation scales the probability of observing an outcome

in the intersection of A and B by the probability of B. From the visual notation of Figure 2.1, this can be seen as the "area" of the event $A \cap B$ normalized by the "area" of B. Essentially, B becomes the new sample space.

2.5.2 Properties of conditional probability

From the definition of this conditional probability, three properties can be deduced:

1. It holds that $\Pr[A|B] \geq 0$;
2. It holds that $\Pr[B|B] = 1$;
3. For a set of disjoint events $A = \{A_1, A_2, \ldots, A_M\}$ it holds that $\Pr[A|B] = \Pr[A_1|B] + \Pr[A_2|B] + \ldots + \Pr[A_M|B]$.

The first and third properties are direct consequences of the probability axioms. The second property is trivial; it simply states that "the probability of having observed an event B after having observed an event B equals 1".

Example 2.4

Let us consider again the experiment of flipping a coin twice, but this time we would like to calculate the probability of flipping two heads if we obtain a head in the first toss. Let A_2 be the event that you observe two heads, and A_1 be the event that you observe a head in the first toss. Using Eq. (4), we can write

$$\Pr[A_2|A_1] = \frac{\Pr[A_1 A_2]}{\Pr[A_1]},$$

To calculate $\Pr[A_1 A_2]$, we can observe that the event space of the intersection is $A_1 \cap A_2 = \{hh\}$, that is the only outcome for which A_1 and A_2 both occur, while the event space of A_1 is $A_1 = \{hh, ht\}$. Thus, we can write

$$\Pr[A_1 A_2] = 0.25,$$

$$\Pr[A_1] = 0.5,$$

$$\Pr[A_2|A_1] = \frac{\Pr[A_1 A_2]}{\Pr[A_1]} = \frac{0.25}{0.5} = 0.5.$$

Not surprisingly, the posterior probability of observing two heads ($\Pr[A_2|A_1]$) is different than its prior probability ($\Pr[A_2]$) due to the observation of an event that influences the final outcome.

2.5.3 Law of total probability

The law of total probability relates marginal probabilities to conditional probabilities, which states that for an event space $\{B_1, B_2, \ldots, B_M\}$ with $\Pr[B_i] > 0$ for all i, it holds that

$$\Pr[A] = \sum_{i=1}^{M} \Pr[A|B_i] \Pr[B_i]. \tag{2.6}$$

This law inevitably follows from substituting the definition of the conditional probability as $\Pr[AB_i] = \Pr[A|B_i] \Pr[B_i]$ in the fifth consequence of the probability axioms.

2.5.4 Bayes' theorem

One of the most important rules in probability theory is Bayes' rule, which is obtained from the definition of conditional probability. The conditional probability can be rewritten as

$$\Pr[A|B]\Pr[B] = \Pr[AB] = \Pr[BA] = \Pr[B|A]\Pr[A]. \tag{2.7}$$

Equality of the middle two terms is obtained because these terms represent the same probability. Rewriting the leftmost and rightmost expressions gives Bayes' rule including the nomenclature of the separate terms as

$$\underbrace{\Pr[B|A]}_{\text{posterior}} = \frac{\overbrace{\Pr[A|B]}^{\text{likelihood}}\overbrace{\Pr[B]}^{\text{prior}}}{\underbrace{\Pr[A]}_{\text{evidence}}}. \tag{2.8}$$

Why is this particular formula so useful? The answer requires you to think in a certain context. Think of a context where the observation of an event A is related to a (nonobservable) underlying event B. An example of this context is where A resembles the observed data and B the model parameters creating this data. In the signal processing field, we would like to obtain the model parameters to draw conclusions about the underlying process (for example in medical diagnostics). We would like to estimate these parameters after observing some data. Therefore we are interested in the probability $\Pr[B|A]$. However, we cannot determine this immediately and therefore we need Bayes' rule. The initial (prior) probability of the model parameters is denoted by $\Pr[B]$ and is determined as an initial guess in terms of probability for the model parameters of the underlying process without having seen the data. The term $\Pr[A|B]$ represents the likelihood of the observed data under the assumed model parameters. Both of these terms can be calculated relatively easily. The last term $\Pr[A]$ represents the evidence, which is the probability of observing some data. This last term is usually more difficult to calculate and is therefore usually calculated by using the law of total probability.

Bayes' theorem originates from a thought experiment, in which Bayes imagines to be sitting with his back at a perfectly flat, square table and he asks his assistant to throw a ball onto the table. The ball can land anywhere on the table, but Bayes wanted to guess where the ball was without looking at it. Then, he would ask the assistant to throw another ball on the table, and asked him whether the second ball fell to the left, right, up or down compared to the first one; he would note this down. Then, he would repeat this a number of times, and by doing so he could keep updating his belief on where the first ball was. Although he could never be completely certain, with each piece of evidence, he would have a more accurate answer on the position of the first ball. Bayes' theorem was in fact never meant as a static formula to be used once and put aside, but rather as a tool to keep updating our estimate as our knowledge about the experiment grows.

Example 2.5

Suppose we have an event A, which indicates that a patient has a lung disease, and an underlying event B, which indicates that the patient smokes. Research has been conducted in a clinic and it has been found that among patients with a lung disease, 30% of the patients smoke. Furthermore, 20% of the people in the clinic smoke, and only 10% of the people in the clinic have a lung disease. Let's suppose that we are interested in the probability that a patient who smokes actually has a lung disease.

If we convert the given information into mathematical notation we can find that the

prior probability (a patient with a lung disease) is $\Pr[A] = 0.1$. Furthermore the evidence (a patient who smokes) is $\Pr[B] = 0.2$. Lastly, we find the likelihood (a patient with a lung disease smoking) as $\Pr[B|A] = 0.3$. From this, we can determine the posterior $\Pr[A|B]$ (the chance of a patient who smokes having a lung disease) as

$$\Pr[A|B] = \frac{\Pr[B|A]\Pr[A]}{\Pr[B]} = \frac{0.3 \cdot 0.1}{0.2} = 0.15. \tag{2.9}$$

Please note that the order of A and B is the opposite of in the above equation because $\Pr[B|A]$ is observed and therefore, the nomenclature changes.

To better understand how Bayes' rule can be used to update our belief as new evidence becomes available, we can also look at the following example.

Example 2.6

Suppose that for the general population, 1 in 5000 people carries the human immunodeficiency virus (HIV). A test for the presence of HIV yields either a positive (+) or negative (-) response. Suppose the test gives the correct answer 99% of the time. (a) What is $\Pr[H|+]$, the conditional probability that a person has the HIV virus, given that the person tests positive? (b) What is $\Pr[H|++]$, the conditional probability that the same person has the HIV virus, if he/she repeats the test and tests positive a second time?

Solution.

Let us first define all the involved probabilities as:

- $\Pr[H]$, the probability of having HIV;
- $\Pr[H^c]$, the probability of not having HIV;
- $\Pr[+]$, the probability of testing positive for HIV;
- $\Pr[+, H]$, the probability of testing positive for HIV and having HIV;
- $\Pr[+, H^c]$, the probability of testing positive for HIV and not having HIV;
- $\Pr[H|+]$, the probability of having HIV given having tested positive for HIV;
- $\Pr[+|H]$, the probability of testing positive for HIV given having HIV;
- $\Pr[+|H^c]$, the probability of testing positive for HIV given not having HIV;
- $\Pr[++]$, the probability of testing positive in a second (independent) test for HIV;
- $\Pr[H|++]$, the probability of having HIV given testing positive at the second test;
- $\Pr[++|H]$, the probability of testing positive at the second test for HIV given having HIV;
- $\Pr[++|H^c]$, the probability of testing positive at the second test for HIV given not having HIV.

(a) The probability that a person who has tested positive for HIV actually has the disease is

$$\Pr[H|+] = \frac{\Pr[+,H]}{\Pr[+]} = \frac{\Pr[+,H]}{\Pr[+,H] + \Pr[+,H^c]},$$

where H^c represents the complement of H and we have used the law of total probability to calculate $\Pr[+]$ in the denominator.

We can use Bayes' rule to evaluate these joint probabilities as

$$\begin{aligned}\Pr[H|+] &= \frac{\Pr[+|H]\Pr[H]}{\Pr[+|H]\Pr[H] + \Pr[+|H^c]\Pr[H^c]} \\ &= \frac{(0.99)(0.0002)}{(0.99)(0.0002) + (0.01)(0.9998)} \\ &= 0.0194.\end{aligned}$$

Even though the test is correct 99% of the time, the probability that a random person who tests positive actually has HIV is less than 2%. The reason this probability is so low is that the a priori probability that a person has HIV is very small.

(b) When the person performs the second test, we can use again Bayes' rule to calculate the probability that he/she has the disease, but this time we need to update the prior probability and the evidence according to what we calculated in the previous step. Since the two tests are independent, and the sensitivity of the test does not change, then $\Pr[++|H] = \Pr[+|H]$. However, now the posterior calculated in (a) becomes the new prior

$$\begin{aligned}\Pr[H|++] &= \frac{\Pr[++|H]\Pr[H|+]}{\Pr[++|H]\Pr[H|+] + \Pr[++|H^c]\Pr[H^c|+]} \\ &= \frac{(0.99)(0.0194)}{(0.99)(0.0194) + (0.01)(0.9806)} \\ &= 0.6620.\end{aligned}$$

Now the probability is more than 65%. This example shows how through Bayes' theorem, we were able to update our belief about the person having HIV as our knowledge about the test grew.

2.5.5 Independency

Another important definition in the field of probability theory is called independence. Two events A and B are independent if and only if the following holds

$$\Pr[AB] = \Pr[A]\Pr[B], \tag{2.10}$$

which is equivalent to $\Pr[A|B] = \Pr[A]$ and $\Pr[B|A] = \Pr[B]$. These equalities simply mean that the probability of an event A remains exactly the same after observing an event B, or vice versa. In other words, we do not get additional information through the occurrence of event B. Combining the previous two equations with the conditional probabilities gives Eq. (2.10).

Note that independent is not the same as disjoint! It is possible that events are both disjoint as independent, but this does not have to be the case. For example, if we chose randomly people

aged between 20 to 30 years, the event of choosing a male person and the event of choosing a person aged 22 are independent but not disjoint.

The definition of independence of two events can be extended to multiple events. Multiple events $\{A_1, A_2, \ldots, A_M\}$ are independent if and only if the following two constraints hold

- Every possible combination of two events is independent;
- It holds that $\Pr[A_1 A_2 \ldots A_M] = \Pr[A_1] \Pr[A_2] \ldots \Pr[A_M]$.

From this, we may automatically conclude that pairwise independence (constraint 1) does not immediately lead to the independence of multiple events, since the second constraint still needs to be satisfied.

Example 2.7

Let us come back once more to the example of tossing a coin twice and the event $A_2 = \{hh\}$. How can we use the notion of independence to calculate $\Pr[A_2]$?

We know that the probability of flipping one head in a single coin toss is $\Pr[A_1 = \{h\}] = 0.5$. Then, assuming that the two coin tosses are independent (no reason to think otherwise) we can use Eq. (2.10) to calculate

$$\Pr[A_2] = \Pr[A_1] \Pr[A_1] = 0.5 \cdot 0.5 = 0.25.$$

2.6 Random Variables

In the field of probability theory, processes that involve uncertainty and therefore have a random outcome are called random processes. These processes have outcomes that can be both numerical and categorical. An example of a numerical outcome can be the voltage measured over a noisy circuit, whereas an example of a categorical outcome can be the suit of the card drawn from a card deck. The latter can produce a card of any of the four suits: spades (♠), clubs (♣), diamonds (♢), and hearts (♡).

In order to perform calculations with these random processes, there is a need to introduce random variables. Random variables map all possible outcomes in the sample space $\mathcal{S}$ to numbers on the real line and are usually denoted by a capital letter. While an event can be both numerical as categorical, random variables are always numerical. In the case of a numerical sample space, the mapping through a random variable usually happens directly. A random variable X for the previous categorical example can be found by assigning four distinct numbers $\{1, 2, 3, 4\}$ to the four suits of cards s. This random variable $X(s)$ can be defined as

$$X(s) = \begin{cases} 1, & \text{for } s = \spadesuit, \\ 2, & \text{for } s = \diamondsuit, \\ 3, & \text{for } s = \clubsuit, \\ 4, & \text{for } s = \heartsuit. \end{cases} \tag{2.11}$$

However, different definitions are also allowed. A visualization of this mapping can be found in Figure 2.4, where all elements in the sample space are mapped to the real line.

In this section, we will first discuss the notation and properties of random variables, after which several statistical properties will be introduced to characterize random variables. We will then extend the concepts introduced for single random variables to pairs of random variables.

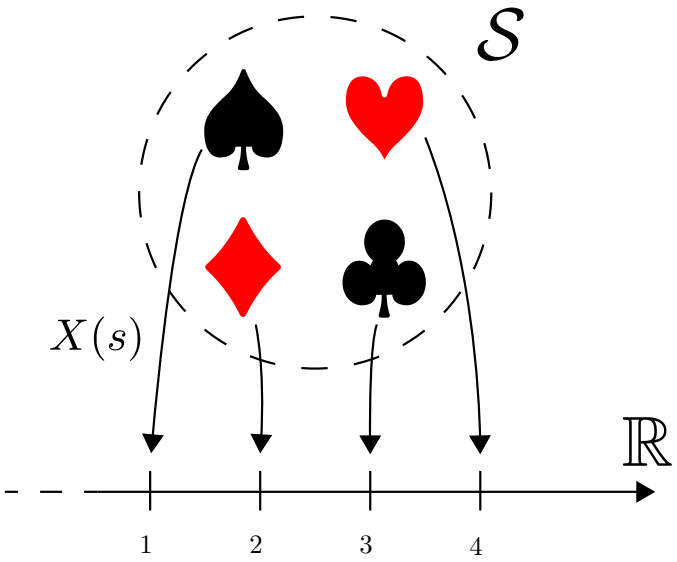


Figure 2.4: Visualization of the mapping of categorical elements (suits of cards) of a sample space $\mathcal{S}$ to numerical values.

2.6.1 Discrete and continuous random variables

Random variables can be either discrete, continuous, or mixed. The difference between them can be found in the allowed values that random variables can take on. A discrete random variable can take any value of a countable list of distinct values, whereas a continuous random variable can take any value on a continuous interval. A mixed random variable takes discrete values in parts of the real axis, and continuous values in other parts.

The previous example of the playing cards is an example of a discrete random variable because the random variable X can only take values in the set $\{1, 2, 3, 4\}$. Note that although the values that a discrete random variable can take on are usually integers, this is not always the case.

An example of a continuous random variable can be the measured voltage of a noisy circuit. Suppose that the noise has a lower and an upper limit of -5 and 5 mV, respectively, then any value within this interval can be measured. Thus, the measured voltage can be regarded as a continuous random variable.

As briefly mentioned before, random variables are denoted by a capital letter, such as $X(s)$. The argument s can be any element on the domain of X, which is the sample space; however, often the argument is dropped and the random variable is only indicated by the capital letter (e.g., X). The set of values or the interval to which the domain is mapped on the real line is called the range S_X of X. A possible value of $X(s)$ on the real line can be denoted by a lowercase x and is called a realization or an observation.

2.7 Probability distributions

Random variables are described by probability distributions. Two distinct categories of probability distributions exist for both discrete and continuous random variables.

y	$\Pr[Y = y]$
2	0.028
3	0.056
4	0.083
5	0.11
6	0.14
7	0.17
8	0.14
9	0.11
10	0.083
11	0.056
12	0.028

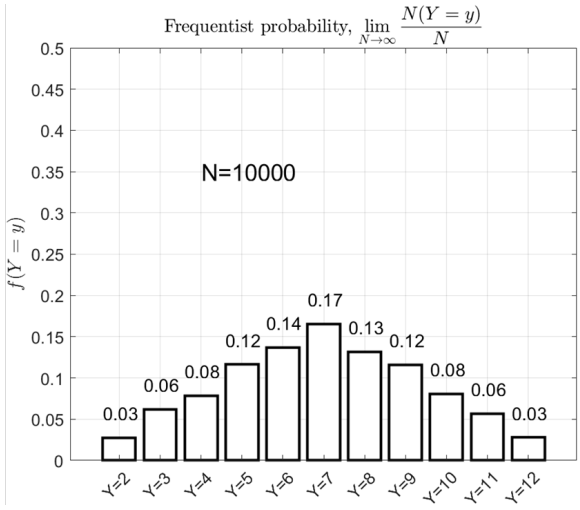


Figure 2.5: Example of PMF describing the random variable defined as the sum of the upward faces after rolling two dice. On the left, the PMF is represented as a table, while on the right a histogram is obtained by simulating the experiment 10000 times.

2.7.1 Discrete random variables

The probability that a discrete random variable X takes on the value x can be expressed by the **probability mass function (PMF)** as

$$p_X(x) = \Pr[X = x]. \tag{2.12}$$

A lower case p is used to indicate that it is the probability for a single realization. The subscript indicates the random variable that is considered and again x resembles the value of the realization. This function immediately returns a probability of x. Essentially, the PMF is a function that describes the probability associated with all possible outcomes of X. This can be given as a formula (see families of random variables in Appendix A and B), in a tabular form, or as a histogram. In Figure 2.5, examples of the latter two are shown. The random process is represented by rolling two dice, and the random variable Y is defined as the sum of the upward faces of the two dice. On the left, the PMF of Y, $P_Y(y)$ is represented in a tabular form, which associates each possible outcome with its probability. On the right, the PMF is represented as a histogram obtained by simulating the experiment for 10000 trials and calculating the relative frequency associated with each outcome.

Besides the PMF, the **cumulative distribution function (CDF)** of a discrete random variable can be defined (cumulative meaning summed/accumulated) as the probability of the discrete random variable X having a value lower than or equal to x. This function is denoted by a capital P as

$$P_X(x) = \Pr[X \leq x] = \sum_{x_i \in X \leq x} \Pr[X = x_i]. \tag{2.13}$$

This function equals the sum of the probabilities of all values in the range of X smaller than or equal to x. As shown on the right-hand side of (2.13), the CDF can be obtained from the PMF by summing all values of the PMF, for each $x_i \leq x$.

2.7.2 Continuous random variables

Similarly, for continuous random variables, the CDF can be defined as

$$P_X(x) = \Pr[X \leq x] = \int_{-\infty}^{x} p_X(x)dx. \tag{2.14}$$

The latter term differentiating the CDF of a continuous random variable from the CDF of a discrete random variable has purposely been rewritten because the range of X is now an interval and not a countable set. Therefore, integration is required instead of a summation. In this context, the term $p_X(x)$ is the probability density function (PDF) and denotes the probability density of an event x. This **probability density function** is defined as

$$p_X(x) = \frac{dP_X(x)}{dx}. \tag{2.15}$$

Please note the difference in meaning between the PMF and the PDF. While the PMF returns a true probability, the PDF returns a probability density. This probability density can be integrated over an interval in order to find the total probability. Suppose a continuous random variable X is uniformly distributed between 0 and 1 and we want to find the probability of X being a single value (e.g. 0.5) on this interval. This probability equals 0, although this might seem counter-intuitive. The interval from 0 to 1 contains an infinite number of distinct values. Although they have an infinitely small spacing between them, they are still distinct. If the probability would have been larger than 0, then all these values would have had this probability. Still, it is required by the probability axioms to have a total probability of 1. Therefore it is not possible with an infinite number of nonzero probabilities to satisfy the probability axiom unless the probability is infinitely close to 0.

2.7.3 Properties of probability distributions

From the definitions of the probability functions and the probability axioms, several consequences can be determined.

Cumulative distribution function

For the CDF, the following properties can be determined from the probability axioms

- $0 \leq P_X(x) \leq 1$ for all x;
- $P_X(-\infty) = 0$ and similarly $P_X(\infty) = 1$;
- $P_X(x_2) \geq P_X(x_1)$ holds when $x_2 \geq x_1$;
- It holds that $P_X(x_2) - P_X(x_1) = \Pr[x_1 < X \leq x_2]$ for $x_2 > x_1$.

In short, the CDF is bounded between 0 and 1, because the total probability cannot be negative nor larger than 1 by the probability axioms. The range of X that is upper-bounded by $-\infty$ is a null set and therefore its probability is equal to 0. Additionally, the range of X that is upper-bounded by ∞ is the entire range and therefore the probability equals 1. The CDF is defined as the cumulative probability and is therefore never decreasing for consecutive numbers. Lastly, the CDF is defined over the interval from $-\infty$ to an upper-bound. Subtracting two CDFs will result in the cumulative probability over an interval that is now also lower-bounded.

Probability mass function

For the probability function, the following properties can be determined from the probability axioms

- If x is not in the range of X then $p_X(x) = 0$;
- $0 \leq p_X(x) \leq 1$ for all x;
- $\sum_x p_X(x) = 1$.

The first property states that the probability of observing an outcome that cannot be observed equals 0. The second and third properties are direct consequences of the probability axioms. In particular, the third property is equivalent to calculating the probability of all possible outcomes and thus of the sample space.

Probability density function

Similarly to the PMF, properties of the PDF are

- It holds that $p_X(x) \geq 0$ for all x;
- By definition $P_X(x) = \int_{-\infty}^{x} p_X(x)dx$;
- It holds that $\int_{-\infty}^{\infty} p_X(x)dx = 1$.

The first and last properties are direct consequences of the probability axioms. Please note that now there is no upper-bound to the value of $p_X(x)$, since this function returns the probability density and not the true probability. The second property is just a repetition of the definition of the CDF.

2.7.4 Mixed random variables

When the PDF of a random variable consists of a combination of both probability mass and density functions, it can be regarded as a mixed random variable. An example is given hereafter.

Example 2.8

Suppose we are monitoring the duration of phone calls. The measured duration x in minutes can be regarded as a continuous random variable X, which is distributed as an exponential random variable with rate parameter $\lambda = 1$ (see Appendix B). This can be written as $X \sim \text{Exp}(\lambda = 1)$ with PDF

$$p_X(x) = \begin{cases} e^{-x}, & \text{for } x \geq 0, \\ 0, & \text{for } x < 0. \end{cases} \tag{2.16}$$

The service provider decides for simplicity to round all phone calls lasting less than 60 seconds to exactly 1 minute. This administrative duration y in minutes can be represented by a mixed random variable Y. The administrative duration can be mathematically determined as

$$y = \begin{cases} 1, & \text{for } x < 1, \\ x, & \text{for } x \geq 1. \end{cases} \tag{2.17}$$

Let us now derive the PDF of Y. In order to do so, the derivation is split for the cases $Y < 1$, $Y = 1$ and $Y > 1$.

We can easily see that $p_Y(y) = 0$ for $y < 1$, because the duration is always rounded to at least 1 minute. For $Y = 1$ a true probability is involved, since there is a nonzero probability that the administrative duration is exactly 1 minute. Because the integral of the PDF should return a nonzero value at $y = 1$ when integrating $\int_{1^-}^{1^+} p_Y(y)dy$, a Dirac delta pulse should be located at $y = 1$ in order to make sure that the integral returns the true probability. This delta pulse should be multiplied with the probability of the occurrence of this event. This probability can be determined as

$$\Pr[Y = 1] = \Pr[X < 1] = \int_{-\infty}^{1} p_X(x)dx = \int_{0}^{1} e^{-x}dx = \left[-e^{-x}\right]_0^1 = 1 - e^{-1}. \tag{2.18}$$

For $Y > 1$ the PDF of Y simply follows the exponential distribution of X. In order to cope with this partial domain, $p_X(x)$ needs to be multiplied with the shifted unit step function.

So to conclude the PDF of Y can be written as

$$p_Y(y) = (1 - e^{-1}) \cdot \delta(y - 1) + e^{-y} \cdot u(y - 1). \tag{2.19}$$

The delta pulse and the unit step function are defined in Appendix C.

2.7.5 Statistical characterization of a random variable

Probability distribution fully characterizes a random variable. However, they are not always easy or possible to obtain. Moreover, even when they are available, it is difficult to fully appreciate the rich information they convey. For these reasons, random variables can be characterized by quantities that summarize some of their characteristics. These quantities are called **moments**. When the probability distribution of a random variable X is available, moments can be calculated by the expectation operator $\mathrm{E}[\cdot]$, and they can be considered as parameters of the probability model of X. However, in most cases, we do not know the probability distribution of X, but we might have a set of experimental data (outcomes of repeated experiments) available. In this case, we can calculate **sample moments** by calculating averages over our data samples. These sample moments can be considered descriptive statistics of the random variable X, and they are only equivalent to the true moments of X asymptotically, meaning when the number of data samples is very large.

Moments of a random variable

Moments of any order m of a random variable can be calculated by the application of the expectation operator $\mathrm{E}[\cdot]$. The m^{th} moment of a discrete random variable is defined as

$$\mathrm{E}[X^m] = \sum_{x \in S_X} x^m p_X(x), \tag{2.20}$$

and for a continuous random variable defined as

$$\mathrm{E}[X^m] = \int_{-\infty}^{\infty} x^m p_X(x)dx. \tag{2.21}$$

The expectation operator has the following properties:

- It is a *linear* operator. For random variables X and Y, and deterministic constants a and b, it holds
$$\mathrm{E}[aX + bY] = a\,\mathrm{E}[X] + b\,\mathrm{E}[Y]; \tag{2.22}$$
- It is a *positive* operator. If $X \geq Y$, then $\mathrm{E}[X] \geq \mathrm{E}[Y]$;
- For a deterministic variable z, $\mathrm{E}[z] = z$.

The first order moment of a random variable X is known as *expected value*, $\mathrm{E}[X]$, or *mean*, μ_X, and can be regarded as the average value of the distribution for an infinite number of observations, or as the center of gravity of a distribution.

For a discrete random variable X, this is calculated as
$$\mathrm{E}[X] = \mu_X = \sum_{x \in S_X} x \cdot p_X(x), \tag{2.23}$$
and for a continuous random variable X as
$$\mathrm{E}[X] = \mu_X = \int_{-\infty}^{\infty} x \cdot p_X(x) dx. \tag{2.24}$$

We can also define **central moments**, which can be found by subtracting from the random variable their first order moment, i.e. by subtracting the mean, as
$$\mathrm{E}[(X - \mu_X)^m] = \sum_{x \in S_X} (x - \mu_X)^m p_X(x), \tag{2.25}$$
for discrete random variables and
$$\mathrm{E}[(X - \mu_X)^m] = \int_{-\infty}^{\infty} (x - \mu_X)^m p_X(x) dx, \tag{2.26}$$
for continuous random variables.

The second central moment is known as the *variance* ($\mathrm{Var}[X]$ or σ_X^2) of a random variable and it is a measure of the spread of the distribution. It can be determined for a discrete random variable X as
$$\mathrm{Var}[X] = \sigma_X^2 = \mathrm{E}[(X - \mu_X)^2] = \sum_{x \in S_X} (x - \mu_X)^2 \cdot p_X(x) \tag{2.27}$$
and for a continuous random variable X as
$$\mathrm{Var}[X] = \sigma_X^2 = \mathrm{E}[(X - \mu_X)^2] = \int_{-\infty}^{\infty} (x - \mu_X)^2 \cdot p_X(x) dx. \tag{2.28}$$

In some cases, it is more convenient to rewrite the variance through
$$\begin{aligned} \mathrm{Var}[X] &= \mathrm{E}[(X - \mu_X)^2] \\ &= \mathrm{E}[X^2 - 2\mu_X X + \mu_X^2] \\ &= \mathrm{E}[X^2] - 2\mu_X\,\mathrm{E}[X] + \mu_X^2 \\ &= \mathrm{E}[X^2] - 2\mu_X^2 + \mu_X^2 \\ &= \mathrm{E}[X^2] - \mu_X^2. \end{aligned} \tag{2.29}$$

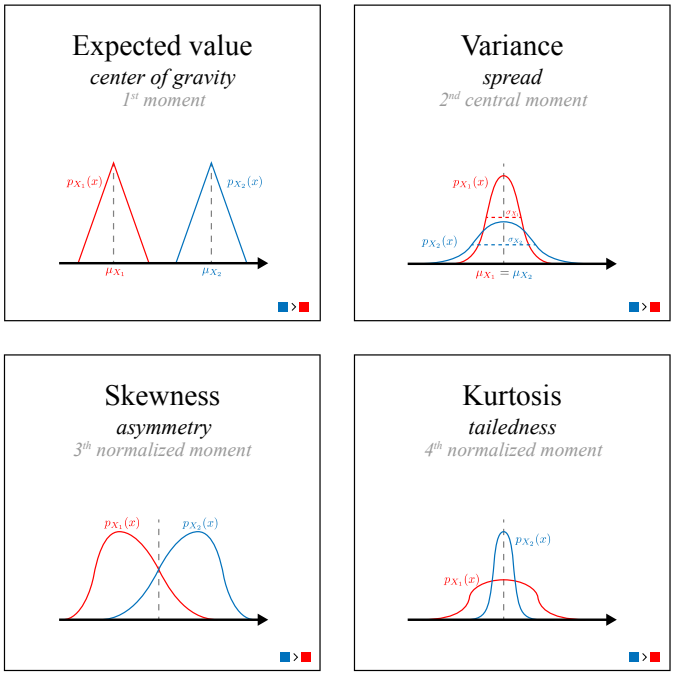


Figure 2.6: Visualization of most common moments. The distribution in blue has a larger m^{th} moment than the distribution in red.

The *standard deviation* is also a commonly used characteristic of a random variable and simply equals the square root of the variance as

$$\text{Std}[X] = \sigma_X = \sqrt{\text{Var}[X]}. \tag{2.30}$$

Finally, the m^{th} **order normalized central moments** can be determined as the central moment divided by the m^{th} order of the standard deviation as

$$\frac{\text{E}[(X - \mu_X)^m]}{\sigma^m} = \frac{1}{\sigma^m} \sum_{x \in S_X} (x - \mu_X)^m p_X(x) \tag{2.31}$$

f X is discrete, and

$$\frac{\text{E}[(X - \mu_X)^m]}{\sigma^m} = \frac{1}{\sigma^m} \int_{-\infty}^{\infty} (x - \mu_X)^m p_X(x) dx, \tag{2.32}$$

if X is continuous.

Well-known normalized central moments are the *skewness* (3^{rd} order), which describes the degree of symmetry of the distribution, and the kurtosis (4^{th} order), which is a measure of how the tails of the distribution are shaped, taking the Gaussian distribution as a reference (zero kurtosis). Moments of probability distributions are represented schematically in Figure 2.6.

Sample moments

When the underlying probabilistic model is not known, we can obtain information on the probability distribution of a random variable by calculating sample moments. Suppose a set of n

observations is defined as

$$X = \{x_0, x_1, \ldots, x_{N-1}\}. \tag{2.33}$$

The *average* or *sample mean* of this set can be determined by summing over all values and dividing by the number of observations through

$$\text{average}(X) = \frac{\sum_{i=1}^{n} x_i}{n}. \tag{2.34}$$

The *median* is another statistic to characterize a set of observations. The median is defined as the value in the middle of the ordered data set. After ordering the data set, at the left and right of the median there should be an equal number of observations. In case the data set has an even number of observations the average of the two observations in the middle is usually taken.

The *mode* of a data set is the observation with the highest number of occurrences. If multiple observations occur both as often, then both observations are the mode of data set and the data set is then called multimodal.

Example 2.9

Suppose the students of the statistical signal processing course have made the final exam and the final course grades are available. A subset of all grades could consist of the following grades

$$\{8, 5, 6, 2, 7, 9, 10, 2, 7, 5, 6, 7\}. \tag{2.35}$$

From this set, the average value can be computed as 6.17. The median can be determined after ordering the set as

$$\{2, 2, 5, 5, 6, \mathbf{6}, \mathbf{7}, 7, 7, 8, 9, 10\}. \tag{2.36}$$

The middle values are 6 and 7 because there are an even number of observations. Therefore the median can be determined as 6.5. From this set the mode can be determined as 7, since this observation occurs 3 times, whereas all other grades occur only once or twice.

More in general, given a set of data samples, sample moments can be calculated as

- *Sample mean*:
$$\mu_X = \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n};$$
- *Sample variance*:
$$\sigma_X^2 = \frac{\sum_{i=1}^{n} (x_i - \mu_X)^2}{n};$$
- *Sample skweness*:
$$\text{skew} = \frac{1}{n} \frac{\sum_{i=1}^{n} (x_i - \mu_X)^3}{\sigma_X^3};$$
- *Sample kurtosis*:
$$\text{kurt} = \frac{1}{n} \frac{\sum_{i=1}^{n} (x_i - \mu_X)^4}{\sigma_X^4}.$$

As mentioned before sample moments are only equivalent to the true moments of X asymptotically, meaning when the number of data samples is very large. For a small data sample,

the sample moments are corrected by using a different normalization. For example, the sample variance is corrected as

$$\sigma_X^2 = \frac{\sum_{i=1}^{n}(x_i - \mu_X)^2}{n-1},$$

where the normalization for $n-1$ instead of n is due to the fact that we already used the data samples to calculate the sample mean, and thus we have one degree of freedom less.

2.8 Families of random variables

As mentioned before, probability distributions can be described as a formula, in a tabular form, or as a histogram. Families of random variables are sets of random variables that can be described by the same probability distribution, and they are typically fully defined by a mathematical formula governed by a set of parameters.

To indicate that a random variable X is distributed according to a certain distribution, e.g., univariate standard normal distribution, we may write $X \sim \mathcal{N}(0,1)$. By this notation, the letter $\mathcal{N}$ indicates the normal distribution, while the numbers in parenthesis indicate the parameters controlling the distribution. In the case of a normal distribution, these are the mean and the variance. Thus, $X \sim \mathcal{N}(0,1)$ reads "the random variable X is normally distributed with zero mean and unitary variance".

A family of random variables can describe experiments that are very different but behave statistically in the same way. Moreover, by changing the controlling parameters, we can describe the same experiment under different conditions.

2.8.1 Families of discrete random variables

An example of a family of discrete random variables is the *Binomial random variable* The Binomial distribution is a discrete probability distribution that models an experiment with a probability of success p. The Binomial distribution gives the probability of observing x successes in n independent trials of the experiment. The distribution is fully characterized by the parameters n and p, and it is denoted as $X \sim \text{Binomial}(n,p)$. The parameter n denotes the number of independent trials and the parameter p denotes the probability of observing a success in n trials.

The PMF of the Binomial(n,p) distribution is given as

$$p_X(x) = \binom{n}{x} p^x (1-p)^{n-x}, \tag{2.37}$$

where $0 < p < 1$ and n is an integer such that $n \geq 1$.

The CDF of the discrete Binomial(n,p) distribution can be determined as

$$P_X(x) = \sum_{m=0}^{x} \binom{n}{m} p^m (1-p)^{n-m}. \tag{2.38}$$

The *expected value* of the discrete Binomial(n,p) distribution can be determined as

$$\mathrm{E}[X] = np. \tag{2.39}$$

The *variance* of the discrete Binomial(n,p) distribution can be determined as

$$\mathrm{Var}[X] = np(1-p). \tag{2.40}$$

More details on the Binomial random variables together with an overview of the most common families of discrete random variables is provided in Appendix A.

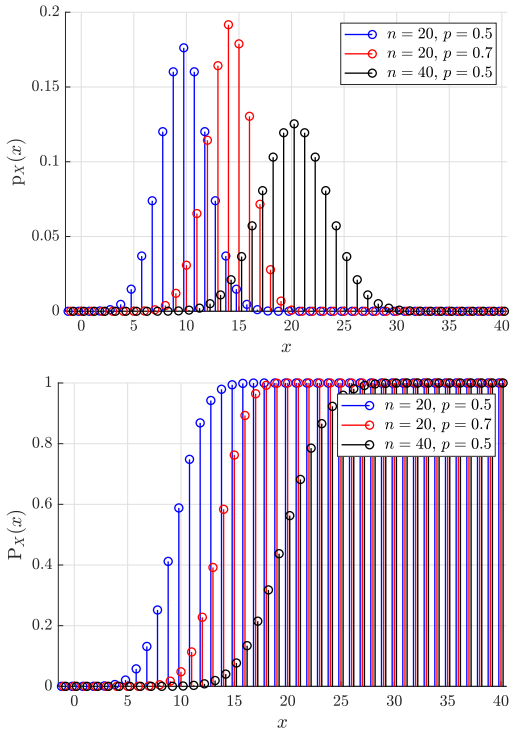


Figure 2.7: Example plot of the (a) PMF and (b) CDF of the Binomial(p) distribution.

2.8.2 Families of continuous random variables

An example of a family of continuous random variables is the Normal or Gaussian $\mathcal{N}(\mu, \sigma^2)$ distribution. The Normal distribution is bell-shaped and symmetric, and it is fully characterized by its mean, μ, and its variance, σ^2.

The PDF of the continuous Gaussian $\mathcal{N}(\mu, \sigma^2)$ distribution is given as

$$p_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \tag{2.41}$$

where $\sigma > 0$.

The CDF of the continuous Gaussian $\mathcal{N}(\mu, \sigma^2)$ distribution can be determined as

$$P_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{x} e^{-\frac{(u-\mu)^2}{2\sigma^2}} du. \tag{2.42}$$

The *expected value* of the continuous Gaussian $\mathcal{N}(\mu, \sigma^2)$ distribution can be determined as

$$\mathrm{E}[X] = \mu. \tag{2.43}$$

The *variance* of the continuous Gaussian $\mathcal{N}(\mu, \sigma^2)$ distribution can be determined as

$$\mathrm{Var}[X] = \sigma^2. \tag{2.44}$$

A specific case of the Normal distribution is the Standard normal distribution, which is simply a Normal distribution with mean $\mu = 0$ and variance $\sigma^2 = 1$. This function can be regarded as

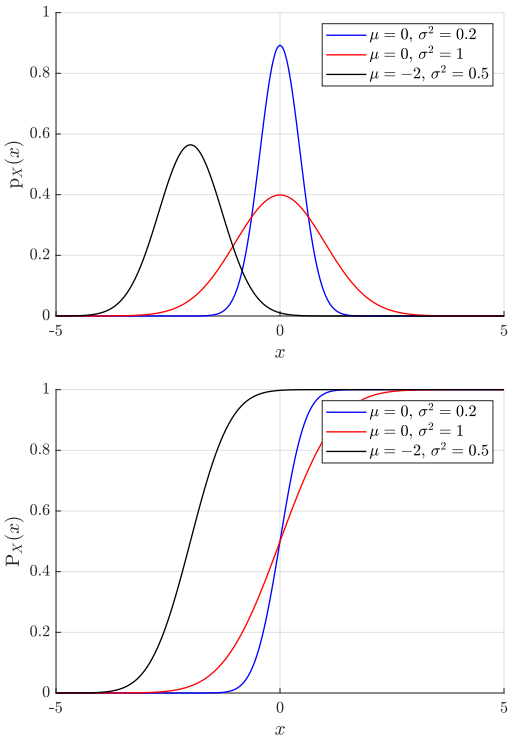


Figure 2.8: Example plot of the (a) PDF and (b) CDF of the Gaussian (μ, σ^2) distribution.

the normalized Gaussian distribution. Any Gaussian random variable $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$ can be transformed to a random variable X under the Standard normal distribution by subtracting its mean and dividing by the standard deviation as $X = \frac{Y-\mu_Y}{\sigma_Y}$.

To easily calculate probabilities from the standard Normal distribution, the **Q-function** is used, which calculates the probability of a Standard normal distributed random variable X exceeding a certain threshold x. It is also known as the right-tail probability of the standard Gaussian distribution since it is calculated by integrating the right side of the Gaussian PDF from the threshold x up to ∞. The Q-function is defined as

$$Q(x) = \Pr[X > x] = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-\frac{u^2}{2}} du. \tag{2.45}$$

The function can be used for all Gaussian distributed random variables; however, the random variable and the corresponding threshold should be normalized first. Additionally, through symmetry follows that $Q(x) = 1 - Q(-x)$, where $Q(-x)$ is equal to the cumulative density function $\Phi_X(x)$ of a standard Normal distribution.

As is implicated by the central limit theorem (which will be explained in Section 2.10), the Gaussian distribution is extremely important. The Gaussian distribution is often used to model measurements in practice and, thanks to the central limit theorem, its use can often be extended to other distributions. As an example, a Gaussian distribution is also often used to model the thermal noise of a band-limited system.

More details on the Gaussian random variables together with an overview of the most common families of continuous random variables is provided in Appendix B.

2.9 Sampling random variables

Most statistical packages in computing software provide a so-called pseudorandom number generator, which is an algorithm to randomly sample a number between 0 and 1 with equal probability. Basically, this means generating random samples from a continuous random variable U which follows a uniform distribution, $U \sim \mathcal{U}(0,1)$ (see Appendix B). More in general, sampling a random variable means generating values $x \in X$ in such a way that the probability of generating x is in accordance with the probability density/mass function $p_X(x)$, or equivalently the CDF $P_X(x)$, associated with X.

Assuming that we have a pseudorandom number generator, how can we generate samples of any random variable X if we know its probability distribution? We need to find a transformation $T : [0,1] \to \mathbb{R}$ such that $T(U) = X$. For continuous random variables, the following theorem can help us with this task.

Theorem. Let X be a continuous random variable with CDF $P_X(x)$ which possesses an inverse P_X^{-1}. Let $U \sim \mathcal{U}(0,1)$ and $Y = P_X^{-1}(U)$, then $P_X(x)$ is the CDF for Y. In other words, Y has the same distribution as X.

Proof. By definition of CDF, and since $Y = P_X^{-1}(U)$, we can write

$$P_Y(x) = [Y \leq x] = \Pr[P_X^{-1}(U) \leq x].$$

Since the CDF is an increasing function, then we can apply this function at both sides of the operator "$\leq$".

$$\Pr[P_X^{-1}(U) \leq x] = \Pr[P_X(P_X^{-1}(U)) \leq P_X(x)]$$

Noting that $P_X(P_X^{-1}(U)) = U$, and that $\Pr[U \leq x] = x$ when $U \sim U(0,1)$, then

$$\Pr[P_X(P_X^{-1}(U)) \leq P_X] = \Pr[U \leq P_X(x)] = P_X(x).$$

□

According to this theorem, the transformation T we were looking for is simply given by P_X^{-1}. Then, to sample x, it is sufficient to follow these steps:

- Generate a random number u from uniform distribution $U \sim \mathcal{U}(0,1)$;
- Find the inverse of the CDF of X, P_X^{-1};
- Compute x as $x = P_X^{-1}(u)$.

This method of sampling a random variable is known as the *inverse transform technique.*

For discrete random variables, however, this technique cannot be applied directly, because when X is discrete, the relationship between X and $P_X^{-1}(u)$ is not univocal. Then $P_X^{-1}(u)$ is defined as the least element $x \in X$ for which $U < P_X(x)$. More formally, let $X = \{x_1, ..., x_n\}$ be a discrete random variable with PMF $p_X(x)$, and where $x_1 \leq ... \leq x_n$. Let us define each value of the CDF of X as

$$q_i = \Pr[X \leq x_i] = \sum_{j=1}^{i} p_X(x_j). \tag{2.46}$$

The sampling formula for X becomes:

$$P_X(x) = \begin{cases} x_1, & \text{if } U < q_1, \\ x_2, & \text{if } q_1 \leq U < q_2, \\ \vdots & \vdots \\ x_{n-1}, & \text{if } q_{n-2} \leq U < q_{n-1}, \\ x_n, & \text{otherwise.} \end{cases}$$

2.10 Functions and pairs of random variables

2.10.1 Functions of random variables

When a random variable generated under a certain probability distribution is manipulated, the probability distribution as well as its moment will change. Although the new probability distribution that follows from this manipulation is usually difficult to determine, the moments of the transformed random variable can be easily calculated. The expected value of a function $g(X)$ of a random variable X can be determined as

$$\mathrm{E}[g(X)] = \begin{cases} \sum_{x \in S_X} g(x) \cdot p_X(x), & \text{if } X \text{ is discrete,} \\ \int_{-\infty}^{\infty} g(x) \cdot p_X(x) dx. & \text{if } X \text{ is continuous.} \end{cases} \tag{2.47}$$

Example 2.10

One common category of transformation consists in the linear combination of a continuous random variable X, with mean μ_X and variance σ_X^2. Suppose that the random variable Y is a linear combination of X defined by

$$Y = aX + b, \tag{2.48}$$

with a and b deterministic constants. The mean of the random variable Y can be determined as

$$\begin{aligned} \mathrm{E}[Y] &= \mathrm{E}[aX + b] \\ &= \int_{-\infty}^{\infty} (ax + b) \cdot p_X(x) dx \\ &= a \int_{-\infty}^{\infty} x \cdot p_X(x) dx + b \int_{-\infty}^{\infty} p_X(x) dx \\ &= a \cdot \mathrm{E}[X] + b = a \cdot \mu_X + b. \end{aligned} \tag{2.49}$$

Similarly, the variance of Y can be determined using the previous result as

$$\begin{aligned}\mathrm{Var}[Y] &= \mathrm{E}[(Y-\mu_Y)^2] \\ &= \mathrm{E}[(aX+b-(a\cdot\mu_X+b))^2] \\ &= \mathrm{E}[a^2(X-\mu_X)^2] \\ &= \int_{-\infty}^{\infty} a^2(x-\mu_X)^2\cdot p_X(x)dx \\ &= a^2\int_{-\infty}^{\infty}(x-\mu_X)^2\cdot p_X(x)dx \\ &= a^2\,\mathrm{Var}[X].\end{aligned} \tag{2.50}$$

2.10.2 Pairs of random variables

So far we have focused on random processes described by a single random variable. However, some random processes can be defined by multiple random variables. As an example, suppose we have an experiment in which we throw two dice consecutively and we collect the total number of eyes after the first and second dice are thrown. Assuming fair dice, the total number of eyes observed after the first throw should be an integer between 1 and 6. But how can we now describe the total number of eyes after throwing the second die? It will be clear that this second observation will be between 2 and 12 eyes. However, we do need to keep in mind that the second observation depends greatly on the first observation. It is for example impossible to first observe a total of 5 eyes after throwing the first dice and then to observe a total of 4 eyes after throwing the second dice. A full description of this random process requires knowledge of both the first throw and the second throw.

Through this example, it is clear that there is a need to combine multiple random variables into a single function to draw conclusions about observations and their probabilities. Multiple random variables that are all associated with an event are called *joint random variables*. These random variables are usually denoted with different capital letters (e.g. X and Y). In this section, we will first discuss the case of pairs of random variables, while in Section 3.2, we will extend these concepts to any number of joint random variables by introducing random vectors.

Joint cumulative distribution function

To describe the probabilities that are associated with the observations (x, y), we will introduce the joint CDF, which is an extension of the CDF for a single random variable. The joint CDF of random variables X and Y is defined as

$$P_{X,Y}(x,y) = \Pr[X \le x, Y \le y], \tag{2.51}$$

where the comma in the probability operator indicates that both conditions need to hold.

The joint CDF has the following properties:

- It holds that $0 \le P_{X,Y}(x,y) \le 1$;
- It holds that $P_{X,Y}(\infty,\infty) = 1$ and that $P_{X,Y}(x,-\infty) = P_{X,Y}(-\infty,y) = 0$;
- It holds that $P_X(x) = P_{X,Y}(x,\infty)$ and similarly $P_Y(y) = P_{X,Y}(\infty,y)$;
- The function is nondecreasing (i.e. for $x \le x_0$ and $y \le y_0$ it holds $P_{X,Y}(x,y) \le P_{X,Y}(x_0,y_0)$).

The first property is a consequence of the total probability axiom, by which a probability cannot be negative nor larger than 1. A total probability of 1 only occurs when both $X \leq x$ and $Y \leq y$ are always satisfied, which is the case for $P_{X,Y}(\infty, \infty)$. Similarly, there is by definition no event where $x \leq -\infty$ or $y \leq -\infty$ and therefore this (cumulative) probability is always 0.

When one of the conditions $X \leq x$ and $Y \leq y$ is always satisfied, the cumulative distribution does not depend on that variable anymore and therefore that variable can be removed from the equation, leaving a distribution that now depends on one single random variable. This process is called **marginalization** and will be discussed later on.

Finally, as seen previously the cumulative distribution is a sum or integral over an ever-increasing domain with nonnegative values, leading to a nondecreasing function.

Joint probability density function

The previous definition of the joint CDF holds both for continuous as well as discrete random variables. For continuous random variables, the joint PDF can be defined through the relationship

$$\int_{-\infty}^{x} \int_{-\infty}^{y} p_{X,Y}(u, v)\ du\ dv = P_{X,Y}(x, y) \tag{2.52}$$

and can be calculated from the joint CDF as

$$p_{X,Y}(x, y) = \frac{\partial^2 P_{X,Y}(x, y)}{\partial x \partial y}. \tag{2.53}$$

Joint probability mass function

In the case of discrete random variables, the joint PMF can be defined explicitly as

$$p_{X,Y}(x, y) = \Pr[X = x, Y = y], \tag{2.54}$$

because the values that the function returns are true probabilities.

The *expected value* of the joint random variables can be determined similarly to the single variable case. Given a new random variable W, which is an arbitrary function of the random variable X and Y through $W = g(X, Y)$, the expected value of W can be determined as

$$\mathrm{E}[W] = \sum_{x \in S_X} \sum_{y \in S_Y} g(x, y) \cdot p_{X,Y}(x, y) \tag{2.55}$$

in the case of discrete random variables and as

$$\mathrm{E}[W] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) \cdot p_{X,Y}(x, y) dx\ dy \tag{2.56}$$

in the case of continuous random variables.

2.10.3 Marginalization

In many engineering problems, you are only interested in the outcome of one random variable. When a joint probability function is available it is possible to marginalize over all other random variables except for the one random variable in which you are interested, leaving you with the probability function of just that random variable. This marginalization is performed for discrete random variables through

$$p_X(x) = \sum_{y \in S_Y} p_{X,Y}(x, y), \tag{2.57}$$

and for continuous random variables through

$$p_X(x) = \int_{-\infty}^{\infty} p_{X,Y}(x, y)dy. \tag{2.58}$$

This operation can be understood when viewing the summation or integration as "taking all possibilities of all other random variables into account". The proof for the continuous random variables follows.

Proof.

$$\begin{aligned} p_X(x) &= \frac{\partial P_X(x)}{\partial x} \\ &= \frac{\partial}{\partial x} \Pr[X \leq x] \\ &= \frac{\partial}{\partial x} \Pr[X \leq x, Y \leq \infty] \\ &= \frac{\partial}{\partial x} \int_{-\infty}^{x} \left(\int_{-\infty}^{\infty} p_{X,Y}(v, u)du \right) dv \\ &= \int_{-\infty}^{\infty} p_{X,Y}(u, v)dv. \end{aligned} \tag{2.59}$$

□

A similar proof can be found for the marginalization of discrete random variables.

2.10.4 Conditional probabilities

When considering events with multiple outcomes we might be able to deduce one outcome from another. Take the example of rolling two dice. If we know that the first die rolls 5 eyes, then we automatically know that the total number of eyes after the second throw can go from 6 to 11 with equal probability. Similarly, we may deduce with 100% certainty that the first throw should have given 6 eyes when 12 eyes are observed after the second throw.

This line of reasoning is an example of conditional probability. Conditional probability involves the probability of some event when another event is observed. In Section 2.4, the definition of conditional probability was given and explained. If this definition is combined with random variables, two separate cases can be identified.

First, it is possible to determine the conditional probability function, which is the PMF or PDF of observing outcomes (x, y), when it is known that the outcomes can be found somewhere in an event set B, which is a subset of the sample space $\mathcal{S}$. When the joint probability function is given by $p_{X,Y}(x, y)$, the conditional probability function of X and Y given B (i.e. $X, Y|B$) can be determined as

$$p_{X,Y|B}(x, y) = \begin{cases} \frac{p_{X,Y}(x,y)}{\Pr[B]}, & \text{when } (x, y) \in B, \\ 0, & \text{otherwise.} \end{cases} \tag{2.60}$$

This operation can be regarded as assigning zero probability to all outcomes outside set B and normalizing the remaining function to satisfy the total probability axiom.

Secondly, it is also possible that one random variable is observed and we would like to find the probability function of the other random variable. Let us now suppose that the random variable

Y is observed as y and we are interested in finding the conditional probability function of X, which is $p_{X|Y}(x|y)$. For discrete random variables, this conditional PMF can be determined as

$$p_{X|Y}(x|y) = \Pr[X = x|Y = y] = \frac{\Pr[X = x, Y = y]}{\Pr[Y = y]} = \frac{p_{X,Y}(x, y)}{p_Y(y)}. \tag{2.61}$$

Keep in mind that the marginalized PMF $p_Y(y)$ can be calculated using marginalization.

Similarly, to the derivation for the discrete random variables, the conditional PDF of continuous random variables can be determined as

$$p_{X|Y}(x|y) = \frac{p_{X,Y}(x, y)}{p_Y(y)} = \frac{p_{X,Y}(x, y)}{\int_{-\infty}^{\infty} p_{X,Y}(x, y)dx} \tag{2.62}$$

Example 2.11

Suppose we are given the joint PDF

$$p_{X,Y}(x, y) = \begin{cases} Ce^{-x}, & \text{for } x \geq 0 \text{ and } -x \geq y \geq x, \\ 0, & \text{otherwise,} \end{cases} \tag{2.63}$$

where C is a constant to be determined. Let us now derive the value of C, the marginal PDFs $p_X(x)$ and $p_Y(y)$ and the conditional PDF $p_{Y|X}(y|x)$.

In order to calculate C we need to make use of the total probability axiom, which is given as

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p_{X,Y}(x, y)\ dx\ dy = 1 \tag{2.64}$$

for the joint random variables case. If we substitute the given joint PDF with the given constraints we find

$$\begin{aligned} \int_0^{\infty} \int_{-x}^{x} Ce^{-x}\ dy\ dx &= C \int_0^{\infty} e^{-x} \int_{-x}^{x} 1\ dy\ dx = C \int_0^{\infty} e^{-x} \left[y\right]_{-x}^{x}\ dx \\ &= 2C \int_0^{\infty} xe^{-x}\ dx = 2C \left[-xe^{-x}\right]_0^{\infty} - 2C \int_0^{\infty} -e^{-x}dx \\ &= -2C \left[xe^{-x}\right]_0^{\infty} - 2C \left[e^{-x}\right]_0^{\infty}, \\ &= -2C(0 - 0) - 2C(0 - 1) = 2C = 1. \end{aligned} \tag{2.65}$$

From this follows that $C = \frac{1}{2}$.

In order to calculate $p_X(x)$ we will marginalize $p_{X,Y}(x, y)$ over Y as

$$\int_{-\infty}^{\infty} p_{X,Y}(x, y)dy = \int_{-x}^{x} Ce^{-x}dy = Ce^{-x} \int_{-x}^{x} 1dy = 2Cxe^{-x}. \tag{2.66}$$

When substituting the value of C and considering that the function only exists for $x \geq 0$, the marginal PDF can be determined as

$$p_X(x) = \begin{cases} xe^{-x}, & \text{for } x \geq 0, \\ 0, & \text{otherwise.} \end{cases} \tag{2.67}$$

Similarly, we can find the marginal PDF $p_Y(y)$. One might say that the marginalization integral has a lower-bound for x, namely $x \geq 0$. However, this is not entirely correct.

When rewriting the constraint $-x \geq y \geq x$ as $x \geq |y|$ it becomes clear that x is not bounded by 0, but by $|y|$. The marginal PDF is not bounded on y by a constant and can therefore be written as

$$p_Y(y) = \int_{|y|}^{\infty} Ce^{-x}dx = C[-e^{-x}]_{|y|}^{\infty} = C(-0 + e^{-|y|}) = \frac{1}{2}e^{-|y|}. \tag{2.68}$$

Now finally we can use our previous results to determine the conditional PDF as

$$\begin{aligned} p_{Y|X}(y|x) &= \frac{p_{X,Y}(x,y)}{p_X(x)} \\ &= \begin{cases} \frac{\frac{1}{2}e^{-x}}{xe^{-x}}, & \text{for } x \geq 0 \text{ and } -x \geq y \geq x \\ 0, & \text{otherwise} \end{cases} \\ &= \begin{cases} \frac{1}{2x}, & \text{for } x \geq 0 \text{ and } -x \geq y \geq x \\ 0, & \text{otherwise.} \end{cases} \end{aligned} \tag{2.69}$$

One could apply the axiom of total probability to all marginalized and conditional PDFs over their domain and should note that the axiom is satisfied.

2.10.5 Statistical characterization of pairs of random variables

When dealing with pairs of random variables, the expectation operator can be applied to understand how the two random variables are related to each other, by calculating quantities such as the covariance, correlation, and correlation coefficient.

Covariance

Previously, we have introduced the second central moment of a univariate random variable as the variance. While the variance denotes the spread in the univariate case, it has a different meaning in the multivariate case. Have a look at Figure 2.9 where two contour plots are shown for two distinct multivariate Gaussian PDFs. A more detailed mathematical description of the multivariate Gaussian distribution is provided in Section 3.2.4.

It can be noted that the distributions in Figure 2.9 are different since the first contour plot shows clear circles and the second contour plot shows tilted ellipses. For both distributions, 10000 random realizations $\mathbf{x} = [x_1, x_2]^\top$ are generated and the marginal distributions of both X_1 and X_2 are shown using a histogram and a fitted Gaussian distribution. It can be seen that the marginal distributions of X_1 and X_2 are exactly equal for both multivariate distributions, but still, the multivariate distributions are different. This difference can be explained by the covariance between the random variables X_1 and X_2.

The covariance is a measure of the relationship between two random variables. The right distribution in Figure 2.9 has a negative covariance between X_1 and X_2, because if X_1 increases as X_2 decreases. No such thing can be said about the left distribution in Figure 2.9, where X_1 and X_2 seem to have no relationship and behave independently from each other.

The formal definition of the **covariance** between two random variables X_1 and X_2 is given by

$$\text{Cov}[X_1, X_2] = \text{E}[(X_1 - \mu_{X_1})(X_2 - \mu_{X_2})], \tag{2.70}$$

which is very similar to the definition of variance; in fact, it actually represents the variance if

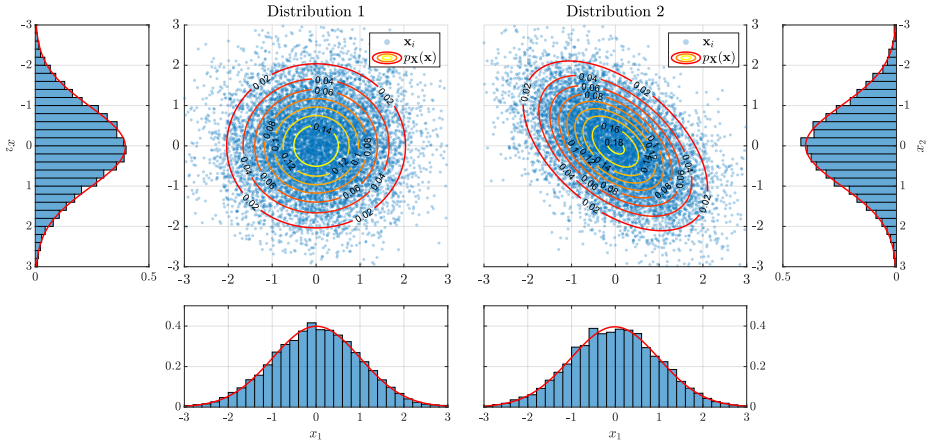


Figure 2.9: Examples of multivariate Gaussian PDFs, generated by sampling 10000 points in both cases, with zero covariance (left) and covariance different than zero (right). The marginal distributions of both the sample elements x_1 and x_2 are approximated using a histogram and a fitted Gaussian distribution. It can be seen that the marginal distribution of X_1 and X_2 are identical; however, the resulting multivariate distributions are different because of the different covariance in the two cases.

$X_1 = X_2$. Intuitively, one might regard the covariance as the expected value of the multiplication of X_1 and X_2 from which the means are subtracted. If both the normalized X_1 and X_2 have the same sign, then their multiplication would be positive and if X_1 and X_2 have different signs, then their multiplication would be negative. The covariance may therefore be regarded as a measure to indicate how X_2 behaves if X_1 increases or decreases.

Correlation

The definition of the covariance can be rewritten as

$$\begin{aligned}\mathrm{Cov}[X_1, X_2] &= \mathrm{E}[(X_1 - \mu_{X_1})(X_2 - \mu_{X_2})] \\ &= \mathrm{E}[X_1X_2 - \mu_{X_1}X_2 - \mu_{X_2}X_1 + \mu_{X_1}\mu_{X_2}] \\ &= \mathrm{E}[X_1X_2] - \mu_{X_1}\,\mathrm{E}[X_2] - \mu_{X_2}\,\mathrm{E}[X_1] + \mu_{X_1}\mu_{X_2} \\ &= \mathrm{E}[X_1X_2] - \mu_{X_1}\mu_{X_2} - \mu_{X_1}\mu_{X_2} + \mu_{X_1}\mu_{X_2} \\ &= \mathrm{E}[X_1X_2] - \mu_{X_1}\mu_{X_2}.\end{aligned} \tag{2.71}$$

The term $\mathrm{E}[X_1X_2]$ is called *correlation* r_{X_1,X_2} of X_1 and X_2 and is defined as

$$r_{X_1,X_2} = \mathrm{E}[X_1X_2]. \tag{2.72}$$

The correlation can be regarded as a noncentralized version of the covariance. These two terms are related through

$$\mathrm{Cov}[X_1, X_2] = r_{X_1,X_2} - \mu_{X_1}\mu_{X_2}. \tag{2.73}$$

It can be noted that the correlation and covariance of two random variables are equal if the mean value of one of the random variables is 0.

Two random variables are called *uncorrelated* if the covariance between them equals 0 as

$$\mathrm{Cov}[X_1, X_2] = 0. \tag{2.74}$$

Although the term suggests that it is related to the correlation between two random variables, it is defined as a zero covariance. On the other hand, two random variables are called *orthogonal* if the correlation between them equals 0 as

$$r_{X_1,X_2} = 0. \tag{2.75}$$

Correlation coefficient

The value of the covariance depends on the variance of both random variables and is therefore unbounded. To express the relationship between two random variables independent of the variances, the correlation coefficient normalizes the covariance as

$$\rho_{X_1,X_2} = \frac{\mathrm{Cov}[X_1, X_2]}{\sqrt{\mathrm{Var}[X_1]\,\mathrm{Var}[X_2]}} = \frac{\mathrm{Cov}[X_1, X_2]}{\sigma_{X_1}\sigma_{X_2}}. \tag{2.76}$$

Please note that this represents the normalized covariance and not the normalized correlation between two random variables, although the name suggests otherwise. Because of this normalization, the correlation coefficient has the property to be bounded between -1 and 1 as

$$-1 \leq \rho_{X_1,X_2} \leq 1. \tag{2.77}$$

In Figure 2.10, realizations of three different probability distributions with negative, zero, and positive correlation coefficients, respectively, are shown.

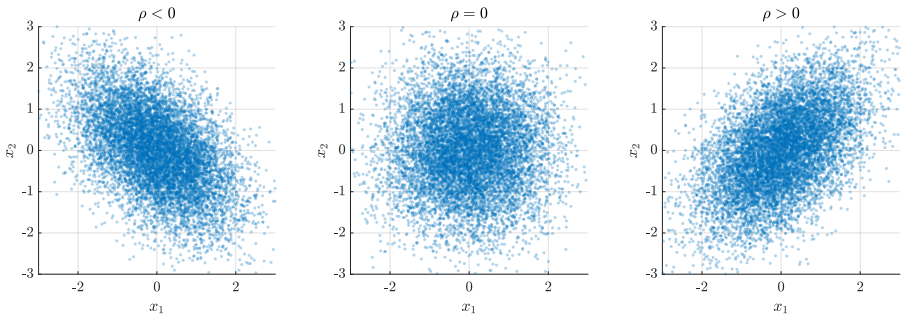


Figure 2.10: Scatter plots of random realizations of random variables X_1 and X_2 with negative, zero, and positive correlation coefficients.

Example 2.12

X and Y are identically distributed random variables with $\mathrm{E}[X] = \mathrm{E}[Y] = 0$, covariance $\mathrm{Cov}[X, Y] = 3$ and correlation coefficient $\rho_{X,Y} = 1/2$. For nonzero constants a and b, $U = aX$ and $V = bY$.

1. Find $\mathrm{Cov}[U, V]$.
2. Find the correlation coefficient $\rho_{U,V}$.

3. Let $W = U + V$. For what values of a and b are X and W uncorrelated?

Solution.

1. Since X and Y have zero expected value, $\mathrm{Cov}[X, Y] = \mathrm{E}[XY] = 3$, $\mathrm{E}[U] = a\,\mathrm{E}[X] = 0$ and $\mathrm{E}[V] = b\,\mathrm{E}[Y] = 0$. It follows that

$$\begin{aligned}\mathrm{Cov}[U, V] &= \mathrm{E}[UV] \\ &= \mathrm{E}[abXY] \\ &= ab\,\mathrm{E}[XY] = ab\,\mathrm{Cov}[X, Y] = 3ab.\end{aligned}$$

2. We start by observing that $\mathrm{Var}[U] = a^2\,\mathrm{Var}[X]$ and $\mathrm{Var}[V] = b^2\,\mathrm{Var}[Y]$. It follows that

$$\begin{aligned}\rho_{U,V} &= \frac{\mathrm{Cov}[U, V]}{\sqrt{\mathrm{Var}[U]\,\mathrm{Var}[V]}} \\ &= \frac{ab\,\mathrm{Cov}[X, Y]}{\sqrt{a^2\,\mathrm{Var}[X] b^2\,\mathrm{Var}[Y]}} = \frac{ab}{\sqrt{a^2 b^2}}\rho_{X,Y} = \frac{1}{2}\frac{ab}{|ab|}.\end{aligned}$$

   Note that $ab/|ab|$ is 1 if a and b have the same sign or is -1 if they have opposite signs.

3. Since $\mathrm{E}[X] = 0$,

$$\begin{aligned}\mathrm{Cov}[X, W] &= \mathrm{E}[XW] - \mathrm{E}[X]\,\mathrm{E}[U] \\ &= \mathrm{E}[XW] \\ &= \mathrm{E}[X(aX + bY)] \\ &= a\,\mathrm{E}[X^2] + b\,\mathrm{E}[XY] \\ &= a\,\mathrm{Var}[X] + b\,\mathrm{Cov}[X, Y].\end{aligned}$$

   Since X and Y are identically distributed, $\mathrm{Var}[X] = \mathrm{Var}[Y]$ and

$$\frac{1}{2} = \rho_{X,Y} = \frac{\mathrm{Cov}[X, Y]}{\sqrt{\mathrm{Var}[X]\,\mathrm{Var}[Y]}} = \frac{\mathrm{Cov}[X, Y]}{\mathrm{Var}[X]} = \frac{3}{\mathrm{Var}[X]}.$$

   This implies $\mathrm{Var}[X] = 6$. From (3), $\mathrm{Cov}[X, W] = 6a + 3b = 0$, or $b = -2a$.

2.10.6 Central limit theorem

We have briefly discussed what happens with the statistics of a random variable, especially the mean and variance, when a random variable is linearly transformed. Now let us focus on the specific situation where we take the sum of N independent random variables. We will define a random variable Y as

$$Y = \frac{X_1 + X_2 + \ldots + X_N}{N} \tag{2.78}$$

and have a look at the PDF of Y, which is the (normalized) sum of N independent random variables.

In Figure 2.11, 100000 samples are generated from three different probability distributions and the resulting histograms are plotted. This is repeated for an increasing number of N, and the histograms are plotted after averaging N similarly generated sets of data. It can be noted that after averaging multiple realizations of a random variable generated from the same (arbitrary) distribution, the distribution of the averaged random variables converges to a Gaussian distribution. This result is known as the central limit theorem.

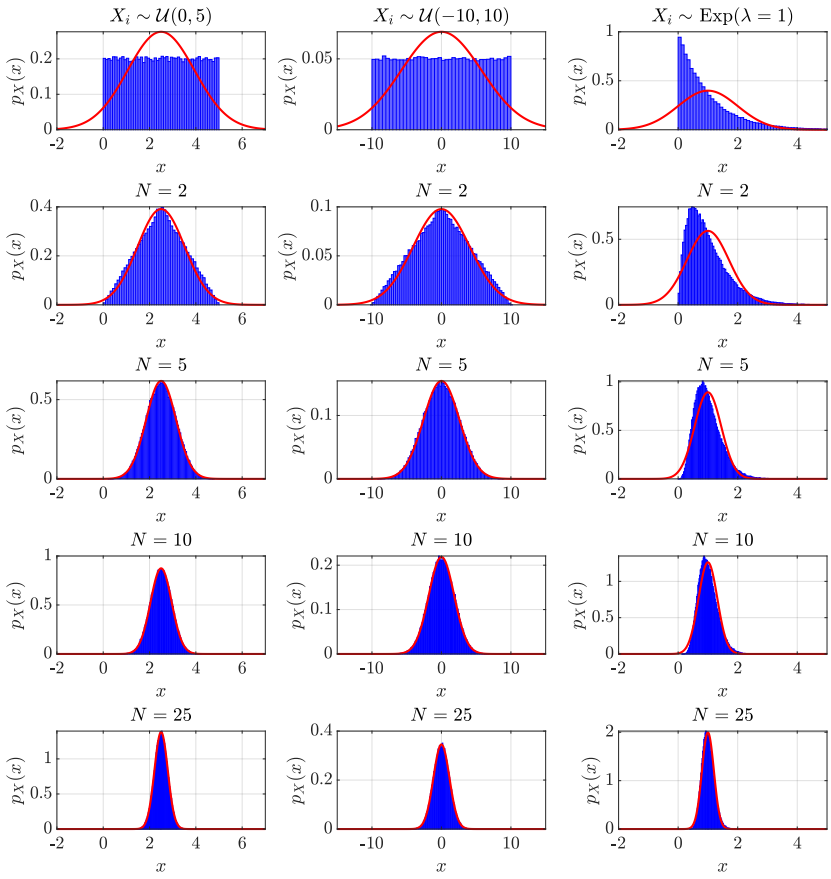


Figure 2.11: Demonstration of the CLT for three different probability distributions, i.e. $\mathcal{U}(0,5)$, $\mathcal{U}(-10,10)$ and $\text{Exp}(\lambda = 1)$, where N sets of 100000 realizations are averaged and the new distributions of the averaged 100000 samples are shown for different values of N. A Gaussian function is fitted over the distributions and it can be seen that after averaging over more realizations, the new distribution will converge to the Gaussian distribution.

The formal definition is given as

Theorem. Let $X_1, X_2, ..., X_n$ be a set of N independent identically-distributed (i.i.d) random

variables and each X_i has an arbitrary probability distribution $p(x_1, x_2, ..., x_n)$ with finite mean $\mu_i = \mu$ and finite standard deviation $\sigma_i = \sigma$. If the sample size N is "sufficiently large", then the CDF of the sum converges to a Gaussian CDF.

In simple words, the CLT states that the normalized sum of a sufficient number of i.i.d. random variables tends to a Gaussian distribution. Under certain conditions, the CLT can be extended for the sum of random variables that are independent but not necessarily identically distributed (i.e., Lyapunov condition). In both cases, the mean or expected value of the approximate Gaussian PDF can be determined as

$$\mathrm{E}[Y] = \frac{1}{N} \sum_{i=1}^{N} \mathrm{E}[X_i], \tag{2.79}$$

which is simply the average value of all individual expected values. Similarly, the variance can be determined as

$$\mathrm{Var}[Y] = \sigma_Y^2 = \frac{1}{N^2} \sum_{i=1}^{N} \sigma_{X_i}^2 = \frac{1}{N^2} \sum_{i=1}^{N} \mathrm{Var}[X_i]. \tag{2.80}$$

The minimum value of N for the CLT approximation to be valid depends on the distribution of the individual random variables. In general, convergence is faster when the summed random variables are i.i.d and the generating distribution is symmetric.

3

Random processes and random signals

3.1 Introduction

When processing and analyzing signals, the noise component makes it more difficult to draw meaningful conclusions from specific signal samples. In order to cope adequately with the uncertainty involved, deterministic signals can better be regarded as random signals, where the exact outcome in time is unknown, but where conclusions can be drawn from the statistical properties of the signal.

In signal processing, we deal with measured signals which are often not ideal. Usually, the signals are corrupted by noise and the exact values of this noise at a certain moment in time cannot be determined. A common source of noise is thermal noise, originating from the electronic components in the measurement equipment. Electronic components inevitably generate noise, because of the free electrons in the materials. These electrons have a certain energy, which is related to the temperature of the material. A higher temperature means that the electrons have a higher energy content. This energy is usually in the form of kinetic energy, meaning that the free electrons have a certain velocity associated with them. Only at a temperature of 0 Kelvin do the electrons not have kinetic energy and therefore they do not move. This random electron movement occurring at non-zero temperatures leads to small local voltage differences, which distorts the signal that is being measured and is commonly known as thermal noise.

In this section, we will discuss the tools from probability theory and random variables that are necessary to deal with random signals. In fact, the ultimate subject of investigation is the process of generating these signals, while the measured signals can be interpreted as individual realizations of these random processes. Thus, we will formally introduce random processes and characterize them with the tools provided by probability theory and random variables.

3.2 Random vectors

Before we can deal with random signals, it is important to extend the concepts we have learned for pairs of random variables to multiple random variables by introducing random vectors. In this case, the outcome of an experiment comprises N observed quantities. An example of such an observation is several variables measured at the hospital to monitor the health of a patient

(heart rate, blood pressure, saturation, etc...) or the different variables measured at a weather station to make weather predictions (temperature, relative humidity, pressure, rain, fall, etc.).

3.2.1 Multivariate joint probability distributions

Let us denote each of these measured quantities by the random variable X_n, where n ranges from 1 to N. Using this definition, the multivariate (meaning that multiple variables are involved) joint probability functions can be introduced. For notation purposes, all random variables X_n can be grouped in a random vector $\mathbf{X} = [X_1, X_2, \ldots, X_N]^T$, where the $\cdot^T$ operator denotes the transpose, turning this row vector into a column vector. The bold capital letter distinguishes the random vector containing multiple random variables from a single random variable. Similarly, a specific realization of this random vector can be written in lower-case as $\mathbf{x} = [x_1, x_2, \ldots, x_N]^T$.

Multivariate joint cumulative distribution function

The multivariate joint CDF of the random vector $\mathbf{X}$ containing random variables $X_1, X_2, \ldots, X_N$ is defined as

$$P_{\mathbf{X}}(\mathbf{x}) = P_{X_1,\ldots,X_N}(x_1, \ldots, x_N) = \Pr[X_1 \leq x_1, \ldots, X_N \leq x_N]. \tag{3.1}$$

This definition holds for both discrete and continuous random variables.

Multivariate joint probability mass function

The multivariate joint PMF of the random vector $\mathbf{X}$ containing discrete random variables $X_1, X_2, \ldots, X_N$ is similarly defined as

$$p_{\mathbf{X}}(\mathbf{x}) = p_{X_1,\ldots,X_N}(x_1, \ldots, x_N) = \Pr[X_1 = x_1, \ldots, X_N = x_N]. \tag{3.2}$$

Multivariate joint probability density function

The multivariate joint PDF of the random vector $\mathbf{X}$ containing continuous random variables $X_1, X_2, \ldots, X_N$ is defined from the multivariate joint CDF as

$$p_{\mathbf{X}}(\mathbf{x}) = p_{X_1,\ldots,X_N}(x_1, \ldots, x_N) = \frac{\partial^N P_{X_1,\ldots,X_N}(x_1, \ldots, x_N)}{\partial x_1 \ldots \partial x_N}. \tag{3.3}$$

As can be observed, the multivariate probability distributions are a straightforward extension of the cases for pairs of random variables.

Generalized probability axioms for multivariate joint distributions

From these definitions, several multivariate joint probability axioms can be determined, which are similar to the case of two random variables.

1. It holds that $p_{\mathbf{X}}(\mathbf{x}) \geq 0$, where $\mathbf{X}$ is a continuous or discrete random vector;
2. From the multivariate joint PDF it follows that

$$P_{X_1,\ldots,X_N}(x_1, \ldots, x_N) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_N} p_{X_1,\ldots,X_N}(x_1, \ldots, x_N)\, dx_1 \ldots dx_N;$$

holds for continuous random vectors.

3. Through the law of total probability it holds that

$$\sum_{x_1 \in S_{X_1}} \cdots \sum_{x_N \in S_{X_N}} p_{X_1,\ldots,X_N}(x_1,\ldots,x_N) = 1$$

for discrete random vectors and

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} p_{X_1,\ldots,X_N}(x_1,\ldots,x_N)\ dx_N \ldots dx_1 = 1$$

for continuous random vectors.

4. The probability of an event A can be determined as: $\Pr[A] = \sum_{\mathbf{X} \in A} p_{X_1,\ldots,X_N}(x_1,\ldots,x_N)$ for discrete random variables and $\Pr[A] = \int_A \cdots \int p_{X_1,\ldots,X_N}(x_1,\ldots,x_N)\ dx_1 \ldots dx_N$ for continuous random variables.

Axiom 1 simply states that a probability (density) cannot be smaller than 0, since no negative probabilities exist by definition. The second axiom is a direct consequence of integrating both sides of the multivariate joint PDF allowing us to determine the multivariate joint CDF from the multivariate joint PDF. The third axiom is a direct consequence of the law of total probability, where the probability of all events together equal 1. The final axiom tells us to sum, or integrate, over all possible outcomes of an event A in order to calculate its probability.

Probability distributions of multiple random vectors

The notation of a random vector allows us to easily include multiple random variables in a single vector. Suppose now that our random vector $\mathbf{Z}$ contains two different types of random variables, where for example each random variable corresponds to a different type of measurement. If we were to distinguish between these types of random variables using two generalized random variables X_i and Y_i, the random vector $\mathbf{Z}$ could be written as $\mathbf{Z} = [X_1, X_2, \ldots, X_N, Y_1, Y_2, \ldots, Y_M]^T$. If we now were to define the random vectors
$\mathbf{X} = [X_1, X_2, \ldots, X_N]^T$ and $\mathbf{Y} = [Y_1, Y_2, \ldots, Y_M]^T$, it becomes evident that we could simplify the random vector $\mathbf{Z}$ as $\mathbf{Z} = [\mathbf{X}^T, \mathbf{Y}^T]^T$.

This shows that it is also possible for joint probability distributions to depend on multiple random vectors, with each vector including a subset of all random variables. This can prove useful in some cases when there is a clear distinction between the subsets and is purely for notation purposes. A probability distribution depending on multiple random vectors can be regarded in all aspects as a probability distribution depending on a single random vector (which is a concatenation of all different random variables). All calculations can be performed by regarding the probability distribution as if it depends on a single (concatenated) random vector. A probability distribution involving multiple random vectors can be written for example as $p_{\mathbf{X},\mathbf{Y}}(\mathbf{x}, \mathbf{Y})$.

3.2.2 Conditional and marginal probabilities

Similarly, as in the case of pairs of random variables, the *conditional probability* can be determined by normalizing the joint probability with the probability of the conditional event through

$$p_{\mathbf{X}|B}(\mathbf{x}) = \begin{cases} \frac{p_{\mathbf{X}}(\mathbf{x})}{\Pr[B]}, & \text{when } \mathbf{x} \in B, \\ 0, & \text{otherwise.} \end{cases} \tag{3.4}$$

Because the notation of a random vector $\mathbf{X}$ is just a shorter notation for the set of random variables $X_1, X_2, \ldots, X_N$, it is possible to calculate the *marginal probability* distribution of a subset of random variables. This subset can also just consist of a single random variable. Again this operation is performed through marginalization as discussed for pairs of random variables. For example, to obtain the marginal probability distribution $p_{X_2,X_3}(x_2, x_3)$, given the probability distribution $p_{\mathbf{X}}(\mathbf{x})$ we can apply the following operation

$$p_{X_2,X_3}(x_2, x_3) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} p_{\mathbf{X}}(\mathbf{x})\ dx_1 dx_4 \cdots dx_N \tag{3.5}$$

for continuous random variables, and

$$p_{X_2,X_3} = \sum_{x_1 \in S_{X_1}} \sum_{x_4 \in S_{X_4}} \cdots \sum_{x_N \in S_{X_N}} p_{\mathbf{X}}(\mathbf{x}) \tag{3.6}$$

for discrete random variables. Here we have integrated or summed over all possible values of all random variables except for the ones that we are interested in.

3.2.3 Independence

Independence is a term in probability theory that reflects that the probability of an event A is not changed after observing an event B, meaning that $\Pr[A|B] = \Pr[A]$. In other words, the occurrence of an event B has no influence on the probability of an event A. Keep in mind that this does not mean that the physical occurrence of event A and B are unrelated, it just means that the probability of the occurrence of event A is unrelated to whether event B occurs or not.

Independent random variables

This notion of independence can be extended to probability functions. The random variables $X_1, X_2, \ldots, X_N$ can be regarded as independent if and only if the following factorization holds

$$p_{X_1,X_2,\ldots,X_N}(x_1, x_2, \ldots, x_N) = p_{X_1}(x_1) p_{X_2}(x_2) \ldots p_{X_N}(x_N). \tag{3.7}$$

This equation states that the total joint probability distribution can be written as a multiplication of the individual probability distribution of each random variable. From a probability point of view (not a physical one) we can conclude that the random variables are independent because the total probability solely depends on all the individual contributions of the random variables. Random variables that satisfy the independence equation and are distributed under the same PDF are regarded as independent and identically distributed (IID or iid) random variables.

Independent random vectors

It is also possible to extend the definition of independence to random vectors. Two random vectors $\mathbf{X}$ and $\mathbf{Y}$ can be regarded as independent if and only if the probability function can be written as

$$p_{\mathbf{X},\mathbf{Y}}(\mathbf{x}, \mathbf{Y}) = p_{\mathbf{X}}(\mathbf{x})\ p_{\mathbf{Y}}(\mathbf{Y}). \tag{3.8}$$

3.2.4 Statistical characterization of random vectors

In Section 2.7.5, we characterized random variables using moments. This section will extend this characterization to random vectors.

Expected value

The expected value of a random vector $\mathbf{X}$ is defined as the vector containing the expected values of the individual random variables $X_1, X_2, \ldots, X_N$ as

$$\mathrm{E}[\mathbf{X}] = \mu_{\mathbf{X}} = [\mu_1, \mu_2, \ldots, \mu_N]^T = [\mathrm{E}[X_1], \mathrm{E}[X_2], \ldots, \mathrm{E}[X_N]]^T . \tag{3.9}$$

Expected value of a function

When we are interested in the expected value of a certain function $g(\mathbf{X})$, which accepts a random vector as an argument and transforms it into another vector, this can be determined by multiplying the function's result with its corresponding probability and summing or integrating over all possible realizations of $\mathbf{X}$. For a discrete random vector $\mathbf{X}$ consisting of random variables $X_1, X_2, \ldots, X_N$ the expected value of a function $g(\mathbf{X})$ can be determined as

$$\mathrm{E}[g(\mathbf{X})] = \sum_{x_1 \in S_{X_1}} \cdots \sum_{x_N \in S_{X_N}} g(\mathbf{x}) \cdot p_{\mathbf{X}}(\mathbf{x}) \tag{3.10}$$

and for a continuous random vector as

$$\mathrm{E}[g(\mathbf{X})] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(\mathbf{x}) \cdot p_{\mathbf{X}}(\mathbf{x}) \; dx_1 \ldots dx_N . \tag{3.11}$$

Covariance matrix

We previously discussed how we could determine the covariance of two random variables. Let us turn now to the covariance of two random vectors $\mathbf{X} = [X_1, X_2, \ldots, X_N]^T$ and $\mathbf{Y} = [Y_1, Y_2, \ldots, Y_N]^T$. Intuitively, one might say that this covariance cannot be described by a single number because there is more than one combination of random variables of which we want to calculate the covariance. As an example, we could determine the covariances of X_1 and Y_1, X_1 and Y_2 and X_N and Y_1. All of these possible combinations can be gathered in the so-called cross-covariance matrix $\mathbf{C_{XY}}$.

The covariance matrix is formally defined as

$$\mathbf{C_{XY}} = \mathrm{E}[(\mathbf{X} - \mu_{\mathbf{X}})(\mathbf{Y} - \mu_{\mathbf{Y}})^T] = \begin{bmatrix} C_{11} & C_{12} & \cdots & C_{1N} \\ C_{21} & C_{22} & \cdots & C_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ C_{N1} & C_{N2} & \cdots & C_{NN} \end{bmatrix}, \tag{3.12}$$

where the individual coefficients correspond to

$$C_{nm} = \mathrm{Cov}[X_n, Y_m] = \mathrm{E}[(X_n - \mu_{X_n})(Y_m - \mu_{Y_m})]. \tag{3.13}$$

The transpose operator in the first equation creates a matrix from the two column vectors filled with the covariances of all possible combinations of random variables. For each of these covariances, the correlation coefficient ρ_{nm} can be calculated similarly using the definition of the correlation coefficient. Two random vectors are called uncorrelated if $\mathbf{C_{XY}} = 0$.

For the special case that $\mathbf{X} = \mathbf{Y}$, the cross-covariance matrix is called the *autocovariance matrix*, which calculates the covariances between all random variables in $\mathbf{X}$. The definition is the same as the definition of the covariance matrix, where $\mathbf{C_{XX}}$ is often simplified as $\mathbf{C_X}$. The main diagonal of the autocovariance matrix is represented by the variances of all random variables in $\mathbf{X}$.

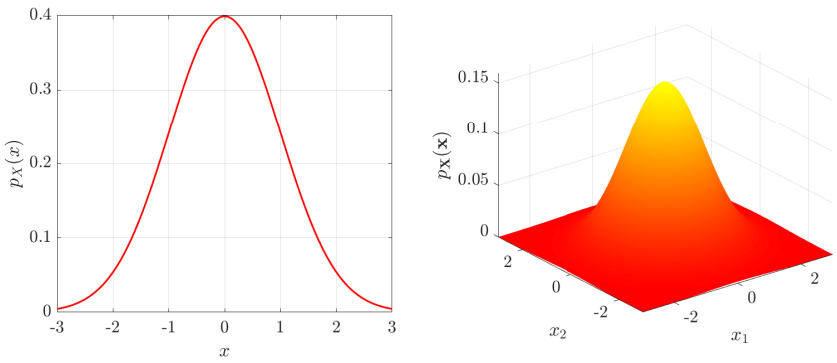


Figure 3.1: Example of a univariate (left) and a multivariate (right) Gaussian PDF.

Multivariate Gaussian distribution

In the case of a single random variable X that is generated according to a Gaussian distribution, defined by its mean μ and variance σ^2, the PDF is defined as

$$p_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \tag{3.14}$$

The left side of Figure 3.1 below shows an example of such univariate Gaussian distribution.

The definition of the univariate Gaussian distribution can be extended to a multivariate distribution. To define the Gaussian distribution the position and its spread are required. These quantities are represented by the mean vector μ and the covariance matrix $\mathbf{\Sigma}$, respectively. Whereas the covariance matrix is defined as $\mathbf{C}$, literature has adopted the $\mathbf{\Sigma}$ notation when discussing multivariate Gaussian distributions, as Σ is the Greek capital letter of σ.

To indicate that a k-dimensional random vector $\mathbf{X}$ is Gaussian distributed, we can write $\mathbf{X} \sim \mathcal{N}_k(\mu, \mathbf{\Sigma})$. The PDF of such a multivariate Gaussian distribution is defined as

$$p_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^k |\mathbf{\Sigma}|}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)^{\mathbf{T}} \mathbf{\Sigma}^{-\mathbf{1}} (\mathbf{x}-\mu)\right\}, \tag{3.15}$$

where $|\mathbf{\Sigma}|$ is the determinant of the covariance matrix. Please note the similarities between the univariate Gaussian distribution and the multivariate distribution. The inverse covariance matrix $\mathbf{\Sigma}^{-\mathbf{1}}$ is often also called the precision matrix, because a low variance (i.e. low spread) relates to a high precision and vice versa, and is denoted by $\mathbf{\Lambda}$.

The covariance matrix of a multivariate Gaussian distribution

The PDF of a Gaussian distribution is fully determined by its mean μ and its covariance matrix $\mathbf{\Sigma}$. In order to give some intuition on how the mean and covariance matrix structure influence the final distribution, we take a look at Figure 3.2 where two multivariate distributions have been plotted. The covariance matrices that were used to plot these distributions in the figure are from left to right:

$$\mathbf{\Sigma_1} = \begin{bmatrix} 25 & 0 \\ 0 & 4 \end{bmatrix}, \qquad \mathbf{\Sigma_2} = \begin{bmatrix} 14.5 & 10.5 \\ 10.5 & 14.5 \end{bmatrix}. \tag{3.16}$$

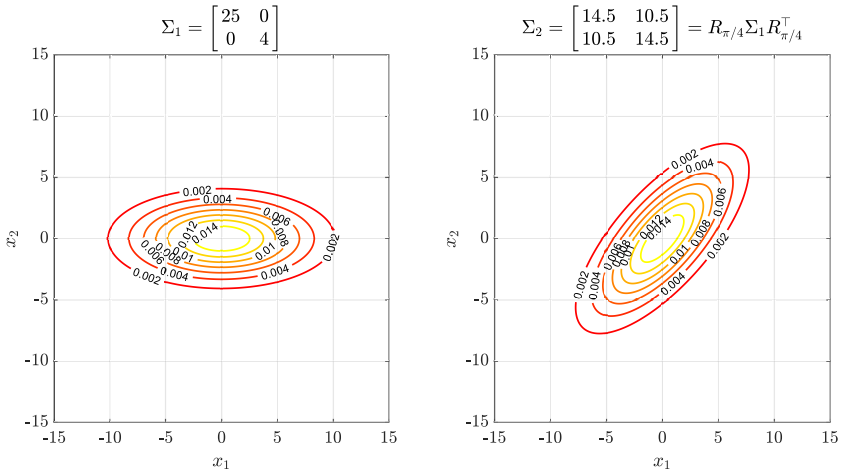


Figure 3.2: Two multivariate Gaussian distributions whose covariance matrices are related through the rotation matrices corresponding to a counter-clockwise rotation of $\pi/4$ radians.

Please note how the off-diagonal entries, referring to $\text{Cov}[X_1, X_2]$ and $\text{Cov}[X_2, X_1]$ influence the shape of the distribution.

In order to understand how the covariance matrix is related to the tilt and the shape of the distribution, we need to first introduce the so-called rotation matrix and the eigenvalue decomposition. The rotation matrix R_θ rotates a coordinate counter-clockwise over an angle θ with respect to the origin. This rotation matrix is defined as

$$R_\theta = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \tag{3.17}$$

and a rotation of θ from the coordinates (x, y) to the coordinates (x', y') can be represented by

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = R_\theta \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x\cos(\theta) - y\sin(\theta) \\ x\sin(\theta) + y\cos(\theta) \end{bmatrix}. \tag{3.18}$$

One of the properties of a rotation matrix is that it is orthogonal. This means that $R_\theta R_\theta^T = I$, where I is the identity matrix. Using the fact that $R_\theta^{-1} = R_{-\theta} = R_\theta^T$ from its definition, the orthogonality property makes complete sense, because rotating a coordinate with the angle $-\theta$ and θ respectively does not change anything.

Besides the rotation matrices, we need to introduce the eigenvalue decomposition in order to better understand the covariance matrix structure. The eigenvalue decomposition states that a square invertible symmetric matrix A can be written as

$$A = Q\Lambda Q^{-1}, \tag{3.19}$$

where the orthogonal matrix Q contains the eigenvectors of A and Λ is a diagonal matrix containing the eigenvalues of A.

Now the general representation of the rotation matrix has been defined as well as the eigenvalue decomposition, we can show that any covariance matrix can be written as the rotations

of a diagonal covariance matrix. This point is very important to understand. To start off, a diagonal covariance matrix can be represented as

$$\mathbf{\Sigma_d} = \begin{bmatrix} a & 0 \\ 0 & b \end{bmatrix}. \tag{3.20}$$

The entries a and b correspond to the individual variances of X_1 and X_2 and are at the same time the eigenvalues of $\mathbf{\Sigma_d}$. An example of a Gaussian distribution that corresponds to a diagonal covariance matrix where $a = 25$ and $b = 4$ is shown on the left in Fig 3.2. Please note that the ratio of $\sqrt{a}$ and $\sqrt{b}$ also represents the ratio of the length (the major axis) and the width (the minor axis) of the distribution.

If we were to apply the eigenvalue decomposition to a covariance matrix $\mathbf{\Sigma}$, we would find that

$$\mathbf{\Sigma} = \mathbf{R}_\theta \mathbf{\Sigma_d} \mathbf{R}\theta^{\mathbf{T}}. \tag{3.21}$$

The right plot of Figure 3.2 shows an example of a multivariate Gaussian distribution whose covariance matrix is a rotated version of the diagonal covariance matrix corresponding to the left side of the same figure. From this we can see that the ratio of eigenvalues of $\mathbf{\Sigma}$ corresponds to the ratio of the lengths of the major and minor axes. Furthermore, we can conclude that the matrix containing the eigenvectors of $\mathbf{\Sigma}$ is at the same time a rotation matrix, implicitly defining the rotation angle.

Example 3.1

An n-dimensional Gaussian vector $\mathbf{W}$ has a block diagonal covariance matrix

$$\mathbf{C_W} = \begin{bmatrix} \mathbf{C_X} & \mathbf{0} \\ \mathbf{0} & \mathbf{C_Y} \end{bmatrix}, \tag{3.22}$$

where $\mathbf{C_X}$ is $m \times m$, $\mathbf{C_Y}$ is $(n-m) \times (n-m)$. Show that $\mathbf{W}$ can be written in terms of component vectors $\mathbf{X}$ and $\mathbf{Y}$ in the form

$$\mathbf{W} = \begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix}, \tag{3.23}$$

such that $\mathbf{X}$ and $\mathbf{Y}$ are independent Gaussian random vectors.

Solution.

As given in the problem statement, we define the m-dimensional vector $\mathbf{X}$, the n-dimensional vector $\mathbf{Y}$ and $\mathbf{W} = \begin{bmatrix} \mathbf{X}^T, & \mathbf{Y}^T \end{bmatrix}^T$.

Note that $\mathbf{W}$ has an expected value of

$$\overline{\mu_W} = \mathrm{E}[\mathbf{W}] = \mathrm{E}\begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix} = \begin{bmatrix} \mathrm{E}[\mathbf{X}] \\ \mathrm{E}[\mathbf{Y}] \end{bmatrix} = \begin{bmatrix} \overline{\mu_X} \\ \overline{\mu_Y} \end{bmatrix}.$$

The covariance matrix of $\mathbf{W}$ is

$$\begin{aligned}\mathbf{C_W} &= \mathrm{E}[(\mathbf{W}-\mu_W)(\mathbf{W}-\mu_W)'] \\ &= \mathrm{E}\left[\begin{bmatrix}\mathbf{X}-\overline{\mu_X}\\ \mathbf{Y}-\overline{\mu_Y}\end{bmatrix}[(\mathbf{X}-\overline{\mu_X})' \quad (\mathbf{Y}-\overline{\mu_Y})']\right] \\ &= \begin{bmatrix}\mathrm{E}[(\mathbf{X}-\overline{\mu_X})(\mathbf{X}-\overline{\mu_X})'] & \mathrm{E}[(\mathbf{X}-\overline{\mu_X})(\mathbf{Y}-\overline{\mu_Y})'] \\ \mathrm{E}[(\mathbf{Y}-\overline{\mu_Y})(\mathbf{X}-\overline{\mu_X})'] & \mathrm{E}[(\mathbf{Y}-\overline{\mu_Y})(\mathbf{Y}-\overline{\mu_Y})']\end{bmatrix} \\ &= \begin{bmatrix}\mathbf{C_X} & \mathbf{C_{XY}} \\ \mathbf{C_{YX}} & \mathbf{C_Y}\end{bmatrix}.\end{aligned}$$

The assumption that $\mathbf{X}$ and $\mathbf{Y}$ are independent implies that

$$\mathbf{C_{XY}} = \mathrm{E}[(\mathbf{X}-\overline{\mu_X})(\mathbf{Y}'-\overline{\mu_Y}')] = \mathrm{E}[(\mathbf{X}-\overline{\mu_X})]\,\mathrm{E}[(\mathbf{Y}-\overline{\mu_Y})'] = \mathbf{0}.$$

This also implies that $\mathbf{C_{YX}} = \mathbf{C}'_{\mathbf{XY}} = \mathbf{0}$. Thus,

$$\mathbf{C_W} = \begin{bmatrix}\mathbf{C_X} & \mathbf{0} \\ \mathbf{0} & \mathbf{C_Y}\end{bmatrix}.$$

Correlation matrix

Similarly, to the cross-covariance matrix, the *cross-correlation matrix* can be defined as containing the correlations of all combinations between the random variables in $\mathbf{X}$ and $\mathbf{Y}$. The cross-correlation matrix of random vectors $\mathbf{X}$ and $\mathbf{Y}$ is denoted by $\mathbf{R_{XY}}$ and is defined as

$$\mathbf{R_{XY}} = \mathrm{E}[\mathbf{X}\mathbf{Y}^{\mathbf{T}}] = \begin{bmatrix}r_{11} & r_{12} & \cdots & r_{1N} \\ r_{21} & r_{22} & \cdots & r_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ r_{N1} & r_{N2} & \cdots & r_{NN}\end{bmatrix}, \tag{3.24}$$

where the individual coefficients correspond to the individual correlations

$$r_{nm} = \mathrm{E}[X_n Y_m]. \tag{3.25}$$

Two random vectors are called orthogonal if $\mathbf{R_{XY}} = 0$. Furthermore, it can be proven that the cross-covariance matrix and the cross-correlation matrix are related through

$$\mathbf{C_{XY}} = \mathbf{R_{XY}} - \mu_{\mathbf{X}}\mu_{\mathbf{Y}}^{\mathbf{T}}. \tag{3.26}$$

For the special case that $\mathbf{X} = \mathbf{Y}$, the cross-correlation matrix is called the *autocorrelation matrix*, which calculates the correlations between all random variables in $\mathbf{X}$. The definition is the same as the definition of the cross-correlation matrix, where $\mathbf{R_{XX}}$ is often simplified as $\mathbf{R_X}$.

3.2.5 Linear transformations of random vectors

Let us define an invertible transformation matrix $\mathbf{A}$, with dimensions $(N \times N)$, which will linearly map a random vector $\mathbf{X}$ of length N to a random vector $\mathbf{Y}$ again with length N after adding an equally long column vector $\mathbf{b}$ through

$$\mathbf{Y} = g(\mathbf{X}) = \mathbf{A}\mathbf{X} + \mathbf{b}. \tag{3.27}$$

Probability density function

From the initial multivariate PDF of $\mathbf{X}$, $p_{\mathbf{X}}(\mathbf{x})$, the new PDF of $\mathbf{Y}$, $p_{\mathbf{Y}}(\mathbf{Y})$, can be determined as

$$p_{\mathbf{Y}}(\mathbf{Y}) = \frac{p_{\mathbf{X}}(g^{-1}(\mathbf{Y}))}{|\det \mathbf{A}|} = \frac{p_{\mathbf{X}}(\mathbf{A}^{-1}(\mathbf{y} - \mathbf{b}))}{|\det \mathbf{A}|}, \tag{3.28}$$

where $|\det \mathbf{A}|$ is the absolute value of the determinant of $\mathbf{A}$.

Mean vector

The new mean vector of the random vector $\mathbf{Y}$ can be determined as

$$\mu_{\mathbf{Y}} = \mathrm{E}[\mathbf{Y}] = \mathrm{E}[\mathbf{A}\mathbf{X} + \mathbf{b}] = \mathbf{A}\,\mathrm{E}[\mathbf{X}] + \mathbf{b} = \mathbf{A}\mu_{\mathbf{X}} + \mathbf{b}. \tag{3.29}$$

Cross-covariance and cross-correlation matrices

By the definition of the cross-covariance matrix, the cross-covariance matrices $\mathbf{C_{XY}}$ and $\mathbf{C_{YX}}$ can be determined from the original autocovariance matrix $\mathbf{C_X}$ of $\mathbf{X}$ through

$$\begin{aligned}
\mathbf{C_{XY}} &= \mathrm{E}[(\mathbf{X} - \mu_{\mathbf{X}})(\mathbf{Y} - \mu_{\mathbf{Y}})^T] \\
&= \mathrm{E}[(\mathbf{X} - \mu_{\mathbf{X}})(\mathbf{A}\mathbf{X} + \mathbf{b} - (\mathbf{A}\mu_{\mathbf{X}} + \mathbf{b}))^T] \\
&= \mathrm{E}[(\mathbf{X} - \mu_{\mathbf{X}})(\mathbf{A}(\mathbf{X} - \mu_{\mathbf{X}}))^T] \\
&= \mathrm{E}[(\mathbf{X} - \mu_{\mathbf{X}})(\mathbf{X} - \mu_{\mathbf{X}})^T\mathbf{A}^T] \\
&= \mathrm{E}[(\mathbf{X} - \mu_{\mathbf{X}})(\mathbf{X} - \mu_{\mathbf{X}})^T]\mathbf{A}^T = \mathbf{C_X}\mathbf{A}^T
\end{aligned} \tag{3.30}$$

and similarly, we can find the result

$$\mathbf{C_{YX}} = \mathbf{A}\mathbf{C_X}. \tag{3.31}$$

The new cross-correlation matrices $\mathbf{R_{XY}}$ and $\mathbf{R_{YX}}$ can be determined as

$$\begin{aligned}
\mathbf{R_{XY}} &= \mathrm{E}[\mathbf{X}\mathbf{Y}^{\mathbf{T}}] \\
&= \mathrm{E}[\mathbf{X}(\mathbf{A}\mathbf{X} + \mathbf{b})^T] \\
&= \mathrm{E}[\mathbf{X}(\mathbf{A}\mathbf{X})^T + \mathbf{X}\mathbf{b}^T] \\
&= \mathrm{E}[\mathbf{X}\mathbf{X}^T\mathbf{A}^T] + \mathrm{E}[\mathbf{X}]\mathbf{b}^T \\
&= \mathbf{R_X}\mathbf{A}^T + \mu_{\mathbf{X}}\mathbf{b}^T
\end{aligned} \tag{3.32}$$

and similarly as

$$\mathbf{R_{YX}} = \mathbf{A}\mathbf{R_X} + \mathbf{b}\mu_{\mathbf{X}}^{\mathbf{T}}. \tag{3.33}$$

Autocovariance and autocorrelation matrix

The autocovariance matrix of $\mathbf{Y}$ can be determined through

$$\begin{aligned}\mathbf{C_Y} &= \mathrm{E}[(\mathbf{Y}-\mu_\mathbf{Y})(\mathbf{Y}-\mu_\mathbf{Y})^T] \\ &= \mathrm{E}[(\mathbf{AX}+\mathbf{b}-(\mathbf{A}\mu_\mathbf{X}+\mathbf{b}))(\mathbf{AX}+\mathbf{b}-(\mathbf{A}\mu_\mathbf{X}+\mathbf{b}))^T] \\ &= \mathrm{E}[(\mathbf{A}(\mathbf{X}-\mu_\mathbf{X}))(\mathbf{A}(\mathbf{X}-\mu_\mathbf{X}))^T] \\ &= \mathrm{E}[\mathbf{A}(\mathbf{X}-\mu_\mathbf{X})(\mathbf{X}-\mu_\mathbf{X})^T\mathbf{A}^T] \\ &= \mathbf{A}\,\mathrm{E}[(\mathbf{X}-\mu_\mathbf{X})(\mathbf{X}-\mu_\mathbf{X})^T]\mathbf{A}^T = \mathbf{A}\mathbf{C_X}\mathbf{A}^T.\end{aligned} \tag{3.34}$$

In a similar fashion, the new autocorrelation matrix of $\mathbf{Y}$ can be calculated as

$$\begin{aligned}\mathbf{R_Y} &= \mathrm{E}[\mathbf{YY}^T] \\ &= \mathrm{E}[(\mathbf{AX}+\mathbf{b})(\mathbf{AX}+\mathbf{b})^T] \\ &= \mathrm{E}[(\mathbf{AX}+\mathbf{b})(\mathbf{X}^T\mathbf{A}^T+\mathbf{b}^T)] \\ &= \mathrm{E}[\mathbf{AXX}^T\mathbf{A}^T+\mathbf{AXb}^T+\mathbf{bX}^T\mathbf{A}^T+\mathbf{bb}^T] \\ &= \mathbf{A}\,\mathrm{E}[\mathbf{XX}^T]\mathbf{A}^T+\mathbf{A}\,\mathrm{E}[\mathbf{X}]\mathbf{b}^T+\mathbf{b}\,\mathrm{E}[\mathbf{X}^T]\mathbf{A}^T+\mathbf{bb}^T \\ &= \mathbf{AR_X}\mathbf{A}^T+\mathbf{A}\mu_\mathbf{X}\mathbf{b}^T+\mathbf{b}\mu_\mathbf{X}^T\mathbf{A}^T+\mathbf{bb}^T.\end{aligned} \tag{3.35}$$

3.3 Random signals

A simple representation of a random signal can be described through an example. Suppose there is a transmitter that transmits a signal $f(t)$. This signal is attenuated over the transmission path, leading to an attenuated signal at the receiver $s(t) = \alpha \cdot f(t)$, where α indicates the degree of attenuation and which is bounded by $0 \leq |\alpha| \leq 1$. Now at the output of the receiver, this received signal $s(t)$ will also be corrupted by noise from the electrical components in the receiver and from other unwanted signals that are received. Let us capture all this noise in a single term called $n(t)$. The signal at the output of the receiver $x(t)$ can now be written as

$$x(t) = s(t) + n(t). \tag{3.36}$$

The noise $n(t)$ can be regarded as a random signal, whose statistics can be known but whose realizations are unpredictable. The signal $s(t)$ is fully deterministic, meaning that we can calculate and predict each sample of this signal if we know the transmitted signal $f(t)$. If the deterministic signal $s(t)$ and the noise $n(t)$ are added to form the received signal $x(t)$, we receive a signal which is partially predictable (the deterministic portion) and which is partially unpredictable (the random noise). Therefore, the signal $x(t)$ is called a random signal. Figure 3.3 shows an example of a deterministic signal to which noise is added to create a random signal.

3.3.1 Additive white Gaussian noise

The random noise $n(t)$ that is part of the random signal $x(t)$ can be described by a PDF. For simplicity, it is often assumed that this noise is additive white Gaussian noise or AWGN noise. All adjectives correspond to certain characteristics of the noise. The noise is additive, meaning that it is added to the deterministic part of the signal. Noise can be regarded as white if it has a uniform power distribution over the relevant frequency spectrum. As an analogy, white light emits all frequencies uniformly across the visible spectrum. Furthermore, it is assumed that the noise is

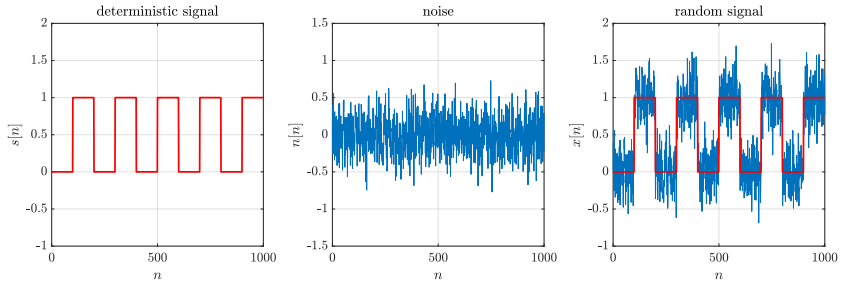


Figure 3.3: An example of a deterministic signal to which random noise is added to create a random signal.

Gaussian distributed, a fair assumption if we consider the central limit theorem. Oftentimes this Gaussian noise distribution has a zero mean and even if this is not the case and the noise actually has a DC-term, then this DC-term is often regarded as part of the deterministic signal. We may therefore write that $n(t) \sim \mathcal{N}(n(t) \mid 0, \sigma_n^2)$, or similarly that $x(t) \sim \mathcal{N}(x(t) \mid s(t), \sigma_n^2)$, meaning that the probability distribution is centered around the deterministic signal.

3.3.2 Discrete-time stochastic processes

The previous description of a random signal is often used in practice. However, it is rather simplistic: the uncertainty in the signal may not be additive or the uncertainty might be inherent in the signal itself. For a more general description, we need to formally introduce the random process. Although here we will focus on discrete-time random processes, the extension to continuous-time random processes is straightforward.

Suppose we would like to measure a random or stochastic signal. Using multiple identical receivers we will sample the received signal at the exact same time instances. We can denote each discrete-time signal sample as $x[n, \xi_k]$, where n corresponds to the discrete-time index or sample number and where ξ_k corresponds to the receiving element k. The uncertainty that is present in the signals results in different sample values for each of the receiving elements. This is described schematically in Figure 3.4.

Formally, a random process can be defined as a mapping between all possible observations in the sample space to the time functions $x[n, \xi_k]$. All possible sequences $x[n, \xi_k]$, which defines the entire stochastic process, are called *ensemble*. If we have a look at the sampled signal that is produced by a single receiver (i.e. we fix ξ), we have a single time function, which is one realization of the random process. Since each sample is affected by uncertainty, if we measure the signal again or with another receiver, we will measure different samples, because of the random noise that is present.

When we fix the time instance (i.e. we fix n) and we look at all the samples received from all receivers at the same time-point, then we have a random variable. In fact, all these receivers should measure the exact same deterministic signal, but a different value of the noise. Each of the measured samples is therefore distributed around the value of the deterministic signal according to the probability distribution of the noise. Finally, if we fix both time n and realization ξ we obtain a single number.

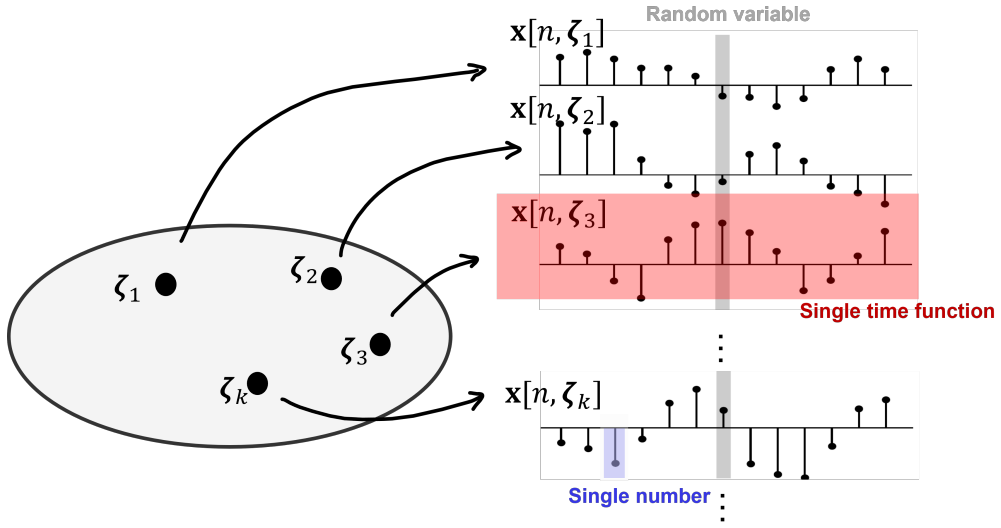


Figure 3.4: Schematic representation of a stochastic process which maps each outcome of the sample space S into a time function $x[n, \xi_k]$.

3.3.3 Statistical characterization of random signals

To characterize the statistical properties of random signals, it is convenient to describe them as random vectors, where each time instance can be considered as a random variable. Let us denote the random variable corresponding to a specific time instance n as $X[n]$. Using this notation the random signal which contains the N samples at and before time instance n can be captured in a random vector $\mathbf{X}[n]$ as

$$\mathbf{X}[n] = [X[n], X[n-1], \ldots, X[n-N+1]]^T. \tag{3.37}$$

The corresponding realization can be written as

$$\mathbf{x}[n] = [x[n], x[n-1], \ldots, x[n-N+1]]^T. \tag{3.38}$$

Please note the different notation with respect to random vectors. Random signals are accompanied by a time index (e.g. $x[n]$). The random processes that generate all these random signals are referred to by a single capital letter $X[n, \xi]$, where the bold ξ indicates a vector of possible realizations. However, ξ is often dropped and the random process is simply indicated as $X[n]$.

The random signals can be written as random vectors and therefore it is also possible to determine relationships between the individual random variables through the covariance and the correlation. We distinguish between autocorrelation and autocovariance, which describe the degree of dependence of the samples in the same signal, and cross-correlation and cross-covariance, which describe the degree of dependence between the samples of two different signals.

Given the random signals $x[n]$ and $y[n]$ corresponding to individual samples of the random processes $X[n]$ and $Y[n]$, several statistical properties can be determined. An overview of all statistical properties is given in Table 3.1. The random process is indicated by the subscript and the sample(s) at which the property is determined is given in between the square brackets $[\cdot]$. Note that for real signals the conjugate (*) can be omitted.

Statistical property	Definition
Mean	$\mu_X[n] = \mathrm{E}\left[X[n]\right]$
Variance	$\sigma_X^2[n] = \mathrm{E}\left[\left\|X[n] - \mu_X[n]\right\|^2\right]$
autocovariance	$C_X[n_1, n_2] = \mathrm{E}\left[\left(X[n_1] - \mu_X[n_1]\right)\left(X[n_2] - \mu_X[n_2]\right)^*\right]$
autocorrelation	$r_X[n_1, n_2] = \mathrm{E}\left[X[n_1] \cdot X^*[n_2]\right]$
Cross-covariance	$C_{XY}[n_1, n_2] = \mathrm{E}\left[\left(X[n_1] - \mu_X[n_1]\right)\left(Y[n_2] - \mu_Y[n_2]\right)^*\right]$
Cross-correlation	$r_{XY}[n_1, n_2] = \mathrm{E}\left[X[n_1] \cdot Y^*[n_2]\right]$
Correlation coefficient	$\rho_{XY}[n_1, n_2] = \frac{C_{XY}[n_1, n_2]}{\sigma_X[n_1]\sigma_Y[n_2]}$

Table 3.1: Statistical characterization of random signals.

3.4 Stationarity and ergodicity

Stationarity and ergodicity are important properties of random processes, which facilitate the characterization of random processes.

3.4.1 Stationarity

A random process is called *stationary* if its statistical properties do not change over time. This means that the signal statistics of random variable $X[n]$ should be identical to the signal statistics of random variable $X[n+k]$ for any value of k. For the underlying probability distribution then holds that

$$p_{X[n]}(x) = p_{X[n+k]}(x) = p_X(x), \tag{3.39}$$

where it is possible to drop the indexing if there is no effect on the distribution.

Stationarity can be defined at different orders. A signal is called an N^{th}-order stationary signal if the N^{th}-order signal statistics do not change over time. The order of the signal statistics can be defined by accounting for the number of random variables, or time points, that are needed to calculate that statistic. For example, the mean and variance can be regarded as first-order signal statistics, since are characteristics of single random variables, while the covariance and correlation are second-order statistics since they describe the relationship between pairs of random variables.

When stationarity applies to the statistics of any order, the random process or signal is defined as **strict-sense stationary**. This also implies that the signal is a N^{th}-order stationary signal for $N = 1, 2, \ldots$.

Wide-sense stationarity (WSS) is a weaker criterion than strict-sense stationarity. Instead of requiring that all N-th order statistics be constant, only the first- as second-order statistics must be independent of the time shift.

Because of the first-order stationarity, we may simplify the notation of the first-order statistics for WSS processes as

$$\mathrm{E}\left[X[n]\right] = \mu_X[n] = \mu_X \tag{3.40}$$

and similarly

$$\mathrm{Var}\left[X[n]\right] = \sigma_X^2[n] = \sigma_X^2, \tag{3.41}$$

which imply that mean and variance are constant.

It is also important to consider the consequences of (wide-sense) stationarity for the covariance and correlation functions. From the definitions of covariance and correlation, these properties depend on the two time indices n_1 and n_2. Taking for example the covariance, stationarity implies that $C_X[n_1, n_2] = C_X[n_1 + k, n_2 + k]$ holds. This leads to the observation that the value of the covariance and correlation only depends on the difference between the two time indices. This time difference is called lag and is defined as

$$l = n_1 - n_2. \tag{3.42}$$

Therefore, for WSS signals the notation of the correlation (and the covariance) may be simplified as

$$r_X[n_1, n_2] = r_X[n_1 - n_2] = r_X[l] = \mathrm{E}\left[X[n_1]X^*[n_2]\right] = \mathrm{E}\left[X[n]X^*[n-l]\right]. \tag{3.43}$$

Using this definition, the value of the autocorrelation function can be determined for $l = 0$ as

$$r_X[0] = \sigma_X^2 + |\mu_X|^2 \geq 0 \tag{3.44}$$

and it can be shown that this correlation value is a maximum of the autocorrelation function since $r_X[0] \geq r_X[l]\ \forall\ l$. Furthermore, it can be found that the autocorrelation function is a complex conjugate symmetric function of its lag

$$r_X^*[-l] = r_X[l]. \tag{3.45}$$

When dealing with real-valued signals, this property can be simplified to $r_X[l] = r_X[-l]$.

Example 3.2

$X(t)$ and $Y(t)$ are independent wide sense stationary processes with expected values μ_X and μ_Y and autocorrelation functions $R_x(\tau)$ and $R_y(\tau)$, respectively. Let $W(t) = X(t)Y(t)$.

1. Find $E[W(t)]$ and $R_W(t, \tau)$ and show that $W(t)$ is wide sense stationary.
2. Are $W(t)$ and $X(t)$ jointly wide sense stationary?

Solution.

1. Since $X(t)$ and $Y(t)$ are independent processes,

$$E[W(t)] = E[X(t)Y(t)] = E[X(t)]E[Y(t)] = \mu_X\mu_Y.$$

In addition,

$$R_W(t, \tau) = R_X(\tau)R_Y(\tau).$$

We can conclude that $W(t)$ is wide sense stationary;

2. To examine whether $W(t)$ and $X(t)$ are jointly wide sense stationary, we calculate

$$R_{WX}(t, \tau) = E[W(t)X(t + \tau)] = E[X(t)Y(t)X(t + \tau)].$$

By independence of $X(t)$ and $Y(t)$,

$$R_{WX}(t, \tau) = E[X(t)X(t + \tau)]E[Y(t)] = \mu_Y R_X(\tau).$$

Since $W(t)$ and $X(t)$ are both wide sense stationary and since $R_{WX}(t,\tau)$ depends only on the time difference τ, we can conclude that $W(t)$ and $X(t)$ are jointly wide sense stationary.

3.4.2 Power spectral density

Besides the covariance and correlation functions, another important second-order statistic is the power spectral density (PSD). Although the power spectral density will be treated more in detail in Part. III, here we introduce some basic concepts.

Let us first introduce the concepts of the energy and the power of a signal. The energy E_s of a signal is defined as the sum of all squared sample magnitudes as

$$E_s = \sum_{n=-\infty}^{\infty} |x[n]|^2. \tag{3.46}$$

The average signal power P_s is the average signal energy per sample and is similarly defined as

$$P_s = \lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} |x[n]|^2, \tag{3.47}$$

where a total of N samples are averaged over.

If an infinitely long signal has finite signal energy it is called an *energy signal*. Since the energy is finite and the time duration infinite, the average signal power is zero. An example of an energy signal is a short pulse that is transmitted only once. An infinitely long signal that has a finite average signal power is called a *power signal*. Because of the finite power that the signal carries over an infinitely long time, the total signal energy is infinite. Any non-zero bounded signal that is infinitely long can be regarded as a power signal. In general, random signals are regarded as aperiodic power signals.

The average signal power yields a limited amount of information because it is just a single number. Oftentimes it is desirable to know how the signal power is distributed over frequencies, and which frequencies are contributing the most to this signal power. This would for example allow for the detection of unwanted signals that cause interference. It is possible to calculate the power spectral density (PSD) of a signal from its Fourier transform of discrete-time signals, which represents the distribution of the signal power of the frequency spectrum. A short review of the Fourier transform is provided in Appendix D.

The power spectral density $P_X(e^{j\theta})$ is defined as the expected value of the squared Fourier transform of $x[n]$, normalized by the number of signal samples as

$$P_X(e^{j\theta}) = \lim_{N\to\infty} \frac{1}{2N+1} \mathrm{E}\left[\left| \sum_{n=-N}^{N} x[n] e^{-jn\theta} \right|^2 \right]. \tag{3.48}$$

If the signal $x[n]$ is real-valued, then the power spectral density is symmetric around its DC component (i.e. $P_X(e^{j\theta}) = P_X(e^{-j\theta})$). Furthermore, the power spectral density is always non-negative and periodic with a period of 2π similarly to $X(e^{j\theta})$. Averaging the area under the power spectral density function will return the average signal power as

$$P_s = \frac{1}{2\pi} \int_{-\pi}^{\pi} P_X(e^{j\theta}) d\theta. \tag{3.49}$$

A special property of the power spectral density is given by the Wiener-Kintchine theorem, which states that the power spectral density of a WSS signal is the Fourier transform of its autocorrelation function as

$$P_X(e^{j\theta}) = \sum_{l=-\infty}^{\infty} r_X[l]e^{-jl\theta}. \tag{3.50}$$

Similarly, the opposite holds

$$r_X[l] = \frac{1}{2\pi}\int_{-\pi}^{\pi} P_X(e^{j\theta})e^{jl\theta}d\theta. \tag{3.51}$$

Some important properties of the PSD are:

- The PSD is a real-valued periodic function of frequency 2π;
- Because of the conjugate symmetric property of the autocorrelation function, it holds that $P_X(e^{j\theta}) = P_X^*(e^{j\theta})$. As a consequence, if the random signal $x[n]$ is real-valued, then $P_X(e^{j\theta})$ is even, that is $P_X(e^{j\theta}) = P_X(e^{-j\theta})$;
- The area under the PSD is non-negative and it equals the average power of $x[n]$:

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} P_X(e^{j\theta})d\theta = r_x[0] = E\{|x[n]|^2\} \geq 0. \tag{3.52}$$

Example 3.3

Suppose we would like to calculate the power spectral density of a zero-mean wide-sense stationary random process $x[n]$, whose autocorrelation function is given as $r_X[l] = \alpha^{|l|}$ with $-1 < \alpha < 1$.

Using the above equation the power spectral density can be determined as

$$\begin{aligned}
P_X(e^{j\theta}) &= \sum_{l=-\infty}^{\infty} r_X[l]e^{-jl\theta} \\
&= \sum_{l=-\infty}^{0} r_X[l]e^{-jl\theta} + \sum_{l=0}^{\infty} r_X[l]e^{-jl\theta} - r_X[0]e^{-j0\theta} \\
&= \sum_{p=0}^{\infty} r_X[-p]e^{jp\theta} + \sum_{l=0}^{\infty} r_X[l]e^{-jl\theta} - 1 \\
&= \sum_{p=0}^{\infty} \left(\alpha e^{j\theta}\right)^p + \sum_{l=0}^{\infty} \left(\alpha e^{-j\theta}\right)^l - 1 \\
&= \frac{1}{1-\alpha e^{j\theta}} + \frac{1}{1-\alpha e^{-j\theta}} - 1 \\
&= \frac{1-\alpha e^{-j\theta} + 1 - \alpha e^{j\theta} - \left(1-\alpha e^{j\theta}\right)\left(1-\alpha e^{-j\theta}\right)}{\left(1-\alpha e^{j\theta}\right)\left(1-\alpha e^{-j\theta}\right)} \\
&= \frac{1-\alpha^2}{1+\alpha^2-2\alpha\cos(\theta)}.
\end{aligned} \tag{3.53}$$

Special case: zero-mean white noise

Given the implications of the central limit theorem, noise is often assumed to be zero-mean additive Gaussian white noise; for simplicity, it is often simply called zero-mean white noise. Thus, it is useful to know the autocorrelation function of such a process, which is determined as

$$r_X[l] = \mathrm{E}\left[X[n]X^*[n-l]\right] = \sigma_X^2\delta[l] = \begin{cases}\sigma_X^2, & \text{for} l = 0,\\ 0, & \text{otherwise.}\end{cases} \tag{3.54}$$

Let us take a close look at Eq. (3.54). At lag 0, the autocorrelation function reduces to the variance plus the squared mean (see Eq. (3.44)). Since the mean is 0, only the variance remains. Intuitively one could also regard the zero lag as multiplying each signal sample with itself, leading always to positive contributions since the multiplication of equal signs always returns a positive number.

Since the signal samples are uncorrelated, the assumption is made that the noise is symmetrically distributed with zero-mean and is completely random. This means that each sample only depends on itself and has no dependence on the samples before or after it. If we were to introduce a certain lag l, each sample would be multiplied by another completely random sample at a distance l. This sample has a random sign and magnitude. Because of the zero-mean property we are half as likely to get a random number with a positive sign as one with a negative sign. The result of the multiplication, therefore, is equally likely to result in a positive or a negative contribution. Because of the random magnitudes that are symmetrically distributed around zero, generally, we may intuitively understand that the total positive and negative contributions for the autocorrelation will cancel each other out, leading to a zero autocorrelation.

In summary, the autocorrelation of zero-mean white noise is a delta pulse at lag 0, with amplitude given by the variance of $x[n]$.

The fact that the signal is white means that, in the frequency domain, there is the same contribution of the signal power at all frequencies. The adjective white comes from an analogy with white light, which contains all colors (all light wavelengths) in equal amounts. It follows that the power spectral density is the same over all frequencies, with a constant value given by the variance of $x[n]$:

$$r_n[l] = \sigma_n^2\delta[l] \Longleftrightarrow P_X(e^{j\theta}) = \sigma_n^2 \quad \forall\theta \tag{3.55}$$

The theoretical autocorrelation function and power spectral density of a white noise sequence are presented in Figure 3.5.

Cross-power spectral density

The power spectral density as discussed up until now was actually the "auto" power spectral density since it was calculated from the autocorrelation function. Similarly, the cross-power spectral density of two random processes can be defined, showing the relationship between two random processes in the frequency domain as

$$P_{XY}(e^{j\theta}) = \sum_{l=-\infty}^{\infty} r_{XY}[l]e^{-jl\theta} \Longleftrightarrow r_{XY}[l] = \frac{1}{2\pi}\int_{-\pi}^{\pi} P_{XY}(e^{j\theta})e^{jl\theta}d\theta. \tag{3.56}$$

Because of the complex conjugate symmetry property of the correlation function, also for the cross power spectral density, it holds that

$$r_{XY}[l] = r_{YX}^*[-l] \Longleftrightarrow P_{XY}(e^{j\theta}) = P_{YX}^*(e^{j\theta}). \tag{3.57}$$

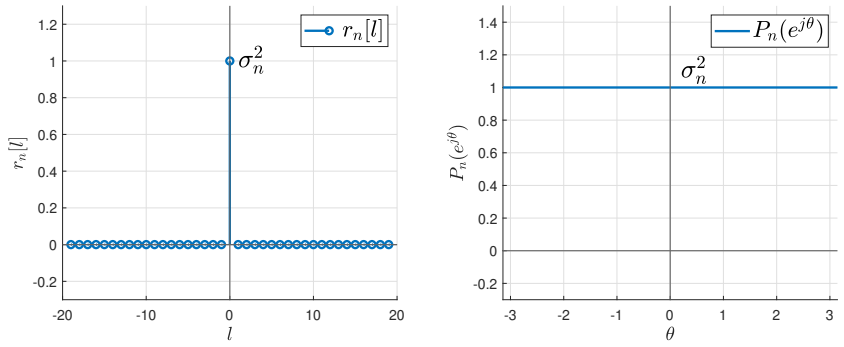


Figure 3.5: Theoretical autocorrelation function (left) and power spectral density (right) of a Gaussian white noise sequence with variance σ_n^2.

Coherence function

Since the cross-correlation function depends on the individual signal amplitudes, also the cross-power spectral density function will. In order to generalize the cross-power spectral density function for all types of signals, the coherence function is introduced as

$$C_{XY}(e^{j\theta}) = \frac{P_{XY}(e^{j\theta})}{\sqrt{P_X(e^{j\theta})}\sqrt{P_Y(e^{j\theta})}} \tag{3.58}$$

which normalizes the cross-power spectral density function. The absolute value of this function is bounded between 0 and 1, which reflects respectively no correlation and total correlation to the point where $x[n] = y[n]$.

Example 3.4

Let $w[n]$ be a zero-mean, uncorrelated Gaussian random sequence with variance $\sigma^2[n] = 1$.

1. Characterize the random sequence $w[n]$;
2. Define $x[n] = w[n] + w[n-1]$, $-\infty < n < \infty$. Determine the mean and autocorrelation of $x[n]$. Also characterize $x[n]$.

Solution.

1. Since the lack of correlation implies independence for Gaussian random variables, $w[n]$ is an independent random sequence. Since its mean and variance are constants, it is at least stationary in the first order. Furthermore, we have

$$r_x[n_1, n_2] = \sigma^2\delta[n_1 - n_2] = \delta[n_1 - n_2].$$

Hence $w[n]$ is also a WSS random process.

2. The mean of $x[n]$ is zero for all n since $w[n]$ is a zero-mean process. Consider

$$\begin{aligned} r_x[n_1, n_2] &= \mathrm{E}\{x[n_1]x[n_2]\} \\ &= \mathrm{E}\{(w[n_1] + w[n_1 - 1])(w[n_2] + w[n_2 - 1])\} \\ &= r_w[n_1, n_2] + r_w[n_1, n_2 - 1] + r_w[n_1 - 1, n_2] + r_w[n_1 - 1, n_2 - 1] \\ &= \sigma^2\delta[n_1 - n_2] + \sigma^2\delta[n_1 - n_2 + 1] + \sigma^2\delta[n_1 - n_2 - 1] \\ &\quad + \sigma^2\delta[n_1 - 1 - n_2 + 1] \\ &= 2\delta[n_1 - n_2] + \delta[n_1 - n_2 + 1] + \delta[n_1 - n_2 - 1]. \end{aligned}$$

Clearly, $r_w[n_1, n_2]$ is a function of $n1 - n2$. Hence

$$r_w[l] = 2\delta[l] + \delta[l + 1] + \delta[l - 1].$$

Therefore, $x[n]$ is a WSS sequence. However, it is not an independent random sequence since both $x[n]$ and $x[n + 1]$ depend on $w[n]$.

3.4.3 Ergodicity

When a signal is stationary, the exact calculation of the expected value operators is usually still cumbersome, because it requires knowledge of the entire random process, which means we need to have the whole ensemble of time functions available (all realizations). In practice, we usually only have a limited number of samples of a single realization of the random process. This means that not only do we have one single realization, but also that this realization is not infinitely-long, but limited in a time window. **If a random process is *ergodic*, the statistical properties of the entire random process can be inferred from any of its realizations**. In order for a signal to be ergodic, it has to be stationary. Ergodicity is usually a strong assumption that cannot always be confirmed. However, without this assumption, the signal statistics could not be approximated in most practical cases.

The formal definition of ergodicity states that a strict-sense stationary process $X[n]$ is strict-sense ergodic if the time average equals the ensemble average. This means that any of the statistical properties of $X[n]$ of any order can be obtained by any of its single realizations $x[n]$, known during an infinite time interval.

As for stationarity, we can define ergodicity at different orders, and for different statistics. A process is (continous-time) ergodic in the mean if

$$\lim_{T\to+\infty} \frac{1}{2T} \int_{-T}^{T} x(t)dt = \mu_x = E[X(t)] = \int_{-\infty}^{+\infty} x p_X(x;t)dx, \tag{3.59}$$

where the equality above only holds if the variance of the time-average tends to zero for $T \to \infty$.

Similarly, a process can be ergodic in the autocorrelation if the autocorrelation can be calculated by time-averaging over one single realization of the process. Assuming a zero-mean process, this can be written:

$$\lim_{T\to+\infty} \frac{1}{2T} \int_{-T}^{T} x(t)x(t+\tau) = R_X(\tau) = E[X(t+\tau)X(t)]. \tag{3.60}$$

Also for the equation above the equality only holds if the variance of the time-average tends to zero for $T \to \infty$.

Note that the above definitions can be easily extended to discrete-time processes by substituting the integral with a sum over the discrete time samples.

3.4.4 Approximate statistics

For ergodic random processes, the signal statistics that are defined in the table of the previous section can be approximated by time-averaging. However, as explained above, we often only have a limited set of samples available, rather than an infinitely-long signal. In practice, this means that the expected value operator is replaced by time-averaging over a sequence of length N. For discrete-time signals, this results in

$$\mathrm{E}[\cdot] \quad \rightarrow \quad \frac{1}{N}\sum_{n=0}^{N-1}[\cdot]. \tag{3.61}$$

Please note that the averaging takes place over the individual realizations $x[n]$ and not over the random process $X[n]$. When applying time-averaging, care should be taken for the calculation of the covariance and correlation functions. These functions need to be calculated at different lags l. When the lag l approaches the length of the sequences N, there is a very limited number of samples to average over, reducing to 1 only for $l = N - 1$. Therefore, in the calculation of the covariance and correlation functions the lag l is limited by a certain upper-lag L, which should be significantly smaller than N.

The approximated signal statistics are denoted by a $\hat{}$ identifier (hat). These approximate signal statistics are defined in Table 3.2. It should be noted that the definitions of $\hat{r}$ and $\hat{C}$ are biased, meaning that they are calculated by normalizing for N values, whereas not always are N values used in the summation, because of the effect of shifting the signal to calculate the different lags. The unbiased estimate would be obtained by normalizing by $N - |l|$ instead of N. However, as we shall see later, the unbiased estimate becomes problematic when estimating the power spectrum from the autocorrelation function (see Chapter 13); this is the reason why we normally use the biased estimate.

By approximating the definitions of the cross-covariance and cross-correlation matrices and by generalizing the definitions for the autocovariance and -correlation matrices, the following approximated covariance and correlation matrices can be determined for the random signal sequences $\mathbf{X}$ and $\mathbf{Y}$ of length N. The approximated cross-covariance matrix can be defined using the definitions in Table 3.2 as

$$\hat{\mathbf{C}}_{XY} = \begin{bmatrix} \hat{C}_{XY}[0] & \hat{C}_{XY}[1] & \hat{C}_{XY}[2] & \cdots & \hat{C}_{XY}[N-1] \\ \hat{C}_{XY}[-1] & \hat{C}_{XY}[0] & \hat{C}_{XY}[1] & \cdots & \hat{C}_{XY}[N-2] \\ \hat{C}_{XY}[-2] & \hat{C}_{XY}[-1] & \hat{C}_{XY}[0] & \cdots & \hat{C}_{XY}[N-3] \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \hat{C}_{XY}[1-N] & \hat{C}_{XY}[2-N] & \hat{C}_{XY}[3-N] & \cdots & \hat{C}_{XY}[0] \end{bmatrix} \tag{3.62}$$

Statistical property	Definition	Constraints
Mean	$\hat{\mu}_X = \frac{1}{N}\sum_{n=0}^{N-1} x[n]$	
Variance	$\hat{\sigma}_X^2 = \frac{1}{N}\sum_{n=0}^{N-1} \lvert x[n] - \hat{\mu}_X\rvert^2$ $= \frac{1}{N}\sum_{n=0}^{N-1} \lvert x[n]\rvert^2 - \lvert\hat{\mu}_X\rvert^2$	
autocovariance	$\hat{C}_X[l] = \frac{1}{N}\sum_{n=0}^{N-1-\lvert l\rvert}(x[n] - \hat{\mu}_X)(x[n-l] - \hat{\mu}_X)^*$	$\lvert l\rvert \leq L-1$
autocorrelation	$\hat{r}_X[l] = \frac{1}{N}\sum_{n=0}^{N-1-\lvert l\rvert} x[n]x^*[n-l] = \hat{C}_X[l] + \lvert\hat{\mu}_X\rvert^2$	$\lvert l\rvert \leq L-1$
Cross-covariance	$\hat{C}_{XY}[l] = \frac{1}{N}\sum_{n=0}^{N-1-\lvert l\rvert}(x[n] - \hat{\mu}_X)(y[n-l] - \hat{\mu}_Y)^*$	$\lvert l\rvert \leq L-1$
Cross-correlation	$\hat{r}_{XY}[l] = \frac{1}{N}\sum_{n=0}^{N-1-\lvert l\rvert} x[n]y^*[n-l]$ $= \hat{C}_{XY}[l] + \hat{\mu}_X \cdot \hat{\mu}_Y^*$	$0 \leq l \leq L-1$
	$\hat{r}_{XY}[l] = \frac{1}{N}\sum_{n=\lvert l\rvert}^{N-1} x[n]y^*[n-l]$ $= \hat{C}_{XY}[l] + \hat{\mu}_X \cdot \hat{\mu}_Y^*$	$-(L-1) \leq l \leq 0$
Correlation coefficient	$\hat{\rho}_{XY}[l] = \frac{\hat{C}_{XY}[l]}{\hat{\sigma}_X \cdot \hat{\sigma}_Y}$	

Table 3.2: Approximate signal statistics calculated on finite-length discrete-time random signals

and the approximated cross-correlation matrix by

$$\hat{\mathbf{R}}_{XY} = \begin{bmatrix} \hat{r}_{XY}[0] & \hat{r}_{XY}[1] & \hat{r}_{XY}[2] & \cdots & \hat{r}_{XY}[N-1] \\ \hat{r}_{XY}[-1] & \hat{r}_{XY}[0] & \hat{r}_{XY}[1] & \cdots & \hat{r}_{XY}[N-2] \\ \hat{r}_{XY}[-2] & \hat{r}_{XY}[-1] & \hat{r}_{XY}[0] & \cdots & \hat{r}_{XY}[N-3] \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \hat{r}_{XY}[1-N] & \hat{r}_{XY}[2-N] & \hat{r}_{XY}[3-N] & \cdots & \hat{r}_{XY}[0] \end{bmatrix}. \tag{3.63}$$

The approximated autocovariance matrix can be defined by

$$\hat{\mathbf{C}}_X = \begin{bmatrix} \hat{C}_X[0] & \hat{C}_X[1] & \hat{C}_X[2] & \cdots & \hat{C}_X[N-1] \\ \hat{C}_X[-1] & \hat{C}_X[0] & \hat{C}_X[1] & \cdots & \hat{C}_X[N-2] \\ \hat{C}_X[-2] & \hat{C}_X[-1] & \hat{C}_X[0] & \cdots & \hat{C}_X[N-3] \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \hat{C}_X[1-N] & \hat{C}_X[2-N] & \hat{C}_X[3-N] & \cdots & \hat{C}_X[0] \end{bmatrix} \tag{3.64}$$

and the approximated autocorrelation matrix by

$$\hat{\mathbf{R}}_X = \begin{bmatrix} \hat{r}_X[0] & \hat{r}_X[1] & \hat{r}_X[2] & \cdots & \hat{r}_X[N-1] \\ \hat{r}_X[-1] & \hat{r}_X[0] & \hat{r}_X[1] & \cdots & \hat{r}_X[N-2] \\ \hat{r}_X[-2] & \hat{r}_X[-1] & \hat{r}_X[0] & \cdots & \hat{r}_X[N-3] \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \hat{r}_X[1-N] & \hat{r}_X[2-N] & \hat{r}_X[3-N] & \cdots & \hat{r}_X[0] \end{bmatrix}. \tag{3.65}$$

Especially the approximated autocorrelation matrix is in practice often used. When dealing with real-valued signals, this leads to a symmetric autocorrelation matrix (i.e. $\hat{\mathbf{R}}_X = \hat{\mathbf{R}}_X^T$) and it turns out that this matrix is at the same time positive semi-definite (i.e. $\hat{\mathbf{R}} \succeq 0$). More specifically, it has a so-called *Toeplitz* structure, which refers to a matrix in which the diagonals are constant.

3.4.5 Ergodicity and approximate statistics

In practice, we cannot observe the signal for an infinite interval. For ergodic random processes, we need to replace the time-averaging over the infinitely-long single realization of the random processes with averaging over a limited number of samples. For a discrete-time signal of length N, the conditions for ergodicity of the mean stated in (3.59) then become

$$\mathrm{E}\left[\frac{1}{N}\sum_{n=0}^{N-1} x[n]\right] = \mathrm{E}[\hat{\mu}_X] = \mu_X, \tag{3.66}$$

and

$$\mathrm{Var}\left\{\frac{1}{N}\sum_{n=0}^{N-1} x[n]\right\} \xrightarrow{N\to\infty} 0. \tag{3.67}$$

Differently from (3.59) and (3.60), here there is no limit to infinity since we are dealing with a limited set of samples.

Note that the second condition does not look at the variance of $X[n]$, but rather at the variance of our estimator for the mean $\hat{\mu}_X$, as given in Table 3.2. By verifying the second condition, we are verifying the consistency of our mean estimator. This concept will be further explained in Part. II.

4

Rational signal models

4.1 Introduction

In many applications, it is useful to generate random signals with desired properties or to obtain a representation of a random signal which captures a set of the signal characteristics. By the term "model" we indicate a mathematical description that provides a representation of certain properties of the signal. If we are able to construct a useful model of a random signal, then we can use this model for various applications. For example, to obtain a better understanding of the physical mechanism generating the signal, to detect changes in the signal for diagnostics purposes, to synthesize artificial signals similar to the real ones, to extract parameters for pattern recognition and machine learning, to obtain a more efficient representation of the signal for data compression.

We often assume a model to be parametric, i.e., a function completely defined by a finite number of model parameters. Here, we further restrict our attention to a particular class of models known as rational signal models, whose system function and power spectrum can be expressed as a ratio between polynomials in z (or equivalently in $e^{j\omega}$). We will particularly focus on stochastic modeling, by which we describe a signal as the hypothetical output of some linear-time invariant system, the input of which is white noise.

Before reading this section, it is important to have a basic understanding of linear-time invariant system. These are reviewed in Appendix E.

4.2 Spectral factorization

A stochastic process may be represented by a stochastic model with given order and parameters, which is able to generate a random signal characterized by well-defined spectral properties. Signal modeling is closely related to spectral factorization which states that most random processes with a continuous power spectral density (PSD) can be generated as the output of a causal filter driven by white noise, the so-called *innovation representation* of the random process. In this section, we first explain how the statistical properties of a random signal are affected when filtering by a LTI system; then we discuss random signals with rational spectra and the innovation representation of such signals; finally, we explain how spectral factorization can be used to provide the innovation representation of a random signal.

4.2.1 LTI with random inputs

When a random signal is filtered by an LTI system, its statistical properties are changed. Focusing on the second-order statistics, here we show the transformation that a discrete random power signal undergoes when processed by a discrete-time LTI system.

Time-domain analysis

The input-output relationship of an LTI system is described using the following convolution

$$y[n] = \sum_{k=-\infty}^{\infty} h[k]x[n-k]. \tag{4.1}$$

Here $x[n]$ is a random stationary signal. Output $y[n]$ converges and is stationary if the system represented by the impulse response $h[k]$ is stable. The condition for stability is that $\sum_{-\infty}^{\infty} |h[k]| < \infty$, or equivalently that all poles are within the unit circle.

The exact output $y[n]$ cannot be calculated, since $y[n]$ is a stochastic process. However, certain statistical properties of $y[n]$ can be determined from knowledge of the statistical properties of the input and the characteristics of the system.

The *mean* value of output $y[n]$ can be determined by taking the expected value from both sides of (4.1) as

$$\mu_y = \sum_{k=-\infty}^{\infty} h[k]\,\mathrm{E}\{x[n-k]\} = \mu_x \sum_{k=-\infty}^{\infty} h[k] = \mu_x H(e^{j0}). \tag{4.2}$$

Note that the filter coefficients are deterministic, and thus $\mathrm{E}\{h[k]\} = h[k]$.

The *input-output correlation* can also be calculated without knowing the exact output. The input-output correlation is defined as $r_{xy}[l] = \mathrm{E}\{x[n+l]y^*[n]\}$. Therefore, to calculate this correlation, we take the complex conjugate of (4.1) and multiply it with $x[n+l]$ before taking the expected value as

$$\begin{aligned}
r_{xy}[l] &= \mathrm{E}\{x[n+l]y^*[n]\} = \sum_{k=-\infty}^{\infty} h^*[k]\,\mathrm{E}\{x[n+l]x^*[n-k]\} \\
r_{xy}[l] &= \sum_{k=-\infty}^{\infty} h^*[k]r_{xx}[l+k] = \sum_{m=-\infty}^{\infty} h^*[-m]r_{xx}[l-m] \\
r_{xy}[l] &= h^*[-l] * r_{xx}[l] \\
r_{yx}[l] &= h^*[l] * r_{xx}[l].
\end{aligned} \tag{4.3}$$

The *output correlation* can be calculated as

$$\begin{aligned}
r_{yy}[l] &= \mathrm{E}\{y[n]y^*[n-l]\} = \sum_{k=-\infty}^{\infty} h[k]\,\mathrm{E}\{x[n-k]y^*[n-l]\} \\
r_{yy}[l] &= \sum_{k=-\infty}^{\infty} h[k]r_{xy}[l-k] = h[l] * r_{xy}[l].
\end{aligned} \tag{4.4}$$

By combining this result with $r_{xy}[l] = h^*[-l] * r_{xx}[l]$ we obtain

$$\begin{aligned}
r_y[l] &= h[l] * h^*[-l] * r_x[l] \\
r_y[l] &= r_h[l] * r_x[l],
\end{aligned} \tag{4.5}$$

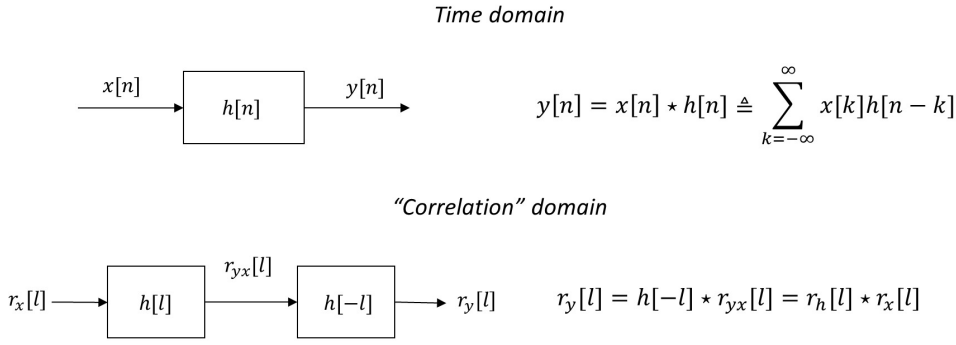


Figure 4.1: LTI with impulse response $h[n]$, random input $x[n]$ and random output $y[n]$. Input-output relationships in the time and "correlation" domains.

where $r_h[l]$ is the autocorrelation of the LTI system with impulse response $h[n]$.

The *output signal power* is equal to $P_y = \mathrm{E}\,|y[n]|^2 = r_y[0]$, which can be calculated using the previous equations,

$$\begin{aligned} P_y &= \mathrm{E}\,|y[n]|^2 = r_y[0] \\ &= r_h[l] * r_x[l=0] = \sum_{k=-\infty}^{\infty} h[k]h^*[-k]r_x[k]. \end{aligned} \tag{4.6}$$

Figure 4.1 describes schematically the input-output relationship of an LTI system in the time domain and in the "correlation" domain.

Transform-domain analysis

The relationships found above in the time domain can be easily transformed in the frequency- and z-domains. Since we are dealing with autocorrelation functions and power spectral densities of mostly real signals, it is useful to recall that

- if $x[n]$ is real, then
$$x[n] = x^*[n] \Longleftrightarrow X(z) = X^*(z^*); \tag{4.7}$$
- if $x[n]$ has even symmetry around the time origin, then
$$x[n] = x[-n] \Longleftrightarrow X(z) = X(1/z); \tag{4.8}$$
- if $x[n]$ is both real and even, then
$$x[n] = x^*[-n] \Longleftrightarrow X(z) = X^*(z^*) = X(1/z) = X^*(1/z^*). \tag{4.9}$$

Using the above properties, we can easily calculate the input-output relationships for an LTI system in the frequency and z-domain, which are provided in the Table 4.1.

Time/correlation domain	Frequency domain	z-domain
$y[n] = h[n] * x[n]$	$Y(e^{j\theta}) = H(e^{j\theta})X(e^{j\theta})$	$Y(z) = H(z)X(z)$
$r_{yx}[l] = h[l] * r_x[l]$	$P_{yx}(e^{j\theta}) = H(e^{j\theta})P_x(e^{j\theta})$	$P_{yx}(z) = H(z)P_x(z)$
$r_{xy}[l] = h^*[-l] * r_x[l]$	$P_{xy}(e^{j\theta}) = H^*(e^{j\theta})P_x(e^{j\theta})$	$P_{xy}(z) = H^*(1/z^*)P_x(z)$
$r_y[l] = h[l] * r_{xy}[l]$	$P_y(e^{j\theta}) = H(e^{j\theta})P_{xy}(e^{j\theta})$	$P_y(z) = H(z)P_{xy}(z)$
$r_y[l] = h[l] * h^*[-l] * r_x[l]$	$P_y(e^{j\theta}) = \lvert H(e^{j\theta})\rvert^2 P_x(e^{j\theta})$	$P_y(z) = H(z)H^*(1/z^*)P_x(z)$

Table 4.1: Summary of input-output relationship for a random signal input to an LTI system in the time/correlation domains and transform-domains

4.2.2 Innovation representation of a random signal

A rational spectrum is a ratio of two rational functions containing $e^{j\theta}$, as

$$P(e^{j\theta}) = P(z) = \frac{\sum_{q=-Q}^{Q} \gamma_1[q]e^{-j\theta q}}{\sum_{p=-P}^{P} \gamma_2[p]e^{-j\theta p}}. \tag{4.10}$$

Here $\gamma_1[q]$ and $\gamma_2[p]$ are two series with even symmetry, similarly to the autocorrelation sequences of real signals. The rational spectrum is actually derived from Wold's decomposition or representation theorem, a very general signal decomposition theorem, which states that every WSS signal can be written as the sum of two components, one deterministic and one stochastic, as given below

$$x[n] = \sum_{k=0}^{\infty} h[k]w[n-k] + y[n], \tag{4.11}$$

where $h[k]$ represents a infinite-length sequence of weights; $w[n]$ is a random signal of uncorrelated samples, often referred to as innovations (e.g., white noise); $y[n]$ is deterministic, thus exactly predictable from the past values. The deterministic part can be subtracted from this signal since it is exactly predictable.

Let us know consider a purely zero-mean WSS signal, from which any deterministic component has already been subtracted. We take a white noise sequence, $i[n]$, as the uncorrelated samples. Then Wold's decomposition theorem becomes

$$x[n] = \sum_{k=0}^{\infty} h[k]i[n-k] = \sum_{k=-\infty}^{n} h[n-k]i[k]. \tag{4.12}$$

The sequence of weights $h[k]$ can now be seen as the samples of the impulse response of a causal LTI filter, which we require to be stable. We also assume this filter to have a rational transfer function of the form

$$H(z) = \frac{B(z)}{A(z)}, \tag{4.13}$$

with

$$A(z) = 1 + a_1 z^{-1} + ... + a_p z^{-p},$$
$$B(z) = 1 + b_1 z^{-1} + ... + b_q z^{-q}.$$

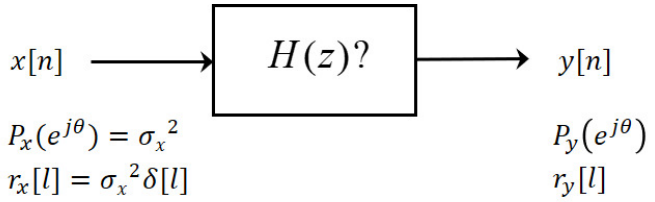


Figure 4.2: Spectral factorization problem: determining $H(z)$ knowing that the input $x[n]$ is white noise, and given the second-order statistics of the output signal $y[n]$.

In (4.2.2), we assume that $a_0 = 1$ and $b_0 = 1$. Note that assuming a rational form for $H(z)$ is not too restrictive. In fact, according to the rational approximation of function, we can always approximate any continuous function by a rational polynomial as closely as we want by increasing the degree of the numerator and denominator in (4.2.2). However, the poles of the transfer function must be inside the unit circle for the filter to be stable. In summary, Wold's decomposition theorem allows us to represent any WSS random signal as the output of an LTI filter driven by white noise. If the filter has transfer function $H(z) = B(z)/A(z)$, then the signal can be rewritten as

$$x[n] = -\sum_{p=1}^{p} a_p x[n-p] + \sum_{q=0}^{Q} b_q i[n-q]. \tag{4.14}$$

Since the input is given by a white-noise sequence, the input autocorrelation is of the form $r_i[l] = \sigma_i^2\delta[l]$. Given that the transfer function is rational, also the power spectrum of the output random signal is rational. Using the results of Table 4.1, the output power spectrum can be calculated as

$$P_x(e^{j\omega}) = \sigma_i^2 \frac{|B(e^{j\omega})|^2}{|A(e^{j\omega})|^2} = \sigma_i^2 |H(e^{j\omega})|^2, \tag{4.15}$$

where

$$\begin{aligned} A(e^{j\omega}) &= 1 + a_1 e^{-j\omega} + ... + a_p e^{-j\omega p}, \\ B(e^{j\omega}) &= 1 + b_1 e^{-j\omega} + ... + b_q e^{-j\omega q}. \end{aligned} \tag{4.16}$$

As we shall see, this means viewing an observed random signal as a stochastic process modeled by an autoregressive moving-average model, whose rational spectrum contains the model parameters.

Spectral factorization

In Chapter 3, we have seen that the PSD and the AC of a random signal are Fourier pairs. Consider the system depicted in Figure 4.2, with input $x[n]$ and output $y[n]$. Spectral factorization tackles the following question: Can we determine $H(z)$ knowing that the input $x[n]$ is white noise, and given the second-order statistics ($r_y[l]$ or equivalently $P_y(e^{j\theta})$) of the output signal $y[n]$? In principle, there is no unique answer to this question. We can see this by considering the following example.

Suppose we are given the PSD of $y[n]$ as $P_y(e^{j\theta}) = \frac{5}{4} - cos(\theta)$ and we know that $y[n]$ is the output of an LTI system driven by white noise, with zero mean and variance σ_i^2. Based on

equation (4.15), we can write:

$$P_y(e^{j\theta}) = \sigma_i^2 |H(e^{j\theta})|^2 = \frac{5}{4} - cos(\theta)$$
$$\sigma_i^2 |H(z)|^2 = \sigma_i^2 H(z)H(z^{-1}) = \frac{5}{4} - \frac{z + z^{-1}}{2}$$

Knowledge of $P_y(e^{j\theta})$ provides us with the magnitude response of the system, i.e. $|H(z)|$. However, there are two possible cases that provide the same magnitude response $|H(z)|^2$:

- In the first case, we can define $H(z)$ and the input noise variance as:

$$H(z) = 1 - \frac{1}{2}z^{-1}, \quad \text{with } \sigma_i^2 = 1.$$

  This choice gives as output $P_y(e^{j\theta})$, as proven below:

$$\begin{aligned}\sigma_i^2 H(z)H(z^{-1}) =& (1 - \frac{1}{2}z^{-1})(1 - \frac{1}{2}z) \\ =& 1 - \frac{z + z^{-1}}{2} + \frac{1}{4} = \frac{5}{4} - \frac{z + z^{-1}}{2}\end{aligned}$$

- Alternatively, we can define $H(z)$ and the input noise variance as:

$$H(z) = 1 - 2z^{-1}, \quad \text{with } \sigma_i^2 = \frac{1}{4}.$$

  This choice gives the same output $P_y(e^{j\theta})$, as proven below:

$$\begin{aligned}\sigma_i^2 H(z)H(z^{-1}) =& \frac{1}{4}(1 - 2z^{-1})(1 - 2z) \\ =& \frac{1}{4} - \frac{z + z^{-1}}{2} + 1 = \frac{5}{4} - \frac{z + z^{-1}}{2}\end{aligned}$$

Therefore, more constraints are needed to uniquely define $H(z)$ which gives $r_y[l]$ or $P_y(e^{j\theta})$. The additional constraint is that the filter $H(z)$ must be minimum-phase. In summary, spectral factorization is defined as the determination of a minimum-phase system from its magnitude response or from its autocorrelation function. In other words, it states: If $P(z)$ is rational, it can be factorized in the following form

$$P(z) = \sigma_i^2 L(z)L(z^{-1}), \tag{4.17}$$

in which the so-called "innovation filter" $L(z)$ is minimum-phase, and σ_i^2 is chosen such that $l[0] = 1$.

One practical method to solve the spectral factorization problem is referred to as the root method. The basic principles are:

- For every rational PSD, that is a PSD that can be expressed as a fraction between two polynomials in $e^{j\omega}$ (or equivalently in z), there exists a unique minimum-phase factorization within a scale factor. To avoid ambiguity, we choose $L(z)$ such that the first filter coefficient of the impulse response is $l[0] = l_0 = 1$;

- For a PSD expressed as a rational polynomial, with a numerator of order Q and a denominator of order P, there are 2^{P+Q} possible rational systems that provide this PSD. From these, we choose the system that is causal, stable, and minimum-phase;

- Not all possible rational systems are valid since for a valid PSD the roots should appear in mirrored pairs, which means that if z_k is a root, then also $1/z_k^*$ is a root.

In the example above, we, therefore, need to choose $H(z) = 1 - \frac{1}{2}z^{-1}$ because it is the minimum-phase choice.

Besides the root-method, another way to find the spectral factorization of a PSD is by polynomial equivalence, as shown in the following example.

Example 4.1

Apply spectral factorization to the following PSD:

$$P(e^{j\theta}) = \frac{5 - 4\cos(\theta)}{10 - 6\cos(\theta)}$$

Give an expression for the innovation filter $L(z)$ and the variance σ_i^2 of the innovation signal.

Solution.

$$P(z) = \frac{5 - 2(z^{-1} + z)}{10 - 3(z^{-1} + z)} = \frac{c_n}{c_d} \cdot \frac{1 - az^{-1}}{1 - bz^{-1}} \cdot \frac{1 - az}{1 - bz} = \sigma_i^2 \cdot L(z) \cdot L(z^{-1})$$

with constants c_n, c_d, a and b; mind that $|a| < 1$ and $|b| < 1$. In this way the innovation filter $L(z)$ is always minimum-phase and the first coefficient of the innovation filter (l_0) equals one, thus $L(z) = 1 + \cdot z^{-1} + \cdots$.

From the equations above, we obtain the following set of equations by polynomial equivalence

$$\begin{aligned} c_n(1 - az^{-1})(1 - az) &= 5 - 2(z^{-1} + z) \\ &\Rightarrow (a = \frac{1}{2} \text{ and } c_n = 4) \text{ or } (a = 2 \text{ and } c_n = 1) \\ c_d(1 - bz^{-1})(1 - bz) &= 10 - 3(z^{-1} + z) \\ &\Rightarrow (b = \frac{1}{3} \text{ and } c_d = 9) \text{ or } (b = 3 \text{ and } c_d = 1) \end{aligned}$$

Since we need to choose that solution that results in a minimum-phase innovation filter, this results in:

$$P(z) = \frac{4}{9} \cdot \left(\frac{1 - \frac{1}{2}z^{-1}}{1 - \frac{1}{3}z^{-1}}\right) \cdot \left(\frac{1 - \frac{1}{2}z}{1 - \frac{1}{3}z}\right) = \sigma_i^2 \cdot L(z) \cdot L(z^{-1})$$

Thus $L(z) = (1 - \frac{1}{2}z^{-1})/(1 - \frac{1}{3}z^{-1})$ and $\sigma_i^2 = 4/9$.

4.3 Autoregressive moving-average models

A stochastic process may be represented by a stochastic model with given order and parameters, which is able to generate a random signal characterized by well-defined spectral properties. In fact, although a stochastic model and process are in principle two different entities, they are sometimes used interchangeably. Stochastic models are fundamental in many applied fields of

science, including engineering, economics, and medicine. Among stochastic models, a special class of models known as autoregressive moving-average (ARMA) is widely used. In the previous section, we saw how any wide-sense stationary random signal can be described as an LTI system driven by white noise. This system, also referred to as the innovation filter, has a rational spectrum. It turns out that this is equivalent to modeling the random process generating the signal as an ARMA(p, q) model, which provides a parsimonious description of a WSS stochastic process in terms of a rational polynomial. The numerator of order q represents the moving-average part, while the denominator of order p represents the autoregressive part.

The ARMA model is based on the observation that stochastic time series often have a dependence on previous time points. This can be described by the autoregressive part of an ARMA model, which permits modeling a "memory" that decades with time. The moving-average part takes into account the new information (innovation) by a linear combination of the present and previous input samples. It smooths the signal by filtering out random short-term fluctuations.

Thanks to their ability to model a wide variety of stochastic processes, ARMA models are useful for:

- understanding the nature of a stochastic process by detecting the mechanism that builds memory into the signal;
- forecasting future values of a signal based on the past values;
- removing from the signal the imprint of some known process, so as to get a more random residual signal to be further analyzed and interpreted (pre-whitening);
- finding a spectral estimate from a random signal. More on this in Chapter 14.

4.3.1 Autoregressive model

An autoregressive process is a special type of ARMA model for which $q = 0$. The AR model is therefore an all-pole filter with the following transfer function

$$H(z) = \frac{1}{1 + \sum_{p=1}^{P} a_p z^{-p}} = \frac{1}{A(z)}. \tag{4.18}$$

The input to this system is white noise, thus an AR random process can be described by the following difference equation

$$x[n] = i[n] - a_1 x[n-1] - a_2 x[n-2] - \ldots - a_p x[n-p]. \tag{4.19}$$

where $i[n]$ is the input white noise and a_i are the filter coefficients. The order of the filter gives an indication on how many previous output samples are used to form a new output.

Example 4.2

One of the first stochastic models was an AR model proposed by Yule in 1927 to describe the motion of a pendulum in a viscous medium. Yule expressed the amplitude $s[n]$ of the oscillation using the following homogeneous difference equation

$$s[n] + a_1 s[n-1] + a_2 s[n-2] = 0, \qquad n = 0, 1, 2, ... \tag{4.20}$$

However, the measured values of $s[n]$ are affected by noise. Therefore, Yule proposed to

use noise as an external driving force determining the pendulum's behavior, resulting in

$$\begin{aligned} s[n] + a_1 s[n-1] + a_2 s[n-2] &= i[n], \qquad n = 0, 1, 2, ..., \\ s[n] &= i[n] - a_1 s[n-1] - a_2 s[n-2], \qquad n = 0, 1, 2, ... \end{aligned} \tag{4.21}$$

which is a 2^{nd} order AR model.

Autocorrelation of an AR(p) process

The autocorrelation function of an AR process can be determined by starting from the definition of autocorrelation. This is a conjugate symmetric function defined as:

$$r_x[l] = r_x^*[-l] = \mathrm{E}\left[x[n]x^*[n-l]\right], \tag{4.22}$$

where we will make use of two properties. First, the autocorrelation function of the independent white Gaussian input noise is defined as

$$r_i[l] = \mathrm{E}\left[i[n]i^*[n-l]\right] = \sigma_i^2 \delta[l] \tag{4.23}$$

and secondly we use the fact that the signal is real, meaning that $x[n] = x^*[n]$.

Before the autocorrelation function of $x[n]$ can be calculated, first the cross-correlation function between the white Gaussian noise input $i[n]$ and the AR-process $x[n]$ has to be determined as

$$\begin{aligned} r_{ix}[l] &= \mathrm{E}\left[i[n]x^*[n-l]\right] \\ &= \mathrm{E}\left[i[n]^* \left(i^*[n-l] - a_1 x^*[n-1-l] - \ldots - a_p x^*[n-p-l]\right)\right] \\ &= \mathrm{E}\left[i[n]i^*[n-l]\right] + \mathrm{E}\left[i[n]\left(-a_1 x^*[n-1-l] - \ldots - a_p x^*[n-p-l]\right)\right] \\ &= \sigma_i^2 \delta[l] - a_1 \mathrm{E}\left[i[n]x^*[n-1-l]\right] - \ldots - a_p \mathrm{E}\left[i[n]x^*[n-p-l]\right], \\ &= \sigma_i^2 \delta[l] \end{aligned} \tag{4.24}$$

where the simplification in the last line requires further clarification. Because the noise is independently distributed, only the expected value of two identically lagged signals $i[n]$ is non-zero. In the latter terms in the equation, the lagged signals $x[n-1+l], \ldots, x[n-p+l]$ only depend on lagged values of the noise $i[n-1+l], \ldots, i[n-p+l]$, which means that this always returns zero, because the white Gaussian noise signals are not calculated at the same lag.

Calculation of the autocorrelation function of the output $x[n]$ now gives

$$\begin{aligned} r_x[l] &= \mathrm{E}\left[x[n]x^*[n-l]\right] \\ &= \mathrm{E}\left[x[n]\left(i^*[n-l] - a_1 x^*[n-1-l] - \ldots - a_p x^*[n-p-l]\right)\right] \\ &= \mathrm{E}\left[x[n]i^*[n-l]\right] - a_1 \mathrm{E}\left[x[n]x^*[n-1-l]\right] - \ldots - a_p \mathrm{E}\left[x[n]x^*[n-p-l]\right] \\ &= r_{i,x}[l] - a_1 r_x[1-l] - \ldots - a_p r_x[p-l] \\ &= \sigma_i^2 \delta[l] - a_1 r_x[1-l] - \ldots - a_p r_x[p-l], \end{aligned} \tag{4.25}$$

which can be rewritten as

$$\sigma_i^2 \delta[l] = \sum_{k=0}^{p} a_k r_x[k-l], \tag{4.26}$$

or, given the symmetric property of the autocorrelation function

$$\sigma_i^2 \delta[l] = \sum_{k=0}^{p} a_k r_x[l-k], \tag{4.27}$$

where the coefficient a_0 equals 1.

Finally, by factoring out $k = 0$ from the sum in (4.27), and remembering that the autocorrelation function for real-valued signals is a symmetric function, we can obtain the modified Yule-Walker equations for calculation of the autocorrelation function of an AR process

$$r[l] = \begin{cases} \sigma_i^2 - \sum_{k=1}^{p} a_k r_x[|l| - k], & \text{for } l = 0, \\ -\sum_{k=1}^{p} a_k r_x[|l| - k], & \text{for } |l| > 0. \end{cases} \tag{4.28}$$

It can be observed that the autocorrelation function of the AR process is recursive. This is caused by the fact that a noise signal that enters the filter will appear at the output and will in this way always be somehow involved in the filter. Because the filter also processes previous outputs, every noise input signal will always be present in this feedback loop.

Alternatively, the autocorrelation function can be calculated starting from its definition through the expectation operator, and using the difference equation, as shown in Example 4.3

Example 4.3

The output of an AR(1) process is given by the following difference equation

$$x[n] = i[n] - a_1 x[n-1].$$

Find the autocorrelation function.

Solution.

The corresponding autocorrelation function can be found as

$$\begin{aligned} r_x[l] &= \mathrm{E}\left[x[n]x^*[n-l]\right] \\ &= \mathrm{E}\left[x[n]\left(i[n-l] - a_1 x[n-1-l]\right)\right] \\ &= \mathrm{E}\left[x[n]i[n-l]\right] - a_1 \mathrm{E}\left[x[n]x[n-1-l]\right] \\ &= \sigma_i^2 \delta[l] - a_1 r_x[1-l]. \end{aligned}$$

Evaluating the equation for the lags 0 and 1 gives

$$r_x[0] = \sigma_i^2 - a_1 r_x[1]$$

and

$$r_x[1] = r_x[-1] = -a_1 r_x[0]$$

respectively, where the autocorrelation function is assumed to be symmetric. From these two equations, we can determine the unknown filter coefficient and the noise variance. This approach will show how the unknown autocorrelation function can be determined. By combining the equations by the substitution of $r_x[1]$ of the second equation into the first gives

$$r_x[0] = \sigma_i^2 - a_1(-a_1 r_x[0]),$$

from which, rearranging the terms, the value of $r_x[0]$ can be determined as

$$r_x[0] = \frac{\sigma_i^2}{1 - a_i^2}$$

and from this $r_x[1]$ can be determined as

$$r_x[1] = -a_1 \frac{\sigma_i^2}{1 - a_i^2}.$$

This recursion can be extended to all lags, leading to the final description of the autocorrelation function as

$$r_x[l] = \frac{\sigma_i^2}{1-a_1^2}(-a_1)^{|l|}.$$

Power spectral density of an AR(p) process

From Table 4.1, we know that filtering an input signal with a filter with frequency response $H(e^{j\theta})$ relates the input and output power spectral densities ($P_I(e^{j\theta})$ and $P_X(e^{j\theta})$, respectively, through

$$P_X(e^{j\theta}) = |H(e^{j\theta})|^2 P_I(e^{j\theta}) = H(e^{j\theta})H^*(e^{j\theta})P_I(e^{j\theta}). \tag{4.29}$$

Note that the transfer function depends on the filter coefficients. This relationship can be determined by calculating the frequency response of the filter. The frequency response $H(e^{j\theta})$ of the filter can be determined by taking the Fourier transform of the difference equation as

$$X(e^{j\theta}) = I(e^{j\theta}) - a_1 X(e^{j\theta})e^{-j\theta} - a_2 X(e^{j\theta})e^{-j2\theta} - \ldots - a_p X(e^{j\theta})e^{-jp\theta}, \tag{4.30}$$

where $X(e^{j\theta})$ and $I(e^{j\theta})$ are the Fourier transforms of the output signal and the white Gaussian process respectively. The terms can be rearranged as

$$I(e^{j\theta}) = X(e^{j\theta})(1 + a_1 e^{-j\theta} + a_2 e^{-j2\theta} + \ldots + a_p e^{-jp\theta}). \tag{4.31}$$

From this, the frequency response of the system can be determined by the fraction of the output transform over the input transform as

$$H(e^{j\theta}) = \frac{X(e^{j\theta})}{I(e^{j\theta})} = \frac{1}{1 + a_1 e^{-j\theta} + a_2 e^{-j2\theta} + \ldots + a_p e^{-jp\theta}}. \tag{4.32}$$

If we combine the result of this derivation with the fact that the PSD of a white Gaussian processes with variance σ_i^2 is given as $P_I(e^{j\theta}) = \sigma_i^2$, we finally obtain using (4.29) that

$$P_X(e^{j\theta}) = \frac{\sigma_i^2}{|1 + a_1 e^{-j\theta} + a_2 e^{-j2\theta} + \ldots + a_p e^{-jp\theta}|^2}. \tag{4.33}$$

4.3.2 Moving-average model

A moving-average process is a special type of ARMA model for which $p = 0$. The MA model is therefore an all-zero filter, with the following transfer function

$$H(z) = 1 + \sum_{q=1}^{Q} b_q z^{-q} = B(z). \tag{4.34}$$

The name moving average can be somewhat misleading. In fact, to actually perform a moving average the coefficient of the filters should be all positive and sum up to unity. However, none of these conditions are valid for a general MA model.

The difference equation of a q^{th}-order MA filter is given by

$$x[n] = i[n] + b_1 i[n-1] + b_2 i[n-2] + \ldots + b_q i[n-q]. \tag{4.35}$$

where $i[n]$ is the input white noise and b_i are the filter coefficients. The filter order determines how many noise (input) samples are combined to form a new sample.

Autocorrelation of an MA(q) process

While the autocorrelation of the AR process is recursive, this is not the case for MA processes. Each sample of the input noise will only be memorized during the time that is present in the filter, determined by the filter length. Therefore, after a certain time, a noise sample will not be present anymore in the output signal. This also leads to the fact that the autocorrelation function will only be non-zero for a certain number of lags, which is a function of the filter length. There is no correlation for lags exceeding the length of the filter. In order to demonstrate this behavior, the autocorrelation function is calculated for several lags.

At $l = 0$ it can be found that

$$\begin{aligned} r_x[0] &= \mathrm{E}\left[x[n]x[n]\right] \\ &= \mathrm{E}\left[(i[n] + b_1 i[n-1] + b_2 i[n-2] + \ldots + b_q i[n-q])^2\right] \\ &= \mathrm{E}\left[i^2[n]\right] + \mathrm{E}\left[b_1^2 i^2[n-1]\right] + \mathrm{E}\left[b_2^2 i^2[n-2]\right] + \ldots + \mathrm{E}\left[b_q^2 i^2[n-q]\right] \\ &= \sigma_i^2(1 + b_1^2 + b_2^2 + \ldots + b_q^2) = \sigma_i^2 \sum_{k=0}^{q} b_k^2, \end{aligned} \tag{4.36}$$

where the terms including the noise signal that were not identically lagged were not shown, because these are zero. Furthermore, the coefficient b_0 is defined as 1. Similarly, at $l = 1$, the autocorrelation function can be determined as

$$\begin{aligned} r_x[1] &= \mathrm{E}\left[x[n]x[n-1]\right] \\ &= \mathrm{E}\left\{(i[n] + b_1 i[n-1] + \ldots + b_q i[n-q]) \right. \\ &\qquad \left. \cdot\, (i[n-1] + b_1 i[n-2] + \ldots + b_q i[n-q-1])\right\} \\ &= b_1 \mathrm{E}\left[i^2[n-1]\right] + b_1 b_2 \mathrm{E}\left[i^2[n-2]\right] + b_2 b_3 \mathrm{E}\left[i^2[n-3]\right] \\ &\qquad + \ldots + b_{q-1} b_q \mathrm{E}\left[i^2[n-q]\right] \\ &= \sigma_i^2(b_1 + b_1 b_2 + b_2 b_3 + \ldots + b_{q-1} b_q) = \sigma_i^2 \sum_{k=1}^{q} b_k b_{k-1}, \end{aligned} \tag{4.37}$$

Please note how the lag has affected the above terms. Furthermore, the number of terms has decreased. The autocorrelation for $l = 2$ can be determined as

$$\begin{aligned} r_x[2] &= \mathrm{E}\left[x[n]x[n-2]\right] \\ &= \mathrm{E}\left\{(i[n] + b_1 i[n-1] + \ldots + b_q i[n-q]) \right. \\ &\qquad \left. \cdot\, (i[n-2] + b_1 i[n-3] + \ldots + b_q i[n-q-2])\right\} \\ &= b_2 \mathrm{E}\left[i^2[n-2]\right] + b_1 b_3 \mathrm{E}\left[i^2[n-3]\right] + b_2 b_4 \mathrm{E}\left[i^2[n-4]\right] \\ &\qquad + \ldots + b_{q-2} b_q \mathrm{E}\left[i^2[n-q]\right] \\ &= \sigma_i^2(b_2 + b_1 b_3 + b_2 b_4 + \ldots + b_{q-2} b_q) = \sigma_i^2 \sum_{k=2}^{q} b_k b_{k-2}. \end{aligned} \tag{4.38}$$

This methodology can be extended for multiple lags, but a pattern should become noticeable, revealing the mathematical structure of the autocorrelation function. The mathematical description of the autocorrelation function can therefore be calculated as

$$r_x[l] = \begin{cases} \sigma_i^2 \sum_{k=|l|}^{q} b_k b_{k-|l|}, & \text{for } 0 \leq |l| \leq q \\ 0 & \text{otherwise} \end{cases} \tag{4.39}$$

Power spectral density of an MA(q) process

From the definition of the difference equation, the Fourier equivalent can be determined as

$$X(e^{j\theta}) = I(e^{j\theta}) + b_1 I(e^{j\theta})e^{-j\theta} + \ldots + b_q I(e^{j\theta})e^{-jq\theta}, \tag{4.40}$$

from which the frequency response of the system can be determined as

$$H(e^{j\theta}) = \frac{X(e^{j\theta})}{I(e^{j\theta})} = 1 + b_1 e^{-j\theta} + \ldots + b_q e^{-jq\theta}. \tag{4.41}$$

Using a similar approach as in the derivation of the AR(p) power spectral density, we can find that the power spectral density of an MA(q) process is given as

$$P_x(e^{j\theta}) = P_I(e^{j\theta})H(e^{j\theta})H^*(e^{j\theta}) = \sigma_i^2|1 + b_1 e^{-j\theta} + \ldots + b_q e^{-jq\theta}|^2. \tag{4.42}$$

4.3.3 Autoregressive moving-average model

The general ARMA model is a mixture of an AR(p) and an MA(q) model, and therefore has both poles and zeros. The resulting transfer function is given by

$$H(z) = \frac{1 + \sum_{q=1}^{Q} b_q z^{-q}}{1 + \sum_{p=1}^{P} a_p z^{-p}} = \frac{B(z)}{A(z)}. \tag{4.43}$$

The input to this system is white noise, thus an ARMA random process can be described by the following difference equation

$$\begin{aligned} x[n] = i[n] + b_1 i[n-1] + b_2 i[n-2] + \ldots + b_q i[n-q] \\ -a_1 x[n-1] - a_2 x[n-2] - \ldots - a_p x[n-p]. \end{aligned} \tag{4.44}$$

where $i[n]$ is the input white noise, a_i are the filter coefficients for the AR part, and b_i are the filter coefficients for the MA part.

Example 4.4

Suppose you have been hired by Signify to manage ligtbulb production. You are asked to estimate each month how many light bulbs to produce next month. Last month the estimation was $l[m-1] = 2500$ and the number of ligtbulbs left in stock were $\epsilon[m-1] = 310$. Modeling the production as an ARMA(1,1), and knowing that you would like to produce 10% more of the needed ligtbulbs to avoid running out of stock, provide an estimate $\hat{l}[m]$ of the ligtbulb to be produced for this month.

Solution.

The ligtbulb production can be modeled using the following ARMA(1,1) process

$$l[m] = \underbrace{a_0 + a_1 l[m-1]}_{\text{AR(1)}} + \underbrace{b_1 \epsilon[m-1] + \epsilon[m]}_{\text{MA(1)}}.$$

In this model, the AR(1) part is a function of the previous month, while the MA(1) incorporates the error to make a new prediction. For both parts, we only look at data from this month and last month's production because we chose order $p = 1$ and $q = 1$.

Since the error for this month cannot be known in advance, we can estimate the

number of ligtbulb needed this month as

$$\hat{l}[m] = a_0 + a_1 l[m-1] + b_1 \epsilon[m-1].$$

To estimate the parameters a_0, a_1, b_1, we can use the information we have from the previous month and the fact that we want to produce 10% more of the needed light bulb. With this information in mind, we can write

$$a_0 + a_1 l[m-1] + b_1 \epsilon[m-1] = 1.1(l[m-1] - \epsilon[m-1]).$$

Although the solution is not unique, equating the coefficients on the left and right hand sides of the above equation leads to $a_0 = 0$, $a_1 = 1.1$, and $b_1 = -1.1$, from which we obtain

$$\hat{l}[m] = 1.1 \cdot 2500 - 1.1 \cdot 310 = 2409.$$

The autocorrelation function of an ARMA process

As we have seen so far, the ARMA process is a combination of a AR and MA process. This also becomes apparent in the autocorrelation function. The derivation is lengthy and will not be shown, but an intuitive description is given. The modified Yule-walker equations can be used to express the relationship between the parameters of the transfer function and the autocorrelation function of an ARMA(p, q) process

$$r_x[l] = \begin{cases} \sigma_i^2 \sum_{k=|l|}^{q} b_k h[k-|l|] - \sum_{k=1}^{p} a_k r_x[|l|-k], & \text{for } 0 \leq |l| \leq q \\ -\sum_{k=1}^{p} a_k r_x[|l|-k]. & \text{for } |l| > q \end{cases} \tag{4.45}$$

where $h[n]$ represents the impulse response of the system. This expression may seem complicated, but it is easily explained by comparison with the autocorrelation function of the AR and MA process. The value of the autocorrelation function is a combination of both AR and MA autocorrelation functions. As with the AR process, the autocorrelation function has a recursive structure as can be seen from the terms with the a_i coefficients. For smaller lags, there is not only an effect of an AR process but there is also the effect of the MA process. In other words, the ARMA process and its autocorrelation function can be interpreted as the super-imposition of the AR and MA processes. For lags larger than q, the MA part does not contribute any more to the autocorrelation function and thus only the AR part is present.

Example 4.5

The random process $x[k]$ is generated by filtering innovation $i[k]$ (white noise sequence with zero mean and unit variance) with filter $H(z) = (1-\frac{1}{3}z^{-1})/(1-\frac{1}{2}z^{-1})$. Calculate the autocorrelation function $r[l]$ of the process $x[k]$ using the modified Yule-Walker equations.

Solution.

From the system function $H(z)$ we can observe that the process is ARMA(1,1) with coefficients $b_1 = -\frac{1}{3}$ and $a_1 = -\frac{1}{2}$. Thus, up to lag ± 1, the autocorrelation function is dependent on both the MA and AR components (first line of the Yule-Walker equations), while for lags larger than ± 1 only the AR component is present. Let's start by calculating

$r[0]$ and $r[-1] = r[1]$.

$$\begin{aligned} r[l=0] &= \sigma_i^2 \sum_{k=0}^{1} b_k h[k] - \sum_{k=1}^{1} a_k r[-k] = \\ &= b_0 h[0] + b_1 h[1] - a_1 r[-1] \\ r[l=1] &= \sigma_i^2 \sum_{k=1}^{1} b_k h[k-1] - \sum_{k=1}^{1} a_k r[1-k] = \\ &= b_1 h[0] - a_1 r[0] \end{aligned} \tag{4.46}$$

In the equations above, we know that $\sigma_i^2 = 1$, $b_0 = 1$, $b_1 = -\frac{1}{3}$ and $a_1 = -\frac{1}{2}$, while $r[0]$ and $r[1] = r[-1]$ are the unknowns we are looking for. However, before we can proceed we need to calculate $h[0]$ and $h[1]$, which are the samples at $k = 0$ and $k = 1$ of the impulse response $h[k]$ of the system. Since we know the system function, we can calculate the first coefficients of the impulse response by polynomial division. In principle, since $H(z)$ is an infinite impulse response system, this division will lead to a polynomial in z^{-1} of infinite length. However, here the task is much easier as we need only the terms at $k = 0$ and $k = 1$. Applying polynomial division, we obtain for the first two terms

$$H(z) = \frac{(1 - \frac{1}{3}z^{-1})}{(1 - \frac{1}{2}z^{-1})} = 1 + \frac{1}{6}z^{-1} + ... \tag{4.47}$$

By taking the inverse Z-transform of both sides of the equation, we obtain:

$$h[k] = \delta[k] + \frac{1}{6}\delta[k-1] + ... \tag{4.48}$$

Thus, $h[0] = 1$ and $h[1] = \frac{1}{6}$. Plugging this into (4.46), we get

$$\begin{aligned} r[0] &= 1 - \frac{1}{3}\frac{1}{6} + \frac{1}{2}r[1] \\ r[1] &= -\frac{1}{3} + \frac{1}{2}r[0] \end{aligned} \tag{4.49}$$

which leads to

$$\begin{aligned} r[0] &= \frac{28}{27} \\ r[1] &= \frac{5}{27}. \end{aligned} \tag{4.50}$$

For lags larger than 1, we need to look at the second line of the Yule-Walker equations, from which we get:

$$\begin{aligned} r[2] &= -a_1 r[1] = \frac{1}{2}r[1] \\ r[3] &= -a_1 r[2] = \frac{1}{2}\frac{1}{2}r[1] \\ r[4] &= -a_1 r[3] = \frac{1}{2}\frac{1}{2}\frac{1}{2}r[1]. \end{aligned} \tag{4.51}$$

From which the recursive structure becomes evident. Thus we can finally write the formula for the autocorrelation functions as

$$r_x[l] = \begin{cases} \frac{28}{27} & \text{for } l = 0 \\ \frac{5}{27}\left(\frac{1}{2}\right)^{|l|-1} & \text{for } |l| \geq 1 \end{cases} \tag{4.52}$$

Power spectral density of an ARMA(p, q) process

From the definition of the difference equation, the Fourier equivalent can be determined as

$$\begin{aligned} X(e^{j\theta}) = I(e^{j\theta}) + b_1 I(e^{j\theta})e^{-j\theta} + \ldots + b_q I(e^{j\theta})e^{-jq\theta} \\ -a_1 X(e^{j\theta})e^{-j\theta} - \ldots - a_p X(e^{j\theta})e^{-jp\theta}, \end{aligned} \tag{4.53}$$

from which the frequency response can be determined as

$$H(e^{j\theta}) = \frac{X(e^{j\theta})}{I(e^{j\theta})} = \frac{1 + b_1 e^{-j\theta} + \ldots + b_q e^{-jq\theta}}{1 + a_1 e^{-j\theta} + \ldots + a_p e^{-jp\theta}}. \tag{4.54}$$

Using a similar approach as in the derivation of the AR(p) and MA(q) PSD functions, we can find that the PSD function of an ARMA(p,q) process is given as

$$P_x(e^{j\theta}) = P_I(e^{j\theta})H(e^{j\theta})H^*(e^{j\theta}) = \sigma_i^2 \frac{|1 + b_1 e^{-j\theta} + \ldots + b_q e^{-jq\theta}|^2}{|1 + a_1 e^{-j\theta} + \ldots + a_p e^{-jp\theta}|^2}. \tag{4.55}$$

Part II

Estimation theory

5

Introduction

In estimation theory, we are interested in determining the value of one or more parameters from some measurements/observations. For example, to use your mobile phone, many quantities such as the carrier frequency offset, initial phase, and timing offset between the mobile device and the base station need to be estimated to establish a connection. Other applications of estimation theory ranges from automotive radars, which estimate the range and the velocity of radar targets, to large physical experiments such as the Large Hadron Collider at CERN.

An estimation process can be described as follows. A parameter θ is mapped to a set of observations $\{x_0, x_1, \ldots, x_{N-1}\}$ by an experiment. The set can be represented by a vector $\mathbf{x} \in \mathbb{R}^N$. Due to measurement inaccuracies or noise, each experiment will result in a different random value. Thus, we can describe the observations by a PDF that is parametrized by the unknown parameters θ. To emphasis the dependency of the PDF on the parameter θ, we denote the PDF by $p(\mathbf{x}; \theta)$. If the PDF is parametrized by multiple parameters, we write $p(\mathbf{x}; \boldsymbol{\theta})$, where $\boldsymbol{\theta} = [\theta_0, \theta_1, \ldots \theta_{K-1}]^T$ are the K parameters to be estimated. The estimator, denoted by $g(\mathbf{x})$, is a mapping from the observations $\mathbf{x}$ to an estimate $\hat{\theta}$. An illustration of the estimation process is provided in Figure 5.1. Note that we can assign many different mappings from the observations to an estimate as shown in the next example.

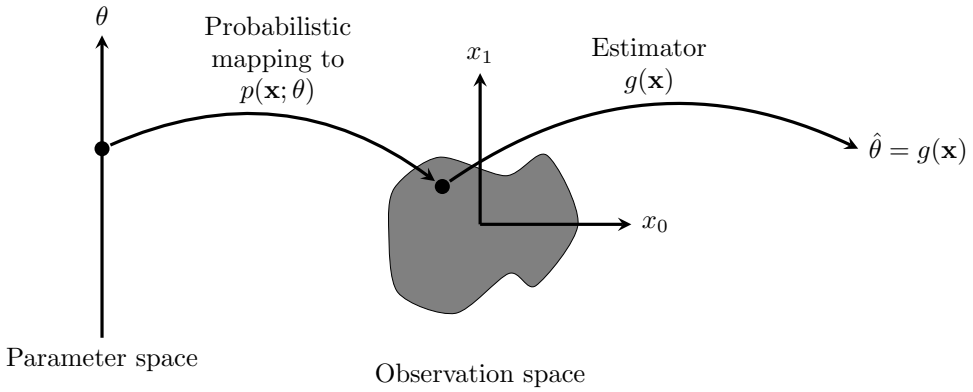


Figure 5.1: Estimation model: The unknown parameter θ is mapped from the parameter space to the observation space via a probabilistic mapping. The estimator itself is a deterministic mapping from the observation space to an estimate $\hat{\theta}$.

Example 5.1

Suppose we want to estimate a DC voltage A which is embedded in noise. Our observation can be modeled as

$$x_n = A + w_n, \tag{5.1}$$

where A is the unknown DC level and $w[n]$ is IID Gaussian noise with zero mean and known variance σ^2. The parametrized PDF of our observation is then

$$p(\mathbf{x}; A) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x_n - A)^2\right]. \tag{5.2}$$

We can define, for example, the following three estimators:

$$\hat{A} = \frac{1}{N}\sum_{n=0}^{N-1} x_n; \tag{5.3}$$

$$\check{A} = x_0; \tag{5.4}$$

$$\tilde{A} = \frac{1}{N+2}\left(2x_0 + \sum_{n=1}^{N-2} x_n - 2x_{N-1}\right). \tag{5.5}$$

The first estimator is the sample mean of the observation. The second estimator considers only the first observation, and the third estimator puts a higher weight on the first and the last observation.

At this point, the question arises as to which of the estimators is the optimal. To answer these question, we need first to define what we mean by optimal. Since the observations $\mathbf{x}$ is a random vector, the estimate $\hat{\theta}$ is a random variable. Thus, we need to evaluate an estimator by its statistic. First of all, it is desirable that an estimate is on average the true parameter, i.e.,

$$\mathrm{E}\left[\hat{\theta}\right] = \theta \tag{5.6}$$

for all θ.

An estimator with this property is said to be *unbiased*. All three estimators presented on Example 5.1 are unbiased, therefore it does not help us to answer which of them is a better choice. A more appropriate criterion for the performance evaluation of an estimator is the variance of the estimate. The variance of a random variable reflects its variability. Thus, we are interested in finding an estimator that has the lowest variance of all unbiased estimators. An estimator whose variance is lower then the variance of all other *unbiased* estimators for all values of θ is referred to as *minimum variance unbiased estimator* (MVUE). In Figure 5.2, the variance of different estimators as a function of the parameter is shown. Among the three different estimators, $\hat{\theta}_3$ has the lowest variance for all values of θ, which makes it the MVUE.

Returning to the three estimators of Example 5.1. The individual variances of the estimates are:

$$\mathrm{Var}\left[\hat{A}\right] = \frac{\sigma^2}{N}; \tag{5.7}$$

$$\mathrm{Var}\left[\check{A}\right] = \sigma^2; \tag{5.8}$$

$$\mathrm{Var}\left[\tilde{A}\right] = \frac{(N+6)}{(N-2)^2}\sigma^2. \tag{5.9}$$

When comparing the different variances, we see that the first estimator has the lowest variance. The variance is N-times smaller, when compared to the second estimator. This is due to the noise averaging effect of the estimator $\hat{A}$. The variance of the last estimator converges to the variance of the first estimator as the sample size increases.

So far, we have seen that we can formulate different estimators, and we can compare their performance by means of their variance. It is logical now to ask if there is a lower bound for the variance of an estimator. In fact there exists a lower bound on the variance for unbiased estimators, the so-called Cramér-Rao lower bound (CRLB). No *unbiased* estimator can have a variance lower than the CRLB. However, depending on the estimation problem, an estimator can attain this lower bound. Such estimators are referred to as *efficient* estimators. For example, the sample mean $\hat{A}$ in the above example is such an estimator. For other estimation problems, we might not be able to find estimators that will attain the CRLB.

Outline

The topics covered in this part are:

- Cramér-Rao lower bound: Just as the data is random, so are the parameter estimations. When the noise properties are known, the lowest possible variance for the parameter estimations can be calculated for unbiased estimators.
- Maximum likelihood estimator: The stochastic nature of the data due to the noise is modeled in terms of the PDF of the noise. The parameter value that maximizes the probability of observing the data at hand is the maximum likelihood estimate.
- Linear models: The minimum variance unbiased estimator can be found for linear problems with Gaussian noise.
- Least squares estimator: This method fits the model to data by minimizing the sum of the squared difference between the data and the model.
- Bayesian estimators: For all estimation techniques so far, the parameter to be estimated is assumed to be deterministic but unknown. Bayesian estimators consider the parameter also as a random variable and utilize the Bayes' Theorem to estimate it.

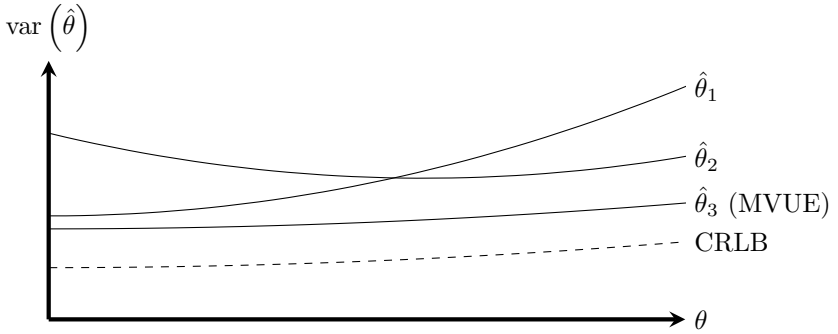


Figure 5.2: Variance of three different estimators. Non of the estimator attains the CRLB. However, the variance of the estimate $\hat{\theta}_3$ is lower than the variance of the other estimates. Thus, $\hat{\theta}_3$ is the MVUE.

- Numerical methods: Not all estimators have closed forms; numerical methods that iteratively estimate the parameters are indispensable tools for implementing estimators.

6

Carmér-Rao lower bound

6.1 Introduction

Many different estimators can be formulated, as shown in the introduction. We also showed that the variance provides a tool to compare the performance of different estimators. We are going to derive a lower bound for the variance of an *unbiased* estimator, the so-called Cramér-Rao lower bound bound (CRLB).

6.2 CRLB for single parameter

The CRLB states that the variance of any unbiased estimator $g(\mathbf{x})$ is lower bounded by

$$\mathrm{Var}[g(\mathbf{x})] \geq \frac{1}{\mathcal{I}(\theta)}, \tag{6.1}$$

where

$$\mathcal{I}(\theta) = \mathrm{E}\left[\left(\frac{\partial \ln p(\mathbf{x};\theta)}{\partial \theta}\right)^2\right] \tag{6.2}$$

$$= -\mathrm{E}\left[\frac{\partial^2 \ln p(\mathbf{x};\theta)}{\partial \theta^2}\right] \tag{6.3}$$

is the so called Fisher information. The quantities $\ln p(\mathbf{x};\theta)$ and $\partial \ln p(\mathbf{x};\theta)/\partial\theta$ are referred to as *log-likelihood* and *score*, respectively.

The inequality (6.1) is valid if the following regularity conditions hold:

$$\frac{d}{d\theta}\int p(\mathbf{x};\theta)d\mathbf{x} = \int \frac{\partial}{\partial\theta} p(\mathbf{x};\theta)d\mathbf{x} \tag{6.4}$$

$$\frac{d^2}{d\theta^2}\int p(\mathbf{x};\theta)d\mathbf{x} = \int \frac{\partial^2}{\partial\theta^2} p(\mathbf{x};\theta)d\mathbf{x} \tag{6.5}$$

The regularity condition in (6.4) and (6.5) ensure that we can interchange the order of differentiation and integration. The condition holds if the domain of integration is independent of the parameter θ, the parameter space or θ is an open interval, and the derivative $p(\mathbf{x};\theta)$ exists and is finite.

Proof. Since we assume that the estimator is unbiased, we have that

$$\mathrm{E}\left[g(\mathbf{x}) - \theta\right] = \int (g(\mathbf{x}) - \theta) p(\mathbf{x}; \theta) d\mathbf{x} = 0. \tag{6.6}$$

Differentiating both sides with respect to θ and using the regularity conditions, we obtain

$$\int (g(\mathbf{x}) - \theta) \frac{\partial}{\partial \theta} p(\mathbf{x}; \theta) d\mathbf{x} - 1 = 0 \tag{6.7}$$

or, equivalently,

$$\int (g(\mathbf{x}) - \theta) \frac{\partial}{\partial \theta} p(\mathbf{x}; \theta) d\mathbf{x} = 1. \tag{6.8}$$

The partial derivative in (6.8) can be expressed as

$$\frac{\partial}{\partial \theta} p(\mathbf{x}; \theta) = p(\mathbf{x}; \theta) \frac{\partial}{\partial \theta} \ln p(\mathbf{x}; \theta). \tag{6.9}$$

Substituting (6.9) in (6.8) yields

$$\int (g(\mathbf{x}) - \theta) p(\mathbf{x}; \theta) \frac{\partial}{\partial \theta} \ln p(\mathbf{x}; \theta) d\mathbf{x} = 1. \tag{6.10}$$

We can rewrite the left-hand side of (6.10) as

$$\int \left[(g(\mathbf{x}) - \theta) \sqrt{p(\mathbf{x}; \theta)} \right] \left[\frac{\partial}{\partial \theta} \ln p(\mathbf{x}; \theta) \sqrt{p(\mathbf{x}; \theta)} \right] d\mathbf{x}, \tag{6.11}$$

which is an inner product between the two expression in the square brackets. The inner product is upper bounded by the Cauchy-Schwarz inequality given as

$$\left(\int \left[(g(\mathbf{x}) - \theta) \sqrt{p(\mathbf{x}; \theta)} \right] \left[\frac{\partial}{\partial \theta} \ln p(\mathbf{x}; \theta) \sqrt{p(\mathbf{x}; \theta)} \right] d\mathbf{x} \right)^2 \\ \leq \int (g(\mathbf{x}) - \theta)^2 p(\mathbf{x}; \theta) d\mathbf{x} \int \left(\frac{\partial}{\partial \theta} \ln p(\mathbf{x}; \theta) \right)^2 p(\mathbf{x}; \theta) d\mathbf{x}. \tag{6.12}$$

From (6.10), we know that the left-hand side of (6.12) is 1. But

$$\int (g(\mathbf{x}) - \theta)^2 p(\mathbf{x}; \theta) d\mathbf{x} = \mathrm{Var}\left[g(\mathbf{x})\right] \tag{6.13}$$

and

$$\int \left(\frac{\partial}{\partial \theta} \ln p(\mathbf{x}; \theta) \right)^2 p(\mathbf{x}; \theta) d\mathbf{x} = \mathrm{E}\left[\left(\frac{\partial}{\partial \theta} \ln p(\mathbf{x}; \theta) \right)^2 \right] \tag{6.14}$$

so that we obtain

$$\mathrm{Var}\left[g(\mathbf{x})\right] \geq \left(\mathrm{E}\left[\left(\frac{\partial}{\partial \theta} \ln p(\mathbf{x}; \theta) \right)^2 \right] \right)^{-1}. \tag{6.15}$$

To obtain the equivalent expression in (6.3), we note that

$$\frac{d}{d\theta} \int p(\mathbf{x}; \theta) d\mathbf{x} = 0. \tag{6.16}$$

This is because $\int p(\mathbf{x};\theta)d\mathbf{x} = 1$ and its derivative is zero. Differentiating a second time with respect to θ and using the (6.5), (6.9), and the chain rule, we get

$$\int \frac{\partial}{\partial\theta} p(\mathbf{x};\theta) \frac{\partial}{\partial\theta} \ln p(\mathbf{x};\theta) d\mathbf{x} + \int p(\mathbf{x};\theta) \frac{\partial^2}{\partial\theta^2} \ln p(\mathbf{x};\theta) d\mathbf{x} = 0. \tag{6.17}$$

Substituting (6.9) again into the first integral yields

$$\int p(\mathbf{x};\theta) \left(\frac{\partial}{\partial\theta} \ln p(\mathbf{x};\theta) \right)^2 d\mathbf{x} + \int p(\mathbf{x};\theta) \frac{\partial^2}{\partial\theta^2} \ln p(\mathbf{x};\theta) d\mathbf{x} = 0 \tag{6.18}$$

or

$$\mathrm{E}\left[\left(\frac{\partial}{\partial\theta} \ln p(\mathbf{x};\theta) \right)^2 \right] = -\mathrm{E}\left[\frac{\partial^2}{\partial\theta^2} \ln p(\mathbf{x};\theta) \right]. \tag{6.19}$$

□

6.3 Efficient estimator

An estimator that attains the lower bound is called *efficient*. In general, there is no guaranty that such an estimator exist at all. However, the Cauchy-Schwarz inequality not only provides a lower bound for the variance but also conditions for equality. Thus, by evaluating the CRLB we might find an *efficient* estimator. If we can express the *score* $\partial \ln p(\mathbf{x};\theta)/\partial\theta$ in the form

$$\frac{\partial}{\partial\theta} \ln p(\mathbf{x};\theta) = \mathcal{I}(\theta)(g(\mathbf{x}) - \theta), \tag{6.20}$$

we can directly determine the expression to find the efficient estimator.

Proof. Equality in (6.12) holds if and only if

$$\frac{\partial}{\partial\theta} \ln p(\mathbf{x};\theta) = a(\theta)(g(\mathbf{x}) - \theta). \tag{6.21}$$

where $a(\theta)$ is an arbitrary function solely depending on θ and not on $\mathbf{x}$. To determine the function $a(\theta)$, we note that

$$\frac{\partial^2}{\partial\theta^2} \ln p(\mathbf{x};\theta) = \frac{\partial}{\partial\theta} \left(a(\theta) \right) (g(\mathbf{x}) - \theta) - a(\theta). \tag{6.22}$$

Taking the expectation and considering the unbiasedness assumption, we obtain

$$-\mathrm{E}\left[\frac{\partial^2}{\partial\theta^2} \ln p(\mathbf{x};\theta) \right] = a(\theta), \tag{6.23}$$

i.e., the function $a(\theta)$ is the Fisher information $\mathcal{I}(\theta)$.

□

Example 6.1

Let us return to the example of estimating the DC voltage embedded in noise presented in the Chapter 5. We want to find the Cramér-Rao lower bound for this estimation problem.

Therefore, we start with the PDF of our observations given as

$$p(\mathbf{x}; A) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x_n - A)^2\right]. \tag{6.24}$$

Taking the logarithm yields

$$\ln p(\mathbf{x}; A) = -\frac{N}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x_n - A)^2, \tag{6.25}$$

and after differentiating with respect to A, we obtain

$$\frac{\partial}{\partial A}\ln p(\mathbf{x}; A) = \frac{1}{\sigma^2}\sum_{n=0}^{N-1}(x_n - A). \tag{6.26}$$

The CRLB can now be found by either evaluating (6.2) or (6.3). For the sake of completeness, let us evaluate both expressions starting with(6.2), which is

$$\mathrm{E}\left[\left(\frac{\partial}{\partial A}\ln p(\mathbf{x}; A)\right)^2\right] = \frac{1}{\sigma^4}\mathrm{E}\left[\left(\sum_{n=0}^{N-1}(x_n - A)\right)\left(\sum_{m=0}^{N-1}(x_m - A)\right)\right] \tag{6.27}$$

$$\begin{aligned}\mathrm{E}\left[\left(\frac{\partial}{\partial A}\ln p(\mathbf{x}; A)\right)^2\right] &= \frac{1}{\sigma^4}\mathrm{E}\left[\left(\sum_{n=0}^{N-1}x_n - NA\right)\left(\sum_{m=0}^{N-1}x_m - NA\right)\right] \\ &= \frac{1}{\sigma^4}\left(\mathrm{E}\left[\sum_{n=0}^{N-1}x_n\sum_{m=0}^{N-1}x_m\right]\right. \\ &\qquad \left. -NA\cdot\left(\sum_{n=0}^{N-1}\mathrm{E}\left[x_n\right] + \sum_{m=0}^{N-1}\mathrm{E}\left[x_m\right]\right) + (NA)^2\right)\end{aligned} \tag{6.28}$$

In the third step, we used linearity property of the expectation operator. Next, we note that $\mathrm{E}\left[x_n\right] = A$, which yields

$$\mathrm{E}\left[\left(\frac{\partial}{\partial A}\ln p(\mathbf{x}; A)\right)^2\right] = \frac{1}{\sigma^4}\left(\mathrm{E}\left[\sum_{n=0}^{N-1}x_n\sum_{m=0}^{N-1}x_m\right] - 2(NA)^2 + (NA)^2\right) \tag{6.29}$$

The first summation evaluates as follows:

$$\begin{aligned}\mathrm{E}\left[\sum_{n=0}^{N-1}x_n\sum_{m=0}^{N-1}x_m\right] &= \mathrm{E}\left[\sum_{n=0}^{N-1}x_n^2 + \sum_{n=0}^{N-1}\sum_{\substack{m=0\\m\neq n}}^{N-1}x_n x_m\right] \\ &= \sum_{n=0}^{N-1}\mathrm{E}\left[x_n^2\right] + \sum_{n=0}^{N-1}\sum_{\substack{m=0\\m\neq n}}^{N-1}\mathrm{E}\left[x_n x_m\right] \\ &= N\left(\sigma^2 + A^2\right) + N(N-1)A^2\end{aligned} \tag{6.30}$$

Here, we first split the two summations into those indices which are equal and those which are different and use the linearity of the operation operator again. Furthermore, we have that $\mathrm{E}\left[x_n^2\right] = \mathrm{Var}[x_n] + \mathrm{E}^2[x_n] = \sigma^2 + A^2$ and since x_n and x_m are independent, we have that $\mathrm{E}\left[x_n x_m\right] = \mathrm{E}\left[x_n\right]\mathrm{E}\left[x_m\right] = A^2$. Substituting (6.30) in (6.29) yields

$$\begin{aligned}\mathrm{E}\left[\left(\frac{\partial}{\partial A}\ln p(\mathbf{x};A)\right)^2\right] &= \frac{1}{\sigma^4}\left(N\sigma^2 + NA^2 + N(N-1)A^2 - 2(NA)^2 + (NA)^2\right)\\ &= \frac{1}{\sigma^4}N\sigma^2\\ &= \frac{N}{\sigma^2}.\end{aligned} \tag{6.31}$$

Evaluating (6.3), we get

$$-\mathrm{E}\left[\frac{\partial^2}{\partial A^2}\ln p(\mathbf{x},A)\right] = -\mathrm{E}\left[-\frac{N}{\sigma^2}\right] = \frac{N}{\sigma^2}, \tag{6.32}$$

which is equivalent to (6.31), and thus, we have that

$$\mathrm{Var}\left[g(\mathbf{x})\right] \geq \frac{\sigma^2}{N}. \tag{6.33}$$

Note that this is the variance of the sample mean estimator presented in the Chapter 5. This result can also be verified by checking if we can express (6.26) in the form of (6.20). Rewriting (6.26) as

$$\frac{\partial}{\partial A}\ln p(\mathbf{x};A) = \underbrace{\frac{N}{\sigma^2}}_{\mathcal{I}(\theta)}\left(\underbrace{\frac{1}{N}\sum_{n=0}^{N-1}x_n}_{g(\mathbf{x})} - \underbrace{A}_{\theta}\right) \tag{6.34}$$

and comparing the single quantities in (6.34) with (6.20), we recognize the expression

$$g(\mathbf{x}) = \frac{1}{N}\sum_{n=0}^{N-1}x_n \tag{6.35}$$

as the efficient estimator for estimating the DC level embedded in white Gaussian noise.

6.4 CRLB for IID observations

The previous example highlighted another property of the CRLB in the case of IID observations. For IID observations, we have that the joint PDF factorizes as

$$p(\mathbf{x};\theta) = \prod_{n=0}^{N-1}p(x_n;\theta), \tag{6.36}$$

and since the logarithm of a product is the sum of the individual logarithms, we further have that

$$\ln p(\mathbf{x};\theta) = \sum_{n=0}^{N-1} \ln p(x_n;\theta). \tag{6.37}$$

Taking the second derivative and the negative expectation over the observations $\mathbf{x}$, we get

$$\begin{aligned} -\operatorname{E}\left[\frac{\partial^2}{\partial\theta^2}\ln p(\mathbf{x};\theta)\right] &= -\sum_{n=0}^{N-1}\operatorname{E}\left[\frac{\partial^2}{\partial\theta^2}\ln p(x_n;\theta)\right] \\ &= -N\operatorname{E}\left[\frac{\partial^2}{\partial\theta^2}\ln p(x_n;\theta)\right] \\ &= Ni(\theta). \end{aligned} \tag{6.38}$$

Thus, the Fisher information provided by all N observations is N-times the Fisher information of a single observation $i(\theta) = -\operatorname{E}[\ln p(x_n;\theta)]$. Consequently, the CRLB decreases as the number of observations increases.

6.5 General CRLB for signals in white Gaussian noise

In many engineering applications, we are confronted with signals corrupted by additive white Gaussian noise. In this case, we can model the observations as

$$x_n = s[n;\theta] + w_n, \qquad n = 0, 1, \ldots, N-1, \tag{6.39}$$

and the corresponding PDF as

$$p(\mathbf{x};\theta) = \frac{1}{(2\pi\sigma^2)^{N/2}}\exp\left(-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x_n - s[n;\theta])^2\right). \tag{6.40}$$

Differentiating the logarithm of the PDF twice yields

$$\frac{\partial^2}{\partial\theta^2}\ln p(\mathbf{x};\theta) = \frac{1}{\sigma^2}\sum_{n=0}^{N-1}\left((x_n - s[n;\theta])\frac{\partial^2}{\partial\theta^2}s[n;\theta] - \left(\frac{\partial}{\partial\theta}s[n;\theta]\right)^2\right) \tag{6.41}$$

Calculating the negative expectation with respect to the observation $\mathbf{x}$ results in

$$-\operatorname{E}\left[\frac{\partial^2}{\partial\theta^2}\ln p(\mathbf{x};\theta)\right] = \frac{1}{\sigma^2}\sum_{n=0}^{N-1}\left(\frac{\partial}{\partial\theta}s[n;\theta]\right)^2 \tag{6.42}$$

and the CRLB can be expressed as

$$\operatorname{Var}[g(\mathbf{x})] \geq \frac{1}{\frac{1}{\sigma^2}\sum_{n=0}^{N-1}\left(\frac{\partial}{\partial\theta}s[n;\theta]\right)^2}. \tag{6.43}$$

We can directly apply this result to the DC voltage estimation problem of Example 5.1 and 6.1. The signal model is given by

$$s[n;A] = A, \tag{6.44}$$

and the partial derivative is simply

$$\frac{\partial s[n,A]}{\partial A} = 1. \tag{6.45}$$

Thus, (6.43) becomes in this particular case

$$\text{Var}\left[g(\mathbf{x})\right] \geq \frac{1}{\frac{1}{\sigma^2}\sum_{n=0}^{N-1} 1} = \frac{\sigma^2}{N}. \tag{6.46}$$

6.6 Parameter transformation

Sometimes it is of interest to estimate a parameter which is related to another parameter. For example, instead of estimating the amplitude A of a signal, we could be interested in estimating the squared amplitude A^2. For such cases and if there exists a one to one mapping between the two parameters, it is sufficient to know the Cramér-Rao lower boundof one of the two parameter to compute the Cramér-Rao lower bound for the other parameter.

Let α be the parameter of interest which is related to θ through the one-to-one mapping $\alpha = h(\theta)$. Then the variance of the estimate $\hat{\alpha}$ is lower bound by

$$\text{Var}[\hat{\alpha}] \geq \left(\frac{dh(\theta)}{d\theta}\right)^2 \frac{1}{\mathcal{I}(\theta)}. \tag{6.47}$$

Proof. We start again with the unbiasedness assumption

$$\text{E}[\hat{\alpha}] = \alpha = h(\theta), \tag{6.48}$$

which can alternatively expressed as

$$\int (\hat{\alpha} - h(\theta))p(\mathbf{x};\theta)d\mathbf{x} = 0. \tag{6.49}$$

Using the regularity condition, we may interchange the order of integration and differentiation to obtain

$$\int \frac{\partial}{\partial\theta}(\hat{\alpha} - h(\theta))p(\mathbf{x};\theta)d\mathbf{x} = 0. \tag{6.50}$$

Next, we apply the chain rule to produce

$$\int (\hat{\alpha} - h(\theta))\frac{\partial}{\partial\theta}p(\mathbf{x};\theta)d\mathbf{x} = \frac{\partial}{\partial\theta}h(\theta). \tag{6.51}$$

After substituting (6.9) and squaring both sides, we have

$$\left(\int (\hat{\alpha} - h(\theta))p(\mathbf{x};\theta)\frac{\partial}{\partial\theta}\ln p(\mathbf{x};\theta)d\mathbf{x}\right) = \left(\frac{\partial}{\partial\theta}h(\theta)\right)^2 \tag{6.52}$$

The left hand side is upper bounded by the Cauchy-Schwarz inequality resulting in

$$\int (\hat{\alpha} - h(\theta))^2 p(\mathbf{x};\theta)d\mathbf{x} \int \left(\frac{\partial}{\partial}\ln p(\mathbf{x};\theta)\right)^2 p(\mathbf{x};\theta)d\mathbf{x} \geq \left(\frac{\partial}{\partial\theta}h(\theta)\right)^2 \tag{6.53}$$

or, equivalently,

$$\text{Var}[\hat{\alpha}] \geq \left(\frac{dh(\theta)}{d\theta}\right)^2 \frac{1}{\mathcal{I}(\theta)} \tag{6.54}$$

□

While efficiency is preserved in a linear (affine) transformation, this is generally not the case. This can be seen from a simple counter example. Therefore consider the Cramér-Rao lower bound of estimating a DC voltage in Example 6.1. For this example, the sample mean was an efficient estimator. However, the transformed sample mean is not even unbiased estimator since

$$\mathrm{E}\left[\left(\frac{1}{N}\sum_{n=0}^{N-1}x_n\right)^2\right] = \mathrm{E}^2\left[\frac{1}{N}\sum_{n=0}^{N-1}x_n\right] + \mathrm{Var}\left[\frac{1}{N}\sum_{n=0}^{N-1}x_n\right] = A^2 + \mathrm{Var}\left[\frac{1}{N}\sum_{n=0}^{N-1}x_n\right] \neq A^2. \quad (6.55)$$

6.7 CRLB for vector parameter

Let $\boldsymbol{\theta} = [\theta_0, \theta_1, \ldots, \theta_{K-1}]^T$ be the vector holding the unknown parameters and let $\hat{\boldsymbol{\theta}} = g(\mathbf{x})$ be the estimate of the parameter vector. Then, for any unbiased estimator, the covariance matrix $\mathbf{C}_{\hat{\boldsymbol{\theta}}}$ given as

$$\mathbf{C}_{\hat{\boldsymbol{\theta}}} = \mathrm{E}\left[(g(\mathbf{x}) - \boldsymbol{\theta})(g(\mathbf{x}) - \boldsymbol{\theta})^T\right], \quad (6.56)$$

is lower bounded by

$$\mathbf{C}_{\hat{\boldsymbol{\theta}}} \geq \mathbf{I}^{-1}(\boldsymbol{\theta}), \quad (6.57)$$

where $\mathbf{I}(\boldsymbol{\theta})$ is the so-called Fisher information matrix with entries

$$\begin{aligned}[\mathbf{I}(\boldsymbol{\theta})]_{i,j} &= \mathrm{E}\left[\left(\frac{\partial}{\partial\theta_i}\ln p(\mathbf{x};\boldsymbol{\theta})\right)\left(\frac{\partial}{\partial\theta_j}p(\mathbf{x};\boldsymbol{\theta})\right)\right] \\ &= -\mathrm{E}\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\ln p(\mathbf{x};\boldsymbol{\theta})\right].\end{aligned} \quad (6.58)$$

The proof is out of the scope of this course.

The notation $\mathbf{C}_{\hat{\boldsymbol{\theta}}} \geq \mathbf{I}^{-1}(\boldsymbol{\theta})$ in (6.57) refers to the condition that the difference of the matrices is *positive semi-definite*, i.e., $\mathbf{a}^T(\mathbf{C}_{\hat{\boldsymbol{\theta}}} - \mathbf{I}^{-1}(\boldsymbol{\theta}))\mathbf{a} \geq 0$ for arbitrary $\mathbf{a} \neq \mathbf{0}$. The diagonal elements of a *positive semi-definite* matrices are nonnegative, thus,

$$\mathrm{Var}\left[\hat{\theta}_i\right] = \left[\mathbf{C}_{\hat{\boldsymbol{\theta}}}\right]_{ii} \geq \left[\mathbf{I}^{-1}(\boldsymbol{\theta})\right]_{ii}. \quad (6.59)$$

Similar to the scalar case in (6.20), equality holds if and only if

$$\frac{\partial}{\partial\boldsymbol{\theta}}\ln p(\mathbf{x},\boldsymbol{\theta}) = \mathbf{I}(\boldsymbol{\theta})(g(\mathbf{x}) - \boldsymbol{\theta}). \quad (6.60)$$

Example 6.2

Suppose we want to estimate in addition to the DC voltage level of the previous example

also the variance σ^2 of the noise. The required expressions are:

$$\frac{\partial}{\partial A}\ln p(\mathbf{x};\boldsymbol{\theta}) = \frac{1}{\sigma^2}\sum_{n=0}^{N-1}(x_n - A); \tag{6.61}$$

$$\frac{\partial}{\partial \sigma^2}\ln p(\mathbf{x};\boldsymbol{\theta}) = -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{n=0}^{N-1}(x_n - A)^2; \tag{6.62}$$

$$\frac{\partial^2}{\partial A^2}\ln p(\mathbf{x};\boldsymbol{\theta}) = -\frac{N}{\sigma^2}; \tag{6.63}$$

$$\frac{\partial^2}{\partial (\sigma^2)^2}\ln p(\mathbf{x};\boldsymbol{\theta}) = \frac{N}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{n=0}^{N-1}(x_n - A)^2; \tag{6.64}$$

$$\frac{\partial^2}{\partial A \partial \sigma^2}\ln p(\mathbf{x};\boldsymbol{\theta}) = \frac{\partial^2}{\partial \sigma^2 \partial A}\ln p(\mathbf{x};A) = -\frac{1}{\sigma^4}\sum_{n=0}^{N-1}(x_n - A)^2. \tag{6.65}$$

The corresponding Fisher information matrix is

$$\mathbf{I}(\boldsymbol{\theta}) = \begin{bmatrix} \frac{N}{\sigma^2} & 0, \\ 0 & \frac{N}{2\sigma^4} \end{bmatrix}, \tag{6.66}$$

and since it is a diagonal matrix, it follows that its inverse is simply the inverse of the main diagonal entries, thus

$$\mathbf{I}^{-1}(\boldsymbol{\theta}) = \begin{bmatrix} \frac{\sigma^2}{N} & 0 \\ 0 & \frac{2\sigma^4}{N} \end{bmatrix}. \tag{6.67}$$

The individual variances are lower bounded by

$$\mathrm{Var}[A] \geq \frac{\sigma^2}{N}, \tag{6.68}$$

and

$$\mathrm{Var}[\sigma^2] \geq \frac{2\sigma^4}{N}. \tag{6.69}$$

6.8 Parameter transformation for vector parameter

Similarly to the scalar case, we can find the Cramér-Rao lower bound for a transformed vector parameter $\boldsymbol{\alpha} = h(\boldsymbol{\theta})$, where $\boldsymbol{\alpha}$ is an r-dimensional vector and $\boldsymbol{\theta}$ is a p-dimensional vector. The Cramér-Rao lower bound for the estimate $\hat{\boldsymbol{\alpha}}$ is then

$$\mathbf{C}_{\hat{\boldsymbol{\alpha}}} \geq \frac{\partial h(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\mathbf{I}^{-1}(\boldsymbol{\theta})\frac{\partial h(\boldsymbol{\theta})^T}{\partial \boldsymbol{\theta}}. \tag{6.70}$$

assuming that the mapping is one-to-one. Here, $\partial h(\boldsymbol{\theta})/\partial \boldsymbol{\theta}$ is the $r \times p$ Jacobian matrix defined as

$$\frac{\partial h(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \begin{bmatrix} \frac{\partial}{\partial \theta_1} h_1(\boldsymbol{\theta}) & \frac{\partial}{\partial \theta_2} h_1(\boldsymbol{\theta}) & \cdots & \frac{\partial}{\partial \theta_p} h_1(\boldsymbol{\theta}) \\ \frac{\partial}{\partial \theta_1} h_2(\boldsymbol{\theta}) & \frac{\partial}{\partial \theta_2} h_2(\boldsymbol{\theta}) & \cdots & \frac{\partial}{\partial \theta_p} h_2(\boldsymbol{\theta}) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial}{\partial \theta_1} h_r(\boldsymbol{\theta}) & \frac{\partial}{\partial \theta_2} h_r(\boldsymbol{\theta}) & \cdots & \frac{\partial}{\partial \theta_p} h_r(\boldsymbol{\theta}) \end{bmatrix}. \tag{6.71}$$

7

Maximum likelihood estimator

7.1 Introduction

The maximum likelihood estimator (MLE) is a popular approach to estimation problems. Firstly, if an *efficient* estimator exists, it is the MLE. Secondly, even if no efficient estimator exists, the mean and the variance converges asymptotically to the true parameter and CRLB as the number of observation increases. Thus, the MLE is *asymptotically unbiased* and *asymptotically efficient*. The principle of the MLE is to find the maximum of the so-called *likelihood* function.

7.2 Maximum likelihood estimation

Before defining the MLE, we define the likelihood function $\mathcal{L}(\mathbf{x};\theta)$. The likelihood function is the PDF $p(\mathbf{x};\theta)$ for a given observation $\mathbf{x}$. Since we fix the observation $\mathbf{x}$, the PDF depends only on the unknown parameter. The value of θ that maximizes the likelihood function is the maximum likelihood estimate $\hat{\theta}_{\mathrm{ML}}$, i.e.,

$$\hat{\theta}_{\mathrm{ML}} = \arg\max_{\theta} p(\mathbf{x};\theta). \tag{7.1}$$

In other words, the maximum likelihood estimate is the value of θ that most likely caused the observation $\mathbf{x}$. Note that depending on the estimation problem, no maximum or multiple maxima exist.

Example 7.1

Let $\{x_0, x_1, \ldots, xN-1\}$ be IID uniformly distributed random variables with PDF

$$p(x_i;\theta) = \begin{cases} \theta^{-1} & 0 \leq x_i \leq \theta, \\ 0 & \text{else.} \end{cases} \tag{7.2}$$

The unknown parameter $\theta > 0$ determines the length of the interval. Due to the IID assumption, we further have that

$$p(\mathbf{x};\theta) = \begin{cases} \theta^{-N}, & \text{if } 0 < x_i < \theta \text{ for } 0 \leq i \leq N-1 \\ 0, & \text{else.} \end{cases} \tag{7.3}$$

The value of θ must be larger than or equal to the largest value in our observed data, otherwise, we have zero probability of obtaining the data. Thus, we can equivalently express the PDF as

$$p(\mathbf{x};\theta) = \begin{cases} \theta^{-N}, & 0 < \max(x_0, x_1, \ldots, x_{N-1}) < \theta \\ 0 & \text{else.} \end{cases} \tag{7.4}$$

The likelihood function is a strictly monotonically decreasing function and is maximized by minimizing the value of θ. However, since $\theta \geq \max(x_0, x_1, \ldots, x_{N-1})$, we get

$$\hat{\theta}_{\text{ML}} = \max(x_0, x_1, \ldots, x_{N-1}). \tag{7.5}$$

Instead of maximizing the likelihood function, we can also maximize $\ln p(\mathbf{x};\theta)$. Since the logarithm is a monotonically increasing function, $\ln p(\mathbf{x};\theta)$ and $p(\mathbf{x};\theta)$ have their maxima for the same value of θ. If the log-likelihood function is continuously differentiable and the maximum is at an interior point, we further have that the derivative is equal to zero at its maxima, i.e.,

$$\frac{\partial}{\partial\theta}\ell(\mathbf{x};\theta)\bigg|_{\theta=\hat{\theta}_{\text{ML}}} = 0. \tag{7.6}$$

The above equation is referred to as the likelihood equation. We already encountered the log-likelihood function in (6.2) and (6.3) in Chapter 6, which indicates its fundamental importance in estimation theory. If the log-likelihood function is not differentiable, other techniques have to be applied.

We have shown in Section 6.3 that an efficient estimator can be obtained if the derivative of the log-likelihood function can be expressed as

$$\frac{\partial}{\partial\theta}\ln p(\mathbf{x};\theta) = \mathcal{I}(\theta)(g(\mathbf{x}) - \theta). \tag{7.7}$$

Combining (7.6) and (7.7) yields

$$\mathcal{I}(\theta)(g(\mathbf{x}) - \theta)|_{\theta=\hat{\theta}_{\text{ML}}} = 0. \tag{7.8}$$

Since the Fisher information is a strictly positive quantity ($\mathcal{I}(\theta) > 0$), we require that

$$g(\mathbf{x}) = \hat{\theta}_{\text{ML}}. \tag{7.9}$$

Consequently, if an efficient estimator exists, then it is the MLE.

Example 7.2

We have already seen that the sample mean is an efficient estimator for estimating the DC level in the presence of additive white Gaussian noise. Moreover, we have just shown that if an efficient estimator exists, it is the MLE; and thus, the sample mean is also the MLE. We can verify this by looking at the partial derivative of the log-likelihood function, which is

$$\frac{\partial}{\partial\theta}p(\mathbf{x};\theta) = \frac{1}{\sigma^2}\sum_{n=0}^{N-1}(x_n - A). \tag{7.10}$$

After equating with zero and solving for A, we have that

$$\hat{A}_{\mathrm{ML}} = \frac{1}{N} \sum_{n=0}^{N-1} x_n, \tag{7.11}$$

which is the efficient estimator found in Chapter 6.

7.3 Asymptotic properties of the maximum likelihood estimator

In general, we would expect that when more samples are available, the more trustworthy our estimate should become. Therefore, we define the following property: an estimator is referred to as consistent, if

$$\lim_{N \to \infty} \Pr\left[|g(\mathbf{x}) - \theta| > \varepsilon\right] = 0 \tag{7.12}$$

for $\varepsilon > 0$ holds, i.e., the estimator converges in probability to the true value. The MLE is an estimator with this property.

Furthermore, if the PDF $p(\mathbf{x}; \theta)$ obeys the regularity condition presented in Chapter 6, then the the following properties of the MLE hold:

$$\mathrm{E}\left[\hat{\theta}_{\mathrm{ML}}\right] \to \theta; \tag{7.13}$$

$$\mathrm{Var}\left[\hat{\theta}_{\mathrm{ML}}\right] \to \mathcal{I}^{-1}(\theta), \tag{7.14}$$

i.e., the MLE is *asymptotically unbiased* and *asymptotically efficient*. Additionally, the MLE is *asymptotically normal*. Combining these properties, we have for $N \to \infty$ that

$$\hat{\theta}_{\mathrm{ML}} \sim \mathcal{N}\left(\theta, \mathcal{I}^{-1}(\theta)\right). \tag{7.15}$$

The proof of this property is rather involved and is out of the scope of this course. Note that *consistency* is a stronger argument than *asymptotic unbiasedness* since it requires the estimator to converge to the true value and not just the expected value.

Example 7.3

Suppose we observe N samples from a signal given as

$$x_n = A + w_n, \tag{7.16}$$

where w_n is IID white Gaussian noise with zero mean and variance A, i.e.,

$$w_n \sim \mathcal{N}(0, A). \tag{7.17}$$

Compared to the problems studied in Chapter 6, here the unknown parameter A is also reflected in the variance of the signal. Since the noise is additive, the likelihood function is given as

$$\ln p(\mathbf{x}; A) = \frac{1}{(2\pi A)^{\frac{N}{2}}} \exp\left(-\frac{1}{2A} \sum_{n=0}^{N-1} (x_n - A)^2\right). \tag{7.18}$$

For this estimation problem, we have that

$$\frac{\partial}{\partial A}\ln p(\mathbf{x};A) = -\frac{N}{2A} + \frac{1}{A}\sum_{n=0}^{N-1}(x_n - A) + \frac{1}{2A^2}\sum_{n=0}^{N-1}(x_n - A)^2, \tag{7.19}$$

and the Fisher information is

$$\mathcal{I}(A) = \frac{N\left(A + \frac{1}{2}\right)}{A^2}. \tag{7.20}$$

If an efficient estimator exists, we should be able to express, according to (6.20), the derivative of the log-likelihood function (7.19) as

$$\frac{\partial}{\partial A}\ln p(\mathbf{x};A) = \mathcal{I}(A)(g(\mathbf{x}) - A), \tag{7.21}$$

which is in this particular case not possible. However, we can still determine the maximum likelihood estimator by equating (7.19) to zero, which yields

$$\begin{aligned} -\frac{N}{2A} + \frac{1}{A}\sum_{n=0}^{N-1}(x_n - A) + \frac{1}{2A^2}\sum_{n=0}^{N-1}(x_n - A)^2 &= 0 \\ -\frac{N}{2A} - N + \frac{1}{2A^2}\sum_{n=0}^{N-1}x_n^2 + \frac{N}{2} &= 0 \\ A^2 + NA - \frac{1}{N}\sum_{n=0}^{N-1}x_n^2 &= 0 \end{aligned} \tag{7.22}$$

After completing the square and solving for A, we obtain

$$\hat{A}_{\mathrm{ML}} = -\frac{1}{2} + \sqrt{\frac{1}{N}\sum_{n=0}^{N-1}x_n^2 + \frac{1}{4}}, \tag{7.23}$$

as the solution for the maximum of the likelihood function. Unfortunately, the estimator is biased since

$$\mathrm{E}\left[-\frac{1}{2} + \sqrt{\frac{1}{N}\sum_{n=0}^{N-1}x_n^2 + \frac{1}{4}}\right] > -\frac{1}{2} + \sqrt{\frac{1}{N}\sum_{n=0}^{N-1}\mathrm{E}\left[x_n^2\right] + \frac{1}{4}} = A, \tag{7.24}$$

which follows from Jensen's inequality. However, by applying the law of large numbers, it follows that for large N

$$\frac{1}{N}\sum_{n=0}^{N-1}x_n^2 \to \mathrm{E}[x_n^2] = \mathrm{E}^2\left[x_n\right] + \mathrm{Var}\left[x_n\right] = A^2 + A, \tag{7.25}$$

Substituting (7.25) in (7.23) results in

$$\hat{A}_{\mathrm{ML}} \to A \tag{7.26}$$

as the number of samples increases. Thus, the MLE is asymptotically consistent.

7.4 Transformed parameters

For some specific problems, we are interested in finding a transformed parameter $\alpha = h(\theta)$, i.e., a parameter which depends on θ. If $h(\theta)$ is a one-to-one mapping, then the maximum likelihood estimate of α is

$$\hat{\alpha}_{\mathrm{ML}} = h\left(\hat{\theta}_{\mathrm{ML}}\right), \tag{7.27}$$

which is known as the invariance property of the maximum likelihood estimator.

7.5 Maximum likelihood estimator for vector parameter

The concept of the MLE can also be used in estimating multiple parameters. If the maximum of the likelihood function is in the interior of the domain on which the PDF is defined and if the partial derivative of the likelihood function with respect to all parameters exists, then a necessary condition for the maximum is

$$\frac{\partial}{\partial \boldsymbol{\theta}} \ln p(\mathbf{x}; \boldsymbol{\theta}) = \mathbf{0}. \tag{7.28}$$

Thus, the maximum can be found by equating the gradient of the log-likelihood function to zero.

7.6 Properties of the maximum likelihood estimator for vector parameter

The MLE for vector parameter possesses the same asymptotic properties as the MLE for scalar parameter. We summarize these properties as follows: for $N \to \infty$, the maximum likelihood estimate is

$$\hat{\boldsymbol{\theta}}_{\mathrm{ML}} \sim \mathcal{N}\left(\boldsymbol{\theta}, \mathbf{I}^{-1}(\boldsymbol{\theta})\right), \tag{7.29}$$

where $\mathbf{I}(\boldsymbol{\theta})$ is the Fisher information matrix evaluated at the true value of the unknown parameter.

Example 7.4

Consider the example of estimating the DC level and the noise variance presented in the Chapter 6. We already evaluated the expressions we needed to calculate the gradient in (6.61) and (6.62), given as

$$\frac{\partial}{\partial A} \ln p(\mathbf{x}; \boldsymbol{\theta})(\mathbf{x}; \boldsymbol{\theta}) = \frac{1}{\sigma^2} \sum_{n=0}^{N-1} (x_n - A), \tag{7.30}$$

$$\frac{\partial}{\partial \sigma^2} \ln p(\mathbf{x}; \boldsymbol{\theta})((\mathbf{x}; \boldsymbol{\theta}) = -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{n=0}^{N-1} (x_n - A)^2. \tag{7.31}$$

We already obtained the maximum likelihood estimate for the DC level, which is the sample mean

$$\hat{A} = \bar{x} = \frac{1}{N} \sum_{n=0}^{N-1} x_n. \tag{7.32}$$

Equating (7.31) to zero and solving for σ^2 using $\bar{x}$ yields

$$\hat{\sigma^2} = \frac{1}{N}\sum_{n=0}^{N-1}(x_n - \bar{x})^2. \tag{7.33}$$

The sample mean $\bar{x}$ is a scaled sum of IID Gaussian random variables, with mean A and variance σ^2/N, i.e.,

$$\bar{x} \sim \mathcal{N}\left(A, \frac{\sigma^2}{N}\right). \tag{7.34}$$

On the other hand, by the central limit theorem, it can be shown that

$$\hat{\sigma^2} \sim \mathcal{N}\left(\frac{N-1}{N}\sigma^2, \frac{2(N-1)}{N^2}\sigma^4\right), \tag{7.35}$$

which, for large N, can be approximated by

$$\hat{\sigma^2} \sim \mathcal{N}\left(\sigma^2, \frac{2}{N}\sigma^4\right). \tag{7.36}$$

Thus, we have that our estimates are asymptotically unbiased, and their variance approaches the CRLB.

8

Linear models

8.1 Introduction

The MVUE has the lowest variance among all unbiased estimators. In general, there is no method to find the MVUE. However, if the signal model is linear, we can find the efficient estimator, and thereby the MVUE.

The chapter begins by introducing the linear signal model and derives the efficient estimators for colored Gaussian noise. Next, we assume white Gaussian noise, which can be considered a special case of colored noise.

8.2 Linear signal model

A linear signal model is any model for which the signal $s[n, \theta]$ can be expressed as a linear combination of the weighted parameters, i.e.,

$$s(\boldsymbol{\theta}) = \sum_{k=0}^{K-1} h_k \theta_k, \tag{8.1}$$

where $h_k[n]$ are the weights for the kth parameter at time instance n. For multiple observations, the model can be expressed as a matrix-vector product of the form

$$\mathbf{s}(\boldsymbol{\theta}) = \mathbf{H}\boldsymbol{\theta}, \tag{8.2}$$

where $\mathbf{s}(\boldsymbol{\theta})$ is the signal vector of length N, $\mathbf{H}$ is the so-called *observation matrix* of size $N \times K$, and $\boldsymbol{\theta}$ is the parameter vector of length K.

Example 8.1

Assume our signal model is a polynomial of degree $K-1$ with coefficients $\theta_0, \theta_1, \ldots, \theta_{K-1}$, i.e.,

$$s_n(\boldsymbol{\theta}) = \theta_0 n^0 + \theta_1 n^1 + \cdots + \theta_{K-1} n^{K-1}, \tag{8.3}$$

and suppose observations for $n = 0, 1, \ldots, N-1$ are given. The corresponding observation

matrix $\mathbf{H}$ is

$$\mathbf{H} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1 & 1 & 1 & & 1 \\ 1 & 2 & 2^2 & & 2^{K-1} \\ \vdots & & & \ddots & \\ 1 & N-1 & (N-1)^2 & \cdots & (N-1)^{K-1} \end{bmatrix}. \tag{8.4}$$

8.3 Linear models and additive Gaussian noise

To find the efficient for linear signals in Gaussian noise, we will take advantage of the equality constraint of the CRLB. Therefore, we need to show that we can obtain following expression from the joint pdf:

$$\frac{\partial}{\partial \boldsymbol{\theta}} \ln p(\mathbf{x}; \boldsymbol{\theta}) = \mathcal{I}(\boldsymbol{\theta})(g(\mathbf{x}) - \boldsymbol{\theta}), \tag{8.5}$$

see Section 6.3, which has been extended here to the vector parameter case. To obtain the derivative, we need to determine first the PDF of the observations. For the linear signal model embedded in additive Gaussian noise, the observations are modeled as

$$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}, \tag{8.6}$$

where $\mathbf{w}$ is zero mean noise with covariance matrix $\mathbf{C}$, i.e., $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$. Thus, the PDF of the observation vector is

$$p(\mathbf{x}, \boldsymbol{\theta}) = \frac{1}{|2\pi\mathbf{C}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^T \mathbf{C}^{-1}(\mathbf{x} - \mathbf{H}\boldsymbol{\theta})\right). \tag{8.7}$$

The corresponding logarithm is

$$\ln p(\mathbf{x}; \boldsymbol{\theta}) = -\frac{1}{2}\ln|2\pi\mathbf{C}| - \frac{1}{2}(\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^T \mathbf{C}^{-1}(\mathbf{x} - \mathbf{H}\boldsymbol{\theta}), \tag{8.8}$$

which after expanding the second term becomes

$$\ln p(\mathbf{x}; \boldsymbol{\theta}) = -\frac{1}{2}\ln|2\pi\mathbf{C}| - \frac{1}{2}\left(\mathbf{x}^T\mathbf{C}^{-1}\mathbf{x} - \mathbf{x}^T\mathbf{C}^{-1}\mathbf{H}\boldsymbol{\theta} - \boldsymbol{\theta}^T\mathbf{H}^T\mathbf{C}^{-1}\mathbf{x} + \boldsymbol{\theta}^T\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H}\boldsymbol{\theta}\right). \tag{8.9}$$

The gradient of this function is

$$\frac{\partial}{\partial \boldsymbol{\theta}} \ln p(\mathbf{x}; \boldsymbol{\theta}) = \mathbf{H}^T\mathbf{C}^{-1}\mathbf{x} - \mathbf{H}^T\mathbf{C}^{-1}\mathbf{H}\boldsymbol{\theta}. \tag{8.10}$$

If $\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H}$ is invertible, (8.10) can be rewritten as

$$\frac{\partial}{\partial \boldsymbol{\theta}} \ln p(\mathbf{x}; \boldsymbol{\theta}) = \mathbf{H}^T\mathbf{C}^{-1}\mathbf{H}\left(\left(\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{C}^{-1}\mathbf{x} - \boldsymbol{\theta}\right). \tag{8.11}$$

By comparing (8.11) with (8.5), we recognize the Fisher information matrix

$$\mathbf{I}(\boldsymbol{\theta}) = \mathbf{H}^T\mathbf{C}^{-1}\mathbf{H} \tag{8.12}$$

and the estimator

$$g(\mathbf{x}) = \left(\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{C}^{-1}\mathbf{x}, \tag{8.13}$$

which is an efficient estimator. In the Chapter 7, we also stated that if an efficient estimator exits, it is the MLE, which is easily verified by equating (8.10) with zero and solving for $\boldsymbol{\theta}$.

8.4 Linear models in additive white Gaussian noise

So far, we have assumed that our noise is colored, i.e., that the noise samples are correlated. For the case of additive white Gaussian noise we can reuse the above result. The individual noise samples are IID with zero mean and variance σ^2. In this case, the covariance matrix $\mathbf{C}$ is a diagonal matrix with elements σ^2 on the main diagonal, i.e.,

$$\mathbf{C} = \sigma^2 \mathbf{I}. \tag{8.14}$$

Substituting (8.14) in (8.11) yields

$$\frac{\partial}{\partial \boldsymbol{\theta}} \ln p(\mathbf{x};\boldsymbol{\theta})(\mathbf{x};\boldsymbol{\theta}) = \frac{\mathbf{H}^T\mathbf{H}}{\sigma^2}\left(\left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{x} - \boldsymbol{\theta}\right), \tag{8.15}$$

where we recognize the estimator

$$g(\mathbf{x}) = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{x}, \tag{8.16}$$

and the Fisher information matrix

$$\mathbf{I}(\boldsymbol{\theta}) = \frac{\mathbf{H}^T\mathbf{H}}{\sigma^2}. \tag{8.17}$$

Example 8.2

Assume we have a signal composed of P sinusoidal signals embedded in white Gaussian noise with zero mean and variance σ^2, i.e.,

$$x_n = \sum_{p=0}^{P-1} a_p \cos\left(2\pi \frac{k_p}{N} n\right) + b_p \sin\left(2\pi \frac{k_p}{N} n\right) + w_n, \tag{8.18}$$

where a_p and b_p are the unknown amplitude and phase, respectively. The frequency parameters k_0 to $k_{(P-1)}$ are assumed to be known, different, and between 0 and $N-1$. The signal model in (8.18) for multiple observation can be expressed as

$$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}. \tag{8.19}$$

The observation matrix $\mathbf{H}$ is

$$\mathbf{H} = \begin{bmatrix} \mathbf{c}_{k_0} & \dots & \mathbf{c}_{k_{(P-1)}} & \mathbf{s}_{k_0} & \dots & \mathbf{s}_{k_{(P-1)}} \end{bmatrix} \tag{8.20}$$

with columns

$$\mathbf{c}_{k_p} = \begin{bmatrix} \cos\left(2\pi \frac{k_p}{N} 0\right) & \cos\left(2\pi \frac{k_p}{N} 1\right) & \dots & \cos\left(2\pi \frac{k_p}{N}(N-1)\right) \end{bmatrix}^T \tag{8.21}$$

and

$$\mathbf{s}_{k_p} = \begin{bmatrix} \sin\left(2\pi \frac{k_p}{N} 0\right) & \sin\left(2\pi \frac{k_p}{N} 1\right) & \dots & \sin\left(2\pi \frac{k_p}{N}(N-1)\right) \end{bmatrix}^T, \tag{8.22}$$

respectively. The parameter vector $\boldsymbol{\theta}$ for this model is

$$\boldsymbol{\theta} = \begin{bmatrix} a_0 & \dots & a_{(P-1)} & b_0 \dots b_{(P-1)} \end{bmatrix} \tag{8.23}$$

Note that the frequencies k_p are multiple of the fundamental frequency $1/N$, and thus, the columns of the observation matrix are orthogonal. The matrix $\mathbf{H}^T\mathbf{H}$ is

$$\mathbf{H}^T\mathbf{H} = \frac{N}{2}\mathbf{I}, \tag{8.24}$$

i.e., a diagonal matrix with entries $N/2$ on the main diagonal. The efficient estimator is

$$\hat{\boldsymbol{\theta}}_{\mathrm{ML}} = \left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^Y\mathbf{x} = \begin{bmatrix} \frac{2}{N}\mathbf{c}_{k_0}^T \\ \vdots \\ \frac{2}{N}\mathbf{c}_{k_{(P-1)}}^T \\ \frac{2}{N}\mathbf{s}_{k_0}^T \\ \vdots \\ \frac{2}{N}\mathbf{s}_{k_{(P-1)}}^T. \end{bmatrix} \mathbf{x} \tag{8.25}$$

The estimate of the parameter $\hat{a}_p$ and $\hat{b}_p$ are

$$\hat{a}_p = \frac{2}{N}\sum_{n=0}^{N-1} x[n]\cos\left(2\pi\frac{k_p}{N}n\right) \tag{8.26}$$

and

$$\hat{b}_p = \frac{2}{N}\sum_{n=0}^{N-1} x[n]\sin\left(2\pi\frac{k_p}{N}n\right). \tag{8.27}$$

9

Least-squares estimation

9.1 Introduction

So far, we have assumed that we have knowledge about the statistic of our observations $\mathbf{x}$. In many applications, however, this statistics is not available. For such scenarios, we can use the least-squares estimator (LSE).

9.2 Least-squares error criterion

As the name suggest, the least-squares approach aims to minimize the squared difference between the observed data and the signal model. Thus, the least-squares error criterion is defined as

$$J(\theta) = \sum_{n=0}^{N-1} e_n^2(\theta), \tag{9.1}$$

where the error between the signal model and the observed data is

$$e_n = x_n - s_n(\theta). \tag{9.2}$$

The least-squares estimate $\hat{\boldsymbol{\theta}}_{\mathrm{LS}}$ is the value of the parameter θ that minimizes (9.1), i.e.,

$$\hat{\boldsymbol{\theta}}_{\mathrm{LS}} = \arg\min_{\theta} J(\theta). \tag{9.3}$$

Example 9.1

Suppose we want to estimate the energy consumption per km of an electric car. For this purpose, we record at each recharge of the car the amount of energy charged E and the distance traveled D. The recordings are visualized in Figure 9.1. From the figure we can see that the energy consumption increases almost linearly with the distance traveled. Thus, we can assume a linear relation between the energy consumption and the distance traveled:

$$s_n(\theta) = \theta D_n. \tag{9.4}$$

In this model, the parameter θ represents the slope of the line. From the figure, however, we can also see that the data points do not exactly lie on a line. The discrepancy between

the model and data comes from the model inaccuracy: the energy consumption depends on more parameters than just the distance. Nonetheless, the linear model is a good approximation.

In this example, the least-squares error criterion becomes

$$J(\theta) = \sum_{n=0}^{N-1} (x_n - \theta D_n)^2. \tag{9.5}$$

The least-squares error criterion is minimized by taking the derivative with respect to θ and equating it to zero, which yields

$$\hat{\boldsymbol{\theta}}_{\mathrm{LS}} = \frac{\sum_{n=0}^{N-1} x_n D_n}{\sum_{n=0}^{N-1} D_n^2}. \tag{9.6}$$

The estimated line is indicated by the dashed line in Figure 9.1.

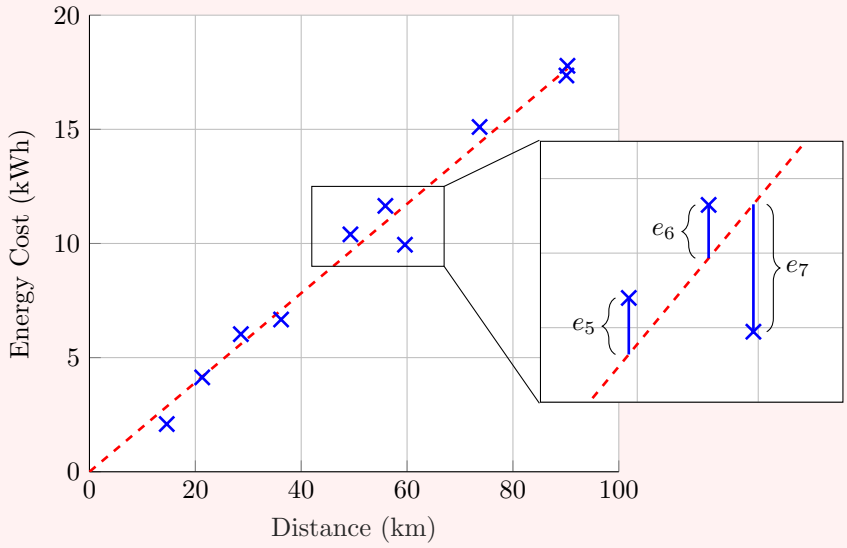


Figure 9.1: Recorded energy consumption vs. distance traveled. The dashed line indicates the slope of the linear model that minimizes the sum of the squared errors.

In the example above, we were only interested in estimating a single parameter θ. In general, the signal model can be dependent on several parameters. In this case we write $s[n;\boldsymbol{\theta}]$. In the following, we will consider vector parameters since the scalar case arise as a special case of the vector parameters.

9.3 Linear least-squares estimator

A linear signal model, as shown in Chapter 8, is any model of the form

$$\mathbf{s}(\boldsymbol{\theta}) = \mathbf{H}\boldsymbol{\theta} = \sum_{k=0}^{K-1} \theta_k \mathbf{h}_k \tag{9.7}$$

where $\mathbf{s}$ is the signal vector, $\mathbf{H}$ is the so called *observation matrix*, $\mathbf{h}_k$ is the kth column of $\mathbf{H}$, and $\boldsymbol{\theta}$ is the parameter vector. To incorporate model inaccuracies or measurement noise, we add a noise term $\mathbf{w}$ to model the observed data, i.e.,

$$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}. \tag{9.8}$$

Using the matrix notation in (9.7), the least-squares error criterion in (9.1) can be expressed as

$$\begin{aligned} J(\boldsymbol{\theta}) &= \|\mathbf{x} - \mathbf{H}\boldsymbol{\theta}\|^2 \\ &= (\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^T(\mathbf{x} - \mathbf{H}\boldsymbol{\theta}) \\ &= (\mathbf{x}^T - \boldsymbol{\theta}^T\mathbf{H}^\mathbf{T})(\mathbf{x} - \mathbf{H}\boldsymbol{\theta}) \\ &= \mathbf{x}^T\mathbf{x} - 2\mathbf{x}^T\mathbf{H}\boldsymbol{\theta} + \boldsymbol{\theta}^T\mathbf{H}^T\mathbf{H}\boldsymbol{\theta}. \end{aligned} \tag{9.9}$$

The corresponding gradient is

$$\frac{\partial}{\partial \boldsymbol{\theta}} J(\boldsymbol{\theta}) = -2\mathbf{H}^T\mathbf{x} + 2\mathbf{H}^T\mathbf{H}\boldsymbol{\theta}. \tag{9.10}$$

The minimum can be found by setting the gradient to zero and solving for $\boldsymbol{\theta}$. The linear least-squares estimator is

$$\hat{\boldsymbol{\theta}}_{\text{LSE}} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{x}, \tag{9.11}$$

which is referred to as *normal equation*.

Returning to the electric car example in Example 9.1, we have the special case of a single parameter θ. In this case, the observation matrix becomes a vector with the distances D_n as its entries. The expression $\mathbf{H}^T\mathbf{H} = \sum_{n=0}^{N-1} D_n^2$. Note that $\sum_{n=0}^{N-1} D_n^2$ is a scalar and its inverse is $1/(\sum_{n=0}^{N-1} D_n^2)$. On the other hand, the expression $\mathbf{H}^T\mathbf{x} = \sum_{n=0}^{N-1} x_n D_n$. Combining these two expressions, we obtain the least-squares estimate in (9.6).

9.4 Geometric interpretation

The previous derivation is based on minimizing the squared error term in (9.1). The *linear* least-squares estimation can also be derived based on a geometrical interpretation. To obtain this geometrical interpretation, we assume that more observations than parameters to be estimated are available, i.e., $N > K$. In this case, the columns of $\mathbf{H}$ span a K dimensional subspace in the N dimensional space. The vector $\mathbf{s}$ is a linear combination of the columns of $\mathbf{H}$, see (9.7), and as such, it lies in this subspace.

From the geometric point of view, the least-squares error criterion in (9.9) describes the length of the error vector $\mathbf{e} = \mathbf{x} - \mathbf{H}\boldsymbol{\theta}$. The LSE $\hat{\boldsymbol{\theta}}_{\text{LS}}$ is the set of coefficients that minimizes the length of the error vector, which is shown in Figure 9.2.

The length of the error vector is minimized if it is orthogonal to the subspace, which is known as the *orthogonality principle*. To see that this is indeed the case, consider the projection matrix

$$\mathbf{P} = \mathbf{H}(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T, \tag{9.12}$$

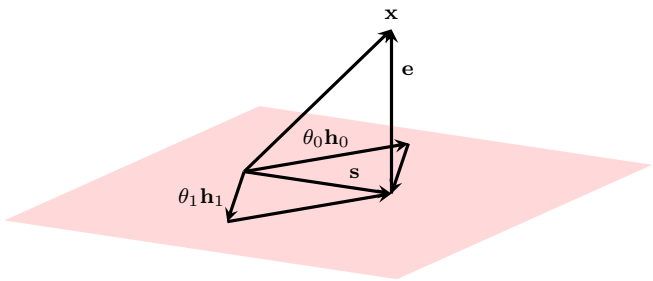


Figure 9.2: The length of the error vector becomes a minimum if it is orthogonal to the subspace spanned by the columns of $\mathbf{H}$.

which projects a vector $\mathbf{x} \in \mathrm{R}^N$ onto the subspace spanned by the columns of $\mathbf{H}$. Using the projection matrix $\mathbf{P}$, the squared norm or the error vector can be expressed as

$$\begin{aligned} \|\mathbf{x}-\mathbf{s}\|^2 &= \|\mathbf{x}-\mathbf{P}\mathbf{x}+\mathbf{P}\mathbf{x}-\mathbf{s}\|^2 \\ &= \|\mathbf{x}-\mathbf{P}\mathbf{x}\|^2 + \|\mathbf{P}\mathbf{x}-\mathbf{s}\|^2 - 2(\mathbf{x}-\mathbf{P}\mathbf{x})^T(\mathbf{P}\mathbf{x}-\mathbf{s}). \end{aligned} \tag{9.13}$$

Note that the vector $\mathbf{P}\mathbf{x}$ and $\mathbf{s}$ are lying in the column space of $\mathbf{H}$. Thus, also the difference $\mathbf{P}\mathbf{x}-\mathbf{s}$ lies in the column space. The vector $\mathbf{x}-\mathbf{P}\mathbf{x}$, on the other hand, is orthogonal to the column space. Consequently, the last term in (9.13) is zero, and thus the squared norm in (9.13) is minimized for the choice

$$\begin{aligned} \mathbf{s} &= \mathbf{P}\mathbf{x} \\ &= \mathbf{H}(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{x}. \end{aligned} \tag{9.14}$$

Comparing (9.7) and (9.14), we see that

$$\hat{\boldsymbol{\theta}}_{\mathrm{LS}} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{x}. \tag{9.15}$$

Note that the *linear* LSE is also a linear estimate since it is formed by a linear combination of our observations $\mathbf{x}$.

9.5 Weighted least-squares estimation

For some estimation problems, we might want to reduce the influence of a portion of the data on our final estimate. For example, the data can be provided by different sensors. Moreover, some sensors may have a higher accuracy than others, and thus, we have more confidence in their measurement result. However, even though the other sensor measurement results are less accurate, they still provide some information. The different confidence in the measurement result can be incorporated in least-squares estimator by assigning them different weights, which leads to the weighted least-squares estimator (WLSE). In general, we can express the weighted least-squares error criterion as

$$J(\boldsymbol{\theta}) = (\mathbf{x}-\mathbf{H}\boldsymbol{\theta})^T\mathbf{W}(\mathbf{x}-\mathbf{H}\boldsymbol{\theta}), \tag{9.16}$$

where $\mathbf{W}$ is a *positive definite* matrix. For the particular case that $\mathbf{W}$ is diagonal, we obtain

$$J(\boldsymbol{\theta}) = \sum_{n=0}^{N-1} [\mathbf{W}]_{n,n}(x_n - s_n(\boldsymbol{\theta}))^2, \tag{9.17}$$

where $[\mathbf{W}]_{n,n}$ is the nth diagonal element of $\mathbf{W}$. The weighted LSE is obtained by setting the gradient to zero, which yields

$$\hat{\boldsymbol{\theta}}_{\mathrm{WLS}} = (\mathbf{H}^T\mathbf{W}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{W}\mathbf{x}. \tag{9.18}$$

9.6 Best linear unbiased estimator

In the above discussion, we have not considered the statistics of the observations. Thus, we were also not able to verify the performance of the LSE. On the other hand, if we are dealing with a signal model (9.8) and have additional knowledge about the mean of the observation which is given as

$$\mathrm{E}[\mathbf{x}] = \mathbf{H}\boldsymbol{\theta} \tag{9.19}$$

and the covariance matrix $\mathbf{C}_\mathbf{x}$ of the observation, we can find an *unbiased* estimator that has the lowest variance within the class of all *linear* estimators. Therefore, consider first that the *weighted* LSE is a linear estimator since the estimate is formed by a linear combination of the observations $\mathbf{x}$. Secondly, under the assumption (9.19), we have that the expectation of the *weighted* LSE is

$$\begin{aligned}\mathrm{E}[\hat{\boldsymbol{\theta}}_{\mathrm{LS}}] &= \mathrm{E}[(\mathbf{H}^T\mathbf{W}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{W}\mathbf{x}] \\ &= (\mathbf{H}^T\mathbf{W}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{W}\,\mathrm{E}[\mathbf{x}] \\ &= (\mathbf{H}^T\mathbf{W}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{W}\mathbf{H}\boldsymbol{\theta} \\ &= \boldsymbol{\theta},\end{aligned} \tag{9.20}$$

i.e., the estimator is unbiased. The covariance matrix of the WLSE is

$$\begin{aligned}\mathbf{C}_{\hat{\boldsymbol{\theta}}_{\mathrm{WLS}}} &= \mathrm{E}[(\hat{\boldsymbol{\theta}}_{\mathrm{WLS}} - \mathrm{E}[\hat{\boldsymbol{\theta}}_{\mathrm{WLS}}])(\hat{\boldsymbol{\theta}}_{\mathrm{WLS}} - \mathrm{E}[\hat{\boldsymbol{\theta}}_{\mathrm{WLS}}])^T] \\ &= \mathrm{E}[(\mathbf{H}^T\mathbf{W}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{W}(\mathbf{x} - \mathrm{E}[\mathbf{x}])(\mathbf{H}^T\mathbf{W}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{W}(\mathbf{x} - \mathrm{E}[\mathbf{x}])^T] \\ &= (\mathbf{H}^T\mathbf{W}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{W}\,\mathrm{E}[(\mathbf{x} - \mathrm{E}[\mathbf{x}](\mathbf{x} - \mathrm{E}[\mathbf{x}])^T])\mathbf{H}^T\mathbf{W}(\mathbf{H}^T\mathbf{W}\mathbf{H})^{-1} \\ &= (\mathbf{H}^T\mathbf{W}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{W}\mathbf{C}_\mathbf{x}\mathbf{W}\mathbf{H}^T(\mathbf{H}^T\mathbf{W}\mathbf{H})^{-1}.\end{aligned} \tag{9.21}$$

In the last step, we made use of the symmetry of $(\mathbf{H}^T\mathbf{W}\mathbf{H})^{-1}$ and $\mathbf{W}$.

From (9.21), we can see that the variance of the WLSE depends on the weighting matrix $\mathbf{W}$ and the covariance matrix of the observations $\mathbf{C}_\mathbf{x}$. This means also that we can influence the variance of the estimator by the choice of the weighting matrix $\mathbf{W}$. If we choose the $\mathbf{W} = \mathbf{C}_\mathbf{x}^{-1}$ then the variance of the WLSE becomes

$$(\mathbf{H}^T\mathbf{C}_\mathbf{x}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{C}_\mathbf{x}^{-1}\mathbf{C}_\mathbf{x}\mathbf{C}_\mathbf{x}^{-1}\mathbf{H}^T(\mathbf{H}^T\mathbf{C}_\mathbf{x}\mathbf{H})^{-1} = (\mathbf{H}^T\mathbf{C}_\mathbf{x}^{-1}\mathbf{H})^{-1}. \tag{9.22}$$

Furthermore, the choice $\mathbf{W} = \mathbf{C}_\mathbf{x}^{-1}$ also minimizes the variance under all possible choices of linear estimators. Therefore, the weighted least-squares with $\mathbf{W} = \mathbf{C}_\mathbf{x}^{-1}$ is the *best linear unbiased estimator*, or short BLUE.

To proof the above statement, we first introduce the matrix

$$\mathbf{A} = (\mathbf{H}^T\mathbf{C}_\mathbf{x}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{C}_\mathbf{x}^{-1}. \tag{9.23}$$

With the above notation, the WLSE and its covariance matrix become

$$\hat{\boldsymbol{\theta}}_{\mathrm{BLUE}} = \mathbf{A}\mathbf{x} \tag{9.24}$$

and

$$\mathbf{C}_{\hat{\boldsymbol{\theta}}_{\text{BLUE}}} = \mathbf{A}\mathbf{C}_{\mathbf{x}}\mathbf{A}^T, \tag{9.25}$$

respectively. Let us now introduce another linear estimator

$$\hat{\boldsymbol{\theta}}' = \mathbf{A}'\mathbf{x}, \tag{9.26}$$

and let us express the matrix $\mathbf{A}'$ as

$$\mathbf{A}' = \mathbf{A} + \mathbf{B}, \tag{9.27}$$

then the expected value of $\hat{\boldsymbol{\theta}}'_{\text{WLS}}$ is

$$\begin{aligned} \text{E}[\mathbf{A}'\mathbf{x}] &= (\mathbf{A} + \mathbf{B})\,\text{E}[\mathbf{x}] \\ &= (\mathbf{A} + \mathbf{B})\mathbf{H}\boldsymbol{\theta} \\ &= (\mathbf{I} + \mathbf{B}\mathbf{H})\boldsymbol{\theta}. \end{aligned} \tag{9.28}$$

For the estimate $\hat{\boldsymbol{\theta}}'_{\text{WLS}}$ to be unbiased, we require that $\mathbf{B}\mathbf{H} = \mathbf{0}$, where $\mathbf{0}$ is the null matrix. The covariance matrix of $\hat{\boldsymbol{\theta}}'_{\text{WLS}}$ is

$$\begin{aligned} \mathbf{C}_{\hat{\boldsymbol{\theta}}'_{\text{WLS}}} &= \mathbf{A}'\mathbf{C}_{\mathbf{x}}\mathbf{A}'^T \\ &= (\mathbf{A} + \mathbf{B})\mathbf{C}_{\mathbf{x}}(\mathbf{A} + \mathbf{B})^T \\ &= \mathbf{A}\mathbf{C}_{\mathbf{x}}\mathbf{A}^T + \mathbf{B}\mathbf{C}_{\mathbf{x}}\mathbf{A}^T + \mathbf{A}\mathbf{C}_{\mathbf{x}}\mathbf{B}^T + \mathbf{B}\mathbf{C}_{\mathbf{x}}\mathbf{B}^T. \end{aligned} \tag{9.29}$$

Inspecting the inner two terms and substituting (9.23), we have

$$\mathbf{B}\mathbf{C}_{\mathbf{x}}\mathbf{A}^T = \mathbf{B}\mathbf{C}_{\mathbf{x}}\mathbf{C}_{\mathbf{x}}^{-1}\mathbf{H}(\mathbf{H}^T\mathbf{C}_{\mathbf{x}}^{-1}\mathbf{H})^{-1} = \mathbf{B}\mathbf{H}(\mathbf{H}^T\mathbf{C}_{\mathbf{x}}^{-1}\mathbf{H})^{-1} = \mathbf{0} \tag{9.30}$$

and

$$\mathbf{A}\mathbf{C}_{\mathbf{x}}\mathbf{B}^T = (\mathbf{H}^T\mathbf{C}_{\mathbf{x}}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{C}_{\mathbf{x}}^{-1}\mathbf{C}_{\mathbf{x}}\mathbf{B}^T = (\mathbf{H}^T\mathbf{C}_{\mathbf{x}}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{B}^T = \mathbf{0}, \tag{9.31}$$

which follows from the fact that $\mathbf{C}_{\mathbf{x}}^{-1} = \left(\mathbf{C}_{\mathbf{x}}^{-1}\right)^T$, $(\mathbf{H}^T\mathbf{C}_{\mathbf{x}}^{-1}\mathbf{H})^{-1} = \left((\mathbf{H}^T\mathbf{C}_{\mathbf{x}}^{-1}\mathbf{H})^{-1}\right)^T$, and the unbiasedness constraint $\mathbf{B}\mathbf{H} = \mathbf{0}$. Using the above results, the covariance matrix reduces to

$$\mathbf{C}_{\hat{\boldsymbol{\theta}}'_{\text{WLS}}} = \mathbf{A}\mathbf{C}_{\mathbf{x}}\mathbf{A}^T + \mathbf{B}\mathbf{C}_{\mathbf{x}}\mathbf{B}^{\cdot} \tag{9.32}$$

The difference between $\mathbf{C}_{\hat{\boldsymbol{\theta}}'_{\text{WLS}}}$ and $\mathbf{C}_{\hat{\boldsymbol{\theta}}_{\text{BLUE}}}$is $\mathbf{B}\mathbf{C}_{\mathbf{x}}\mathbf{B}^T$, which is positive semidefinite, meaning that the diagonal elements that are the differences between the variances of $\hat{\boldsymbol{\theta}}_{\text{WLS}}$ and $\hat{\boldsymbol{\theta}}_{\text{BLUE}}$, are greater than or equal to zero. This concludes the proof.

9.7 Nonlinear least-squares estimator - transformation of parameters

Until now, we dealt with linear signal models, i.e., functions that are linear in the parameter vector $\boldsymbol{\theta}$. For this particular case, we were able to derive a closed form expression for the least-squares estimate based on the observation matrix $\mathbf{H}$. In general, the signal model can also be nonlinear in the parameter vector $\boldsymbol{\theta}$. For most nonlinear problems, we rely on numerical methods.

However, optimization problems have the property that they can be carried out in a transformed space that is obtained by a one-to-one mapping. This property can be used to produce a linear signal model from a nonlinear signal model. Therefore, let

$$\boldsymbol{\alpha} = g(\boldsymbol{\theta}) \tag{9.33}$$

be a function whose inverse exists. If we can find a function $g(\boldsymbol{\theta})$ such that

$$\mathbf{s}(\boldsymbol{\theta}(\boldsymbol{\alpha})) = \mathbf{s}(g^{-1}(\boldsymbol{\alpha})) = \mathbf{H}\boldsymbol{\alpha}, \tag{9.34}$$

then the signal model will be linear in $\boldsymbol{\alpha}$. We can use the linear LS formulations we derived so far and find the parameter θ through the inverse transform by

$$\hat{\boldsymbol{\theta}} = g^{-1}(\hat{\boldsymbol{\alpha}}), \tag{9.35}$$

where

$$\boldsymbol{\alpha} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{x}. \tag{9.36}$$

10

Bayesian estimation

10.1 Introduction

So far, we viewed the parameters $\boldsymbol{\theta}$ as a deterministic but unknown quantity. This point of view changes when considering Bayesian estimators. In Bayesian estimation, the unknown parameters are considered random variables and have a defined PDF $p(\boldsymbol{\theta})$. Considering the parameters to be random variables allows incorporating prior knowledge in the estimator. For example, some physical quantities can only be in a specific range, which can be modeled by a uniform prior of the parameter. The prior PDF is either obtained by physical modeling or empirically by collecting many observations.

Additionally to the prior probability, we introduce a cost or loss function. The cost function penalizes the estimation error and models which errors should be avoided. We will introduce three different cost functions resulting in three different estimates. Out of these three cost functions, two are of particular interest and lead to the minimum means square error (MMSE) estimate and the maximum a posteriori (MAP) estimate.

10.2 Cost function and Bayes risk

For the classical estimators, we evaluated the performance by their mean and variance. This approach, however, does not apply if we consider the parameters to be random variables. Instead, we are interested in assigning a cost to the estimation error between all pairs $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}$. Therefore, we define the estimation error as

$$\mathbf{e} = \hat{\boldsymbol{\theta}} - \boldsymbol{\theta} \tag{10.1}$$

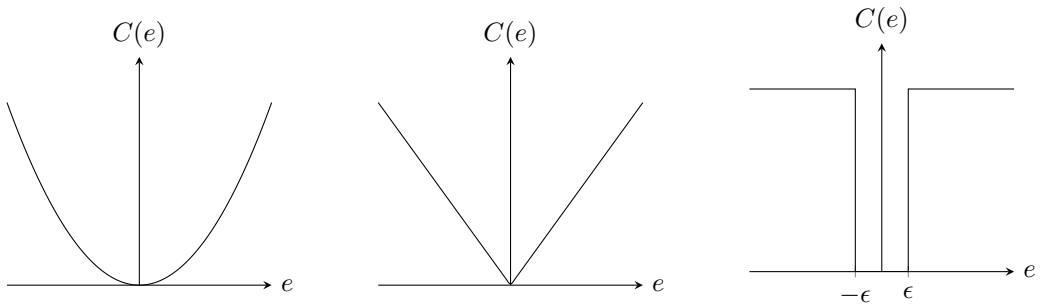
and denote the cost of the error by $C(\mathbf{e})$. The most commonly used costs are

1. Square error:

$$C(\mathbf{e}) = \|\mathbf{e}\|_2^2 = \sum_{k=0}^{K-1} e_k^2. \tag{10.2}$$

2. Absolute error:

$$C(\mathbf{e}) = \|\mathbf{e}\|_1 = \sum_{k=0}^{K-1} |e_k|. \tag{10.3}$$



(a) Quadratic error (b) Absolute error (c) Uniform error

Figure 10.1: Examples of the different cost function: (a) squared error, (b) absolute error, and (c) uniform cost.

3. Uniform (notch or hit-or-miss) error:

$$C(\mathbf{e}) = \begin{cases} 0 & \|\mathbf{e}\|_\infty = \max\limits_{0\le k\le K-1} |e_k| < \epsilon, \\ 1 & \text{otherwise.} \end{cases} \tag{10.4}$$

All three cost functions are illustrated in Figure 10.1 for the case of $K = 1$. The square error cost increases the cost with the square of the error vector's magnitude, thus penalizing larger errors more severely. The absolute error cost function, on the other hand, penalizes the error according to its magnitude. The uniform cost function assigns the same penalty to all errors larger than a certain threshold.

After specifying the cost function, we can compute the average cost, also called the Bayes risk, given by

$$\mathrm{E}[C(\mathbf{e})] = \int\int C(\mathbf{e}) p(\boldsymbol{\theta}, \mathbf{x}) d\boldsymbol{\theta} d\mathbf{x}. \tag{10.5}$$

Note that the expectation is taken with respect to the parameters $\boldsymbol{\theta}$ and the observations $\mathbf{x}$. Expressing $p(\boldsymbol{\theta}, \mathbf{x}) = p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})$ we can express (10.5) equivalently as

$$\mathrm{E}[C(\mathbf{e})] = \int\int C(\mathbf{e}) p(\boldsymbol{\theta}|\mathbf{x}) d\boldsymbol{\theta} p(\mathbf{x}) d\mathbf{x}. \tag{10.6}$$

The Bayesian estimate is the estimate that minimizes the Bayes risk, i.e.,

$$\hat{\boldsymbol{\theta}} = \arg\min_{\hat{\boldsymbol{\theta}}} \int\int C(\mathbf{e}) p(\boldsymbol{\theta}|\mathbf{x}) d\boldsymbol{\theta} p(\mathbf{x}) d\mathbf{x}. \tag{10.7}$$

Since $p(\mathbf{x}) \ge 0$, it follows that the Bayes risk is minimized if the inner integral is minimized, which is shown next for the three cost discussed cost functions.

10.3 Minimum mean square error

To minimize the mean square error we set the gradient of the inner integral

$$\frac{\partial}{\partial \hat{\boldsymbol{\theta}}} \int \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2 p(\boldsymbol{\theta}|\mathbf{x}) d\boldsymbol{\theta} = 2 \int (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{x}) d\boldsymbol{\theta} \tag{10.8}$$

to zero. Solving for $\hat{\boldsymbol{\theta}}$ yields

$$\hat{\boldsymbol{\theta}}_{\text{MMSE}} = \int \boldsymbol{\theta} p(\boldsymbol{\theta}|\mathbf{x}) d\boldsymbol{\theta}, \tag{10.9}$$

which is the conditional mean of $\boldsymbol{\theta}$ given the observation $\mathbf{x}$.

10.4 Minimum absolute error

To find the Bayes estimator for the minimum absolute error (MAE) cost function, we note that the kth component of the gradient is

$$\frac{\partial}{\partial \theta_k} \int \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_1 p(\boldsymbol{\theta}|\mathbf{x}) d\boldsymbol{\theta} = \int \text{sign}(\hat{\theta}_k - \theta_k) p(\theta_k|\mathbf{x}) d\theta_k, \tag{10.10}$$

where the sign function is defined as

$$\text{sign}(x) = \begin{cases} 1 & x \geq 0, \\ -1 & x < 0. \end{cases} \tag{10.11}$$

We can split the integral in (10.10) into two parts as follows:

$$\int \text{sign}(\hat{\theta}_k - \theta_k) p(\theta_k|\mathbf{x}) d\theta_k = \int_{-\infty}^{\hat{\theta}_k} p(\theta_k|\mathbf{x}) d\theta_k - \int_{\hat{\theta}_k}^{\infty} p(\theta_k|\mathbf{x}) d\theta_k. \tag{10.12}$$

The integral becomes zero if both integrals become $1/2$, or equivalently if we set $\hat{\theta}$ to the *median* of the posterior PDF $p(\theta_k|\mathbf{x})$. Thus, the minimum absolute error estimate $\hat{\boldsymbol{\theta}}_{\text{MAE}}$ is the *median* of the posterior distribution $p(\boldsymbol{\theta}|\mathbf{x})$.

10.5 Uniform error

To find the Bayes estimate for the uniform error criterion, we first rewrite the integral we wish to minimize. Since the uniform error criterion is 1 for all errors $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_\infty \geq \epsilon$ and since $\int p(\boldsymbol{\theta}|\mathbf{x}) d\boldsymbol{\theta} = 1$, we simply subtract the part of the integral for which the uniform error is 0, i.e.,

$$\int C(\mathbf{e}) p(\boldsymbol{\theta}|\mathbf{x}) d\boldsymbol{\theta} = 1 - \int_{\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_\infty < \epsilon} p(\boldsymbol{\theta}|\mathbf{x}) d\boldsymbol{\theta}. \tag{10.13}$$

The integral is minimized if the expression

$$\int_{\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_\infty < \epsilon} p(\boldsymbol{\theta}|\mathbf{x}) d\boldsymbol{\theta} \tag{10.14}$$

is maximized. For a small ϵ, the integral is maximized if $\hat{\boldsymbol{\theta}}$ is chosen as the maximum of the posterior PDF $p(\boldsymbol{\theta}|\mathbf{x})$. Therefore, the uniform error criterion leads to the *maximum aposterior* (MAP) estimator, which is

$$\hat{\boldsymbol{\theta}}_{\text{MAP}} = \arg\max_{\boldsymbol{\theta} \in \mathbb{R}^K} p(\boldsymbol{\theta}|\mathbf{x}). \tag{10.15}$$

Example 10.1

We wish to estimate a quantity which has a Gaussian prior probability and the measurement itself is corrupted by an additive Gaussian measurement error. Such an estimation problem is given, for example, when we assume the voltage in Example 5.1 to have a Gaussian prior probability.

To find the Bayesian estimates of the discussed cost functions, we need to find either the *mean*, the *median*, or the *maximum* of the posterior distribution $p(\theta|\mathbf{x})$. To find the posterior distribution $p(\theta|\mathbf{x})$, we use Bayes' theorem and express the posterior distribution by means of the prior probability and the likelihood, i.e.,

$$p(\theta|\mathbf{x}) = \frac{p(\mathbf{x}|\theta)p(\theta)}{p(\mathbf{x})} = \frac{p(\mathbf{x}|\theta)p(\theta)}{\int p(\mathbf{x}|\theta)p(\theta)d\theta}. \tag{10.16}$$

The prior probability is given as

$$p(\theta) = \frac{1}{\sqrt{2\pi\sigma_\theta^2}} \exp\left(-\frac{1}{2\sigma_\theta^2}(\theta - \mu_\theta)^2\right), \tag{10.17}$$

and the likelihood is

$$p(\mathbf{x}|\theta) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left(-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n] - \theta)^2\right). \tag{10.18}$$

Both the prior and the likelihood are Gaussian distributed with means μ_θ and θ, and variances σ_θ^2 and σ^2, respectively. The product of the prior and the likelihood distribution is again Gaussian. This can be shown by defining $\bar{x} = \frac{1}{N}\sum_{n=0}^{N-1} x[n]$ and noting that the denominator in (10.16) is independent of the parameter θ. The numerator is then

$$\begin{aligned} p(\theta|\mathbf{x}) &= \frac{\exp\left(-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}x^2[n] - \frac{1}{2\sigma^2}\left(N\theta^2 - 2\theta N\bar{x}\right) - \frac{1}{2\sigma_\theta^2}(\theta - \mu_\theta)^2\right)}{\int \exp\left(-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}x^2[n] - \frac{1}{2\sigma^2}\left(N\theta^2 - 2\theta N\bar{x}\right) - \frac{1}{2\sigma_\theta^2}(\theta - \mu_\theta)^2\right)d\theta} \\ &= \frac{\exp\left(-\frac{1}{2}\left(\frac{1}{\sigma^2}\left(N\theta^2 - 2\theta N\bar{x}\right) + \frac{1}{\sigma_\theta^2}(\theta - \mu_\theta)^2\right)\right)}{\int \exp\left(-\frac{1}{2}\left(\frac{1}{\sigma^2}\left(N\theta^2 - 2\theta N\bar{x}\right) + \frac{1}{\sigma_\theta^2}(\theta - \mu_\theta)^2\right)\right)d\theta} \\ &= \frac{\exp\left(-\frac{1}{2}Q(\theta)\right)}{\int \exp\left(-\frac{1}{2}Q(\theta)\right)d\theta}, \end{aligned} \tag{10.19}$$

where

$$Q(\theta) = \frac{1}{\sigma^2}\left(N\theta^2 - 2\theta N\bar{x}\right) + \frac{1}{\sigma_\theta^2}(\theta - \mu_\theta)^2. \tag{10.20}$$

The denominator is a normalization constant and independent of θ. To show that $p(\theta|\mathbf{x})$ is a Gaussian distribution, we need to show that the numerator has a quadratic exponential form with respect to the parameter θ. Therefore, we express $Q(\theta)$ as

$$Q(\theta) = \left(\frac{N}{\sigma^2} + \frac{1}{\sigma_\theta^2}\right)\theta^2 - 2\left(\frac{N\bar{x}}{\sigma^2} + \frac{\mu_\theta}{\sigma_\theta^2}\right)\theta + \frac{\mu_\theta^2}{\sigma_{\theta|\mathbf{x}}^2}, \tag{10.21}$$

where we define

$$\sigma^2_{\theta|\mathbf{x}} = \frac{1}{\frac{N}{\sigma^2} + \frac{1}{\sigma^2_\theta}} \tag{10.22}$$

and

$$\mu_{\theta|\mathbf{x}} = \left(\frac{N\bar{x}}{\sigma^2} + \frac{\mu_\theta}{\sigma^2_\theta}\right)\sigma^2_{\theta|\mathbf{x}}. \tag{10.23}$$

Now, completing the square we get

$$\begin{aligned} Q(\theta) &= \frac{1}{\sigma^2_{\theta|\mathbf{x}}}\left(\theta^2 - 2\mu_{\theta|\mathbf{x}}\theta\right) + \frac{\mu^2_\theta}{\sigma^2_{\theta|\mathbf{x}}} + \frac{\mu^2_{\theta|\mathbf{x}}}{\sigma^2_{\theta|\mathbf{x}}} - \frac{\mu^2_{\theta|\mathbf{x}}}{\sigma^2_{\theta|\mathbf{x}}} \\ &= \frac{1}{\sigma^2_{\theta|\mathbf{x}}}\left(\theta^2 - 2\mu_{\theta|\mathbf{x}}\theta + \mu^2_{\theta|\mathbf{x}}\right) + \frac{\mu^2_\theta}{\sigma^2_{\theta|\mathbf{x}}} - \frac{\mu^2_{\theta|\mathbf{x}}}{\sigma^2_{\theta|\mathbf{x}}} \\ &= \frac{1}{\sigma^2_{\theta|\mathbf{x}}}\left(\theta - \mu_{\theta|\mathbf{x}}\right)^2 + \frac{\mu^2_\theta}{\sigma^2_{\theta|\mathbf{x}}} - \frac{\mu^2_{\theta|\mathbf{x}}}{\sigma^2_{\theta|\mathbf{x}}}. \end{aligned} \tag{10.24}$$

Note that the last two terms on the right hand side are independent of θ. Therefore we have

$$\begin{aligned} p(\theta|\mathbf{x}) &= \frac{\exp\left(-\frac{1}{2\sigma^2_{\theta|\mathbf{x}}}\left(\theta - \mu_{\theta|\mathbf{x}}\right)^2\right)}{\int \exp\left(-\frac{1}{2\sigma^2_{\theta|\mathbf{x}}}\left(\theta - \mu_{\theta|\mathbf{x}}\right)^2\right) d\theta} \\ &= \frac{1}{\sqrt{2\pi\sigma^2_{\theta|\mathbf{x}}}} \exp\left(-\frac{1}{2\sigma^2_{\theta|\mathbf{x}}}\left(\theta - \mu_{\theta|\mathbf{x}}\right)^2\right). \end{aligned} \tag{10.25}$$

The posterior distribution is a Gaussian distribution with mean $\mu_{\theta|\mathbf{x}}$ and variance $\sigma^2_{\theta|\mathbf{x}}$. Hence, the mean, the median, and the maximum are the same. Note that this is in general not true for all posterior probabilities.

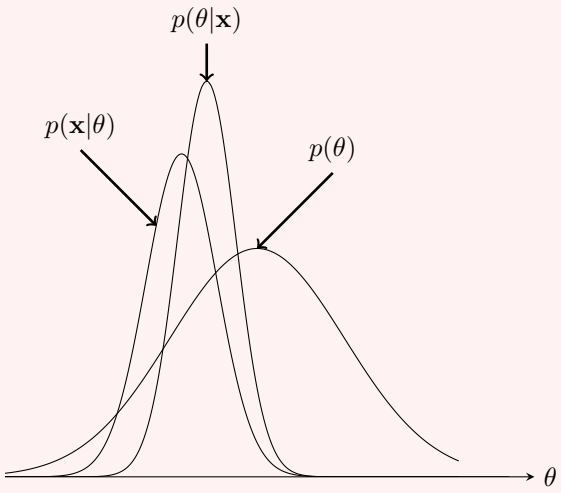


Figure 10.2: The posterior probability of Gaussian prior and a Gaussian likelihood is again Gaussian.

11

Numerical methods

11.1 Introduction

For each estimator we investigate in this course, we introduce the formal description and demonstrate the derivation of a closed-form estimator when the signal model and noise distribution are suitable. However, the closed-form solution may not be available in many situations. Moreover, the data might be following a time-dependent distribution, which precludes using a single signal model for a data batch. In this section, we will see several solution methods to cover such situations. It can be said that a significant portion of the signal processing research focuses on devising solutions to problems that deviate from the standard methods described in this course. Thus, this chapter is far from a comprehensive investigation of all such methods. The goal of this chapter is to give insight into the practical side of the estimation problems.

11.2 Grid search

Closed-form solutions for estimators like the MLE or the LSE only exist for particular cases such as linear signal models. For all other cases, we have to resort to other methods to find the estimate. For the maximum likelihood estimator, for example, we can evaluate the likelihood function numerically at equally spaced points for the parameter vector $\boldsymbol{\theta}$. If we choose the spacing between the points sufficiently small, we can find a value close to the real solution. This method is referred to as grid search. The application of the gird search, however, is restricted to simple problems. For example, if the range of possible values for the parameter is unbounded, the grid search becomes infeasible.

11.3 The Newton-Raphson method

The Newton-Raphson method is a method to approximate a function at a given point or to solve an equation. In maximum likelihood estimation, we want to solve the likelihood equation, and thus, we are interested in the latter case. The idea is to approximate the function's behavior at a particular point by a linear function. The derivative of the function determines the slope of the linear function at this point. Thus, the function is approximated as

$$f(x) \approx f(a) + f'(a)(x - a), \tag{11.1}$$

where $f'(a)$ denotes the derivative of $f(x)$ evaluated at point a. In the particular case of solving the equation $f(x) = 0$, we obtain

$$x \approx a - \frac{f(a)}{f'(a)}. \tag{11.2}$$

While the solution to the equation above is closer to the actual zero of the function, there is still room for improvement. The approximation can be utilized again. This time the first derivative is evaluated on the new estimate. The iterative formulation is

$$x_{m+1} = x_m - \frac{f(x_m)}{f'(x_m)}. \tag{11.3}$$

A solution is found if $x_{m+1} = x_m$. Note that, depending on the starting point, the method may converge to a local minimum. Hence, to find a global minimum, the method should be initialized at different values.

11.4 Newton-Raphson method for maximum likelihood estimation

We consider the Newton-Raphson method for the MLE, where it is used to numerically calculate the parameter value at which the derivative of the log-likelihood function is zero. Thus, we will be dealing with both the first and the second derivatives of the log-likelihood function. The iterative MLE for the parameter vector θ is given by

$$\theta_{m+1} = \theta_m - \left(\frac{\partial^2 \ln p(\mathbf{x}; \theta_m)}{\partial \theta^2} \right)^{-1} \frac{\partial \ln p(\mathbf{x}; \theta + m)}{\partial \theta} \tag{11.4}$$

11.5 Method of scoring

An extension to the Newton-Raphson method for the MLE is the method of scoring. The idea is to replace the second derivative of the log-likelihood function with its expected value, negative of which is the Fisher information matrix. Thus, the method of scoring yields the iterative estimator

$$\theta_{m+1} = \theta_m + \mathcal{I}^{-1}(\theta_m) \frac{\partial \ln p(\mathbf{x}; \theta_m)}{\partial \theta}. \tag{11.5}$$

The Fisher information does not depend on the data, while the second derivative of the log-likelihood function is data-dependent. It is possible to end up with an ill-conditioned second derivative expression that is almost singular. The iteration can become very unstable and unable to converge. The Fisher information alleviates this problem by eliminating the dependence on the observation.

11.6 Extension to vector parameter

The Newton-Raphson method can be generalized to vector parameter $\boldsymbol{\theta}$ which gives the following general iteration rule for the MLE:

$$\boldsymbol{\theta}_{m+1} = \boldsymbol{\theta}_m - \left(\mathbf{H}^{-1}(\boldsymbol{\theta}_,) \frac{\partial}{\partial \boldsymbol{\theta}} \ln p(\mathbf{x}; \boldsymbol{\theta}_m) \right) \tag{11.6}$$

where $\mathbf{H}(\boldsymbol{\theta})$ is the Hessian matrix with elements

$$[\mathbf{H}(\boldsymbol{\theta})]_{i,j} = \frac{\partial \ln(p(\mathbf{x};\boldsymbol{\theta}))}{\partial\theta_i\theta_j}. \tag{11.7}$$

Equivalently, the scoring method becomes

$$\boldsymbol{\theta}_{m+1} = \boldsymbol{\theta}_m + \mathbf{I}^{-1}(\boldsymbol{\theta}_m)\frac{\partial}{\partial\boldsymbol{\theta}}\ln p(\mathbf{x};\boldsymbol{\theta}_m), \tag{11.8}$$

where $\mathbf{I}(\boldsymbol{\theta})$ is the Fisher information matrix.

11.7 The Gauss-Newton method for least squares

The above introduced Newton-Raphson method can also be applied to nonlinear least-squares problems. However, we can derive another method by using the linear approximation of a function. The cost function of the LSE is given as

$$J(\theta) = \sum_{n=0}^{N-1}(x_n - s_n(\theta))^2. \tag{11.9}$$

Using the approximation (11.1), we obtain the following approximation of the cost function

$$\begin{aligned} J(\theta) &\approx \sum_{n=0}^{N-1}\left(x_n - s[n;\theta_0] - \left.\frac{\partial s_n(\theta)}{\partial\theta}\right|_{\theta=\theta_0}(\theta-\theta_0)\right)^2 \\ &= \sum_{n=0}^{N-1}\left(x_n - s_n(\theta_0)] + \left.\frac{\partial s[n;\theta]}{\partial\theta}\right|_{\theta=\theta_0}\theta_0 - \left.\frac{\partial s_n(\theta)}{\partial\theta}\right|_{\theta=\theta_0}\theta\right)^2 \\ &= \left(\mathbf{x} - \mathbf{s}(\theta_0) + \mathbf{h}(\theta_0)\theta_0 - \mathbf{h}(\theta_0)\theta\right)^T\left(\mathbf{x} - \mathbf{s}(\theta_0) + \mathbf{h}(\theta_0)\theta_0 - \mathbf{h}(\theta_0)\theta\right), \end{aligned} \tag{11.10}$$

with

$$\mathbf{h}(\theta_0) = \left.\left[\frac{\partial s[0;\theta]}{\partial\theta} \quad \frac{\partial s[1;\theta]}{\partial\theta} \quad \dots \quad \frac{\partial s[N-1;\theta]}{\partial\theta}\right]^T\right|_{\theta=\theta_0}. \tag{11.11}$$

Note that the expression $\mathbf{x} - \mathbf{s}(\theta_0) + \mathbf{h}(\theta_0)\theta_0$ is known. To find the minimum, we equate the derivative of (11.10) with zero and solve for θ, which yields

$$\hat{\theta} = \left(\mathbf{h}^T(\theta_0)\mathbf{h}(\theta_0)\right)^{-1}\mathbf{h}^T(\theta_0)\left(\mathbf{x} - \mathbf{s}(\theta_0) + \mathbf{h}(\theta_0)\theta_0\right), \tag{11.12}$$

or, equivalently,

$$\hat{\theta} = \theta_0 + \left(\mathbf{h}^T(\theta_0)\mathbf{h}(\theta_0)\right)^{-1}\mathbf{h}^T(\theta_0)\left(\mathbf{x} - \mathbf{s}(\theta_0)\right). \tag{11.13}$$

This method is the Gauss-Newton method.

To obtain a more accurate result, this procedure can be carried out iteratively using the following iteration steps

$$\theta_{m+1} = \theta_m + \left(\mathbf{h}^T(\theta_m)\mathbf{h}(\theta_m)\right)^{-1}\mathbf{h}^T(\theta_m)\left(\mathbf{x} - \mathbf{s}(\theta_m)\right). \tag{11.14}$$

The presented method can be generalized for vector parameter. The corresponding iteration rule is

$$\boldsymbol{\theta}_{m+1} = \boldsymbol{\theta}_m + \left(\mathbf{h}^T(\boldsymbol{\theta}_m)\mathbf{h}(\boldsymbol{\theta}_m)\right)^{-1}\mathbf{h}^T(\boldsymbol{\theta}_m)\left(\mathbf{x} - \mathbf{s}(\boldsymbol{\theta}_m)\right), \tag{11.15}$$

where $\mathbf{h}(\boldsymbol{\theta})$ is the Jacobian matrix with elements

$$\mathbf{h}(\boldsymbol{\theta})_{i,j} = \frac{\partial s[i;\boldsymbol{\theta}]}{\partial \theta_j}. \tag{11.16}$$

Starting from an initial value $\boldsymbol{\theta}_0$, the LSE can be iteratively calculated to converge to a minimum value. Of course, there have to be criteria to stop the iterations, such as a limited change in the parameter estimate, i.e., $(\boldsymbol{\theta}_{m+1} - \boldsymbol{\theta}_m)^T(\boldsymbol{\theta}_{m+1} - \boldsymbol{\theta}_m) < \epsilon$, or reaching a maximum number of iterations. The selection of the initial parameter vector $\boldsymbol{\theta}_0$ can be based on a grid search over the squared error function to ensure that the numerical solution begins close to the global maximum.

Part III

Spectral estimation

12

Introduction

Spectral estimation is concerned with estimating the power spectrum of a stochastic signal from a limited set of observations. The power spectrum describes how the power is distributed across frequencies and can thus provide important information about the signal itself. The distribution of the signal power across its frequency range can help in characterizing unwanted noise sources. Other purposes of spectral estimation include the detection of periodicities, signal compression, and signal classification. Manipulation of the spectrum through filtering allows for the denoising, analysis, and classification of signals. The key underlying aspect of all these signal processing techniques in the frequency domain is a good understanding of the spectrum of a signal.

As we shall see in this part, spectral estimation can be performed by two inherently different approaches. Non-parametric approaches do not assume any knowledge of the structure that correlates with the signal samples. They are simply based on operations performed on the available signal samples. Although several different non-parametric methods were developed, here, we will focus on methods based on the Fourier transform. On the other hand, parametric methods assume a model that generates the observed random signal. Spectral estimation then reduces to finding the model parameters and calculating the power spectral density (PSD) from the model. In here, we will restrict our attention to ARMA modeling for parametric spectral estimation.

12.1 Energy and power spectral distributions

This chapter will describe the main methods used to calculate the energy and power spectrum of a discrete-time signal. Initially, we will assume that the signal has zero-mean, is stationary, and is infinitely long. However, since especially the latter is usually not the case, we will later analyze the consequences of these non-ideal conditions.

12.1.1 Energy signals

As described in Section 3.3, if an infinitely long signal has finite signal energy it is called an energy signal. Since the energy is finite and the time duration infinite, the average signal power is zero. An example of an energy signal is a short pulse that is transmitted only once. An energy signal can be characterized by its energy spectral distribution, which describes the energy distribution of the signal in the frequency domain. The ideal energy spectral density is denoted by $\mathcal{E}(e^{j\theta})$ and can be calculated using a direct or an indirect method.

The total energy E_s of an energy signal $x[n]$ can be calculated as

$$E_s = \sum_{n=-\infty}^{\infty} |x[n]|^2, \tag{12.1}$$

which is bounded between 0 and ∞ by the definition of an energy signal.

Energy spectral distribution: direct method

The energy spectral distribution of a discrete-time energy signal can be determined by squaring the magnitude spectrum of signal spectrum $X(e^{j\theta})$ as

$$\mathcal{E}(e^{j\theta}) = \left|X(e^{j\theta})\right|^2 = \left| \sum_{n=-\infty}^{\infty} x[n]e^{-jn\theta} \right|^2, \tag{12.2}$$

where the second equality results from the definition of the Fourier transform for discrete-time signals. In order to show that this definition is valid, we first need to make use of Parseval's theorem, which states that the energy of a signal in the time-domain should be equal to the energy of the signal in the frequency-domain as

$$\sum_{n=-\infty}^{\infty} |x[n]|^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left|X(e^{j\theta})\right|^2 d\theta. \tag{12.3}$$

Secondly, the intuitive definition of the total signal energy is required, which is the integral of the energy spectral distribution $\mathcal{E}(e^{j\theta})$ over all frequencies as

$$E_s = \frac{1}{2\pi} \int_{-\pi}^{\pi} \mathcal{E}(e^{j\theta}) d\theta. \tag{12.4}$$

Combining (12.1), (12.3) and (12.4) leads to the desired direct calculation method as described by (12.2).

Energy spectral distribution: indirect method

Using the definition of the Fourier transform for discrete-time signals, the direct method can be rewritten as

$$\begin{aligned}
\mathcal{E}(e^{j\theta}) &= \left|X(e^{j\theta})\right|^2, \\
&= X(e^{j\theta})X^*(e^{j\theta}), \\
&= \left(\sum_{n=-\infty}^{\infty} x[n]e^{-jn\theta} \right) \left(\sum_{p=-\infty}^{\infty} x^*[p]e^{jp\theta} \right), \\
&= \sum_{p=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} x[n]x^*[p]e^{-j(n-p)\theta}, \\
&\overset{p=n-l}{=} \sum_{l=-\infty}^{\infty} \left(\sum_{n=-\infty}^{\infty} x[n]x^*[n-l]e^{-jl\theta} \right), \\
&= \sum_{l=-\infty}^{\infty} r_x[l]e^{-jl\theta}
\end{aligned} \tag{12.5}$$

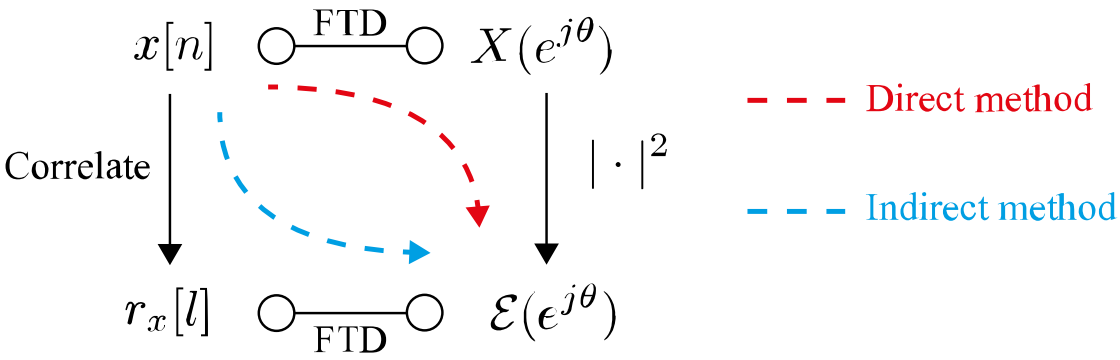


Figure 12.1: Schematic overview of the direct and indirect methods for calculating the energy spectral density of an energy signal.

which is the Fourier transform for discrete-time signals of the ideal autocorrelation function $r_x[l]$, which is defined as

$$r_x[l] = \mathrm{E}\left[x[n]x^*[n-l]\right] = \sum_{n=-\infty}^{\infty} x[n]x^*[n-l] = r_x^*[-l], \tag{12.6}$$

where l represents the lag between the signal $x[n]$ and its shifted version $x[n-l]$. As it is shown, the indirect method calculates the energy spectral distribution by calculating the Fourier transform for discrete-time signals of the autocorrelation function. This relation is described by the Wiener-Khintchine theorem. Figure 12.1 schematically shows both the direct and indirect method for calculating the energy spectral density.

Example 12.1

Given the signal $x[n] = a^n \cdot u[n]$, where $u[n]$ represents the unit step function and $|a| < 1$, calculate the frequency spectrum, autocorrelation function, energy spectral density and signal energy of $x[n]$.

Solution.

The frequency spectrum $X(e^{j\theta})$ can be calculated as

$$X(e^{j\theta}) = \sum_{n=-\infty}^{\infty} x[n]e^{-jn\theta} = \sum_{n=0}^{\infty} \left(ae^{-j\theta}\right)^n = \frac{1}{1-ae^{-j\theta}}.$$

The autocorrelation function can be determined as

$$\begin{aligned} r_x[l] &= \sum_{n=-\infty}^{\infty} x[n]x[n-l] \\ &= \sum_{n=-\infty}^{\infty} a^n u[n] a^{n-l} u[n-l] \\ &\overset{l\geq 0}{=} a^{-l} \sum_{n=l}^{\infty} a^{2n} \\ &= a^{-l} \left(\sum_{n=0}^{\infty} a^{2n} - \sum_{n=0}^{l-1} a^{2n} \right) \\ &= a^{-l} \left(\frac{1}{1-a^2} - \frac{1-a^{2l}}{1-a^2} \right) \\ &= \frac{a^l}{1-a^2}. \end{aligned}$$

Here only the case where $l \geq 0$ is discussed. Because the signal $x[n]$ is a real signal (i.e. $x[n] = x^*[n]$) the autocorrelation function is symmetric around $l = 0$ and is therefore fully given as

$$r_x[l] = \frac{a^{|l|}}{1-a^2}.$$

The energy spectral density can be determined as

$$\begin{aligned} \mathcal{E}(e^{j\theta}) = |X(e^{j\theta})|^2 &= \sum_{l=-\infty}^{\infty} r_x[l] e^{-jl\theta} \\ &= \frac{1}{1-ae^{-j\theta}} \frac{1}{1-ae^{j\theta}} = \frac{1}{(1+a^2) - 2a\cos(\theta)}. \end{aligned}$$

From this the total signal energy can be determined as

$$E_s = \frac{1}{2\pi} \int_{-\pi}^{\pi} \mathcal{E}(e^{j\theta}) d\theta = \sum_{n=-\infty}^{\infty} |x[n]|^2 = \sum_{n=0}^{\infty} a^{2n} = \frac{1}{1-a^2}.$$

12.1.2 Power signals

An infinitely long signal that has a finite average signal power is called a power signal. Because of the finite power that the signal carries over an infinitely long time, the total signal energy is infinite. Any non-zero bounded signal that is infinitely long can be regarded as a power signal. A power signal can be characterized by its PSD, which describes the power distribution of the signal in the frequency domain. The PSD is denoted by $P(e^{j\theta})$ and it can be calculated using a direct and an indirect method, which are known as the *periodogram* and *correlogram*, respectively.

Since the signal power is defined as the energy density over time, the average signal power

P_s of a windowed power signal $\tilde{x}[n]$ of length N can be calculated as

$$P_s = \lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} |\tilde{x}[n]|^2, \tag{12.7}$$

which is bounded between 0 and ∞ by the definition of a power signal. The windowing operation is explained in detail in Ch. 12.2.

Power spectral distribution: direct method

Similarly to the above, the average signal power can be intuitively understood as the integral of the power spectral distribution over frequencies as

$$P_s = \frac{1}{2\pi} \int_{-\pi}^{\pi} P(e^{j\theta}) d\theta = \lim_{N\to\infty} \frac{1}{2\pi} \int_{-\pi}^{\pi} P_N(e^{j\theta}) d\theta, \tag{12.8}$$

where the subscript N denotes the length of the windowed signal $\tilde{x}[n]$. The ideal power spectral distribution can be determined as

$$P(e^{j\theta}) = \lim_{N\to\infty} P_N(e^{j\theta}). \tag{12.9}$$

This definition cannot be evaluated directly since it requires an infinitely-long signal window. In order to define $P_N(e^{j\theta})$, first $X_N(e^{j\theta})$ has to be defined as the Fourier transform for discrete-time signals of the windowed signal $x[n]$ with length N as

$$X_N(e^{j\theta}) = \sum_{n=0}^{N-1} \tilde{x}[n] e^{-jn\theta} \;\; \circ\!\!-\!\!\circ \;\; \tilde{x}[n] = \frac{1}{2\pi} \int_{-\pi}^{\pi} X_N(e^{j\theta}) e^{jn\theta} d\theta. \tag{12.10}$$

By rewriting (12.7) using (12.10), the following expansion can be obtained

$$\begin{aligned}
P_s &= \lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} |\tilde{x}[n]|^2, \\
&= \lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \tilde{x}[n] \tilde{x}^*[n] \\
&= \lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \tilde{x}[n] \left(\frac{1}{2\pi} \int_{-\pi}^{\pi} X_N^*(e^{j\theta}) e^{-jn\theta} d\theta \right) \\
&= \lim_{N\to\infty} \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{1}{N} \left(\sum_{n=0}^{N-1} \tilde{x}[n] e^{-jn\theta} \right) X_N^*(e^{j\theta}) d\theta \\
&= \lim_{N\to\infty} \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{1}{N} X_N(e^{j\theta}) X_N^*(e^{j\theta}) d\theta \\
&= \lim_{N\to\infty} \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{1}{N} \left| X_N(e^{j\theta}) \right|^2 d\theta,
\end{aligned} \tag{12.11}$$

from which the direct method of calculating the power spectral distribution follows by comparison with (12.8) as

$$\hat{P}_N(e^{j\theta}) = \frac{1}{N} \left| X_N(e^{j\theta}) \right|^2 = \frac{1}{N} \left| \sum_{n=0}^{N-1} \tilde{x}[n] e^{-jn\theta} \right|^2. \tag{12.12}$$

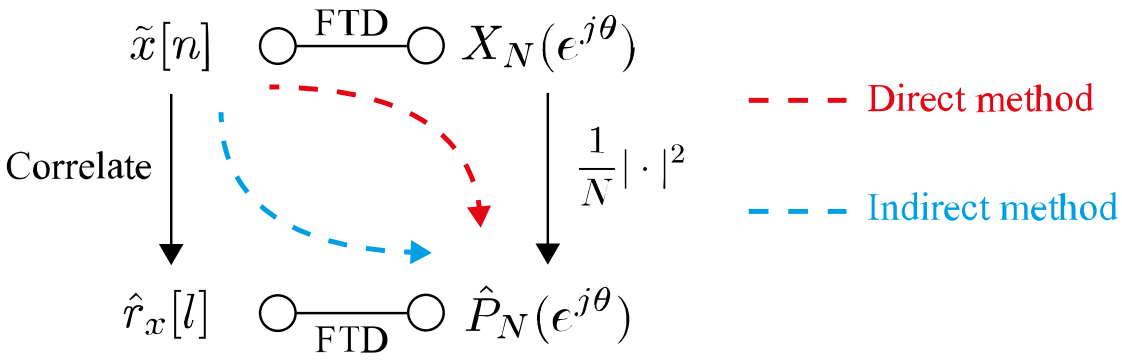


Figure 12.2: Schematic overview of the direct and indirect methods for calculating the PSD of a windowed power signal.

It should be noted that this definition is different from the energy spectral distribution as the spectrum is only calculated using a finite window of length N. In order to calculate the true PSD N should approach ∞.

Power spectral distribution: indirect method

By expanding the definition of $P_N(e^{j\theta})$, the indirect method of determining the power spectral distribution can be found as

$$\begin{aligned}
\hat{P}_N(e^{j\theta}) &= \frac{1}{N}\left|X_N(e^{j\theta})\right|^2 \\
&= \frac{1}{N}X_N(e^{j\theta})X_N^*(e^{j\theta}) \\
&= \frac{1}{N}\left(\sum_{n=0}^{N-1}\tilde{x}[n]e^{-jn\theta}\right)\left(\sum_{p=0}^{N-1}\tilde{x}^*[p]e^{jp\theta}\right) \\
&= \frac{1}{N}\sum_{p=0}^{N-1}\sum_{n=0}^{N-1}\tilde{x}[n]\tilde{x}^*[p]e^{-j(n-p)\theta} \\
&\overset{p=n-l}{=} \sum_{l=-(N-1)}^{N-1}\left(\frac{1}{N}\sum_{n=0}^{N-1-|l|}\tilde{x}[n]\tilde{x}^*[n-|l|]e^{-jl\theta}\right) \\
&= \sum_{l=-(N-1)}^{N-1}\hat{r}_x[l]e^{-jl\theta},
\end{aligned} \tag{12.13}$$

where $\hat{r}_x[l]$ is an estimate of the autocorrelation function of $x[n]$. As explained in Chapter 3, this can be obtained as

$$\hat{r}_x[l] = \frac{1}{N}\sum_{n=0}^{N-1-|l|}\tilde{x}[n]\tilde{x}^*[n-|l|]. \tag{12.14}$$

Figure 12.2 schematically shows the direct and indirect methods for calculating the PSD of a windowed power signal $\tilde{x}[n]$.

Example 12.2

Consider the power signal $x[n] = u[n]$. Calculate the frequency spectrum, autocorrelation function, PSD and average power of the windowed signal $x[n]$ for $0 \leq n < N-1$.

Solution.

The frequency spectrum can be calculated as

$$\begin{aligned} X_N(e^{j\theta}) &= \sum_{n=-\infty}^{\infty} \tilde{x}[n]e^{-jn\theta}, \\ &= \sum_{n=0}^{N-1} e^{-jn\theta}, \\ &= \frac{1-e^{-jN\theta}}{1-e^{-j\theta}}, \\ &= \frac{(e^{j\frac{N}{2}\theta} - e^{-j\frac{N}{2}\theta})e^{-j\frac{N}{2}\theta}}{(e^{j\frac{1}{2}\theta} - e^{-j\frac{1}{2}\theta})e^{-j\frac{1}{2}\theta}} \\ &= \frac{2j}{2j}\frac{\sin(\frac{N}{2}\theta)}{\sin(\frac{1}{2}\theta)}e^{-j\frac{N-1}{2}\theta} \\ &= \frac{\sin(\frac{N}{2}\theta)}{\sin(\frac{1}{2}\theta)}e^{-j\frac{N-1}{2}\theta}. \end{aligned}$$

The autocorrelation function for $|l| < N$ can be estimated as

$$\begin{aligned} \hat{r}_x[l] &= \frac{1}{N}\sum_{n=0}^{N-1-|l|} \tilde{x}[n]\tilde{x}[n-|l|], \\ &= \frac{1}{N}\sum_{n=0}^{N-1-|l|} 1, \\ &= \frac{N-|l|}{N}. \end{aligned}$$

The PSD can be calculated as

$$\hat{P}_N(e^{j\theta}) = \frac{1}{N}|X_N(e^{j\theta})|^2 = \frac{1}{N}\frac{\sin^2(\frac{N}{2}\theta)}{\sin^2(\frac{1}{2}\theta)}.$$

The average signal power can be determined as

$$P_s = \hat{r}_x[0] = 1.$$

12.2 Windowing and zero padding

The methods presented in the previous sections for the calculation of the power spectrum ideally require an infinitely-long signal for the most accurate estimation. However, in practice, all real-

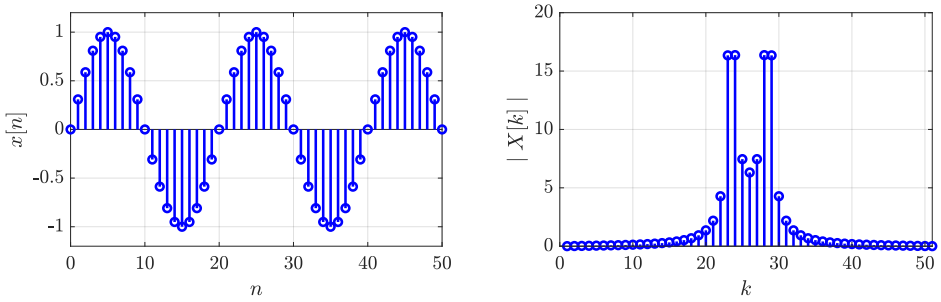


Figure 12.3: A sinusoidal discrete-time signal and the corresponding discrete-time Fourier transform.

world signals are windowed, which means that only a finite portion of the signal is considered, instead of the entire signal. Windowing can result from many different causes. First, it is highly unlikely that an infinite number of signal samples are available for analysis and therefore it is inherently windowed by the limited observation time. Secondly, windowing makes an algorithm less computationally expensive. There are operations, such as matrix inversion, whose computational costs do not scale linearly with the number of samples. Therefore it is important to use a limited number of samples. Finally, there are cases in which the signal is not stationary, but most analysis methods assume stationarity. In these cases, the signal is windowed such that it can be assumed to be locally stationary.

Calculating the spectrum of signal $x[n]$ when only the windowed signal $\tilde{x}[n]$ is available leads to several issues. First of all, in most practical cases the spectrum will be calculated through the N-point discrete-time Fourier transform (DFT), which is actually a sampled version of the Fourier transform for discrete-time signals (FTD). Therefore, it is possible that important signal characteristics will get lost. Secondly, windowing causes spectral leakage, which means that single frequency components will be spread throughout the frequency spectrum. Finally, windowing leads to loss of resolution, which relates to the ability to distinguish between different spectral components.

A good understanding of the Fourier transform and its variants is a necessary prerequisite for spectral estimation. A short review of the Fourier transforms is provided in Appendix D. For a more in-depth treatment, students are referred to material from previous courses.

In fact, careful understanding of the Fourier transform and the effects of windowing are necessary in order to avoid inaccurate estimation and/or misinterpretation of the obtained results. An example of such a situation is represented in Figure 12.3. Here a single discrete-time sinusoidal signal is plotted over time. Suppose that an ambitious student tries to determine the frequency of this sinusoid by simply applying the fft()-function in MATLAB. He or she would expect to find only 2 peaks at the positive and negative frequency of the sinusoid, but instead, the right-hand plot in Figure 12.3 is obtained. Here, there are multiple frequency components present, which seem to be unrelated to the original sinusoidal signal; moreover, the frequency axis does not immediately correspond to the actual frequency, but to some index k.

After more careful studying, the student comes to the realization that the operation he performed just gave him an estimate of the frequency content. In this specific case, the student could easily have obtained the expected frequency spectrum. The student neglected the fact that the fft()-operation assumes a periodic signal. In this case, the signal is windowed, meaning that the infinitely long signal is cropped to a finite length, which is no longer periodic, leading to potentially undesirable results in the frequency domain.

From this we can conclude that just calculating the spectrum of a signal, especially when it is comprised of multiple signals, is not as easy as it sounds. Since any real-world signal can be observed only for a limited amount of time, all real-world signals are virtually windowed. Besides the effects of windowing, we cannot know in advance whether the signal is periodic or not. Because of these reasons, it is usually impossible to determine the exact spectrum of an observed random signal. As we shall in Ch. 13, our spectral estimate will be affected by bias and variance.

12.2.1 Rectangular window

To illustrate the consequences of windowing, we will show the case of the simplest window: the rectangular window. The rectangular window $w[n]$ can be defined as

$$w[n] = \begin{cases} 1, & \text{for } n = 0, 1, \ldots, N-1 \\ 0, & \text{otherwise} \end{cases} \tag{12.15}$$

where N is the length of the window.

The windowed signal $\tilde{x}[n]$ is obtained by multiplying the original infinitely-long signal $x[n]$ with the window in the time domain:

$$\tilde{x}[n] = w[n]x[n]. \tag{12.16}$$

For a rectangular window, this operation can be regarded as taking only a finite set of samples into account.

Suppose that we know that the Fourier transform of discrete-time signals of $x[n]$ is ideally $X(e^{j\theta})$ and we would like to see the effect of windowing on the estimated spectrum. Recall that a convolution in the time-domain results in a multiplication in the frequency domain. It can be shown that the opposite relation holds as well: a multiplication in the time domain will result in a periodic convolution in the frequency domain. This relationship leads to the following frequency response of the windowed signal $\tilde{X}(e^{j\theta})$ as

$$\tilde{X}(e^{j\theta}) = X(e^{j\theta}) * W(e^{j\theta}), \tag{12.17}$$

where $W(e^{j\theta})$ is the frequency response of the window. So the frequency response of the window determines how the original spectrum is influenced by windowing. From the definition of the Fourier transform for discrete-time signals, the spectrum of the rectangular window can be

determined as

$$\begin{aligned} W(e^{j\theta}) &= \sum_{n=-\infty}^{\infty} w[n]e^{-jn\theta}, \\ &= \sum_{n=0}^{N-1} e^{-jn\theta}, \\ &= \frac{1-e^{-jN\theta}}{1-e^{-j\theta}}, \\ &= \frac{\left(e^{j\frac{N}{2}\theta} - e^{-j\frac{N}{2}\theta}\right)e^{-j\frac{N}{2}\theta}}{\left(e^{j\frac{1}{2}\theta} - e^{-j\frac{1}{2}\theta}\right)e^{-j\frac{1}{2}\theta}}, \\ &= \frac{\frac{1}{2j}\left(e^{j\frac{N}{2}\theta} - e^{-j\frac{N}{2}\theta}\right)}{\frac{1}{2j}\left(e^{j\frac{1}{2}\theta} - e^{-j\frac{1}{2}\theta}\right)}e^{-j\frac{N-1}{2}\theta}, \\ &= \frac{\sin(N\theta/2)}{\sin(\theta/2)}e^{-j\frac{N-1}{2}\theta}, \end{aligned} \tag{12.18}$$

where the ratio $\frac{\sin(N\theta/2)}{\sin(\theta/2)}$ is the Dirichlet or periodic and aliased sinc function. This function is convolved with the true spectrum $X(e^{j\theta})$ to obtain the estimated spectrum $\tilde{X}(e^{j\theta})$ from the windowed signal $\tilde{x}[n]$. The frequency response of the rectangular window is also a function of the window length N. Figure 12.4 shows three rectangular windows with their respective normalized magnitude responses on the logarithmic scale for increasing N. It can be noted that an increase of N results in a better resolution of the estimated spectrum, which means that the width of the main lobe is smaller. However, the height of the side lobes does not change with N. This has important consequences for spectral leakage, as will be discussed hereafter.

From Figure 12.5 it can be noted that the frequency response always consists of several lobes. The largest lobe is called the main lobe and the peak of this lobe is usually normalized to 1, which is 0 dB. This lobe should be as narrow as possible in order to get an accurate approximation of the real frequency spectrum. Its width can be described by the so-called -3 dB bandwidth, which is the width of the frequency interval on which the spectral power is reduced by less than a factor 2. In other words, the -3 dB bandwidth includes all frequencies whose power is still larger than half its original value. It turns out that the -3 dB bandwidth can be approximately calculated, as shown hereafter.

12.2.2 Different types of windows

The rectangular window is used in most cases because of its high resolution. However, the relatively high side lobe level can cause problems when small frequency components are to be detected. Changing the shape of the window can result in different frequency characteristics, such as a different resolution or side lobe level. Numerous different windows have been proposed over the years. Figure 12.6 shows the rectangular, triangular, and Hanning windows for the same window length. In the corresponding normalized frequency spectra, it can be seen that there is a clear trade-off between the -3 dB bandwidth (resolution) and the side lobe levels (spectral leakage). This means that you can only improve either the resolution or the spectral leakage when estimating a spectrum using windowing.

As said before, the -3 dB bandwidth is dependent on the length of the window and it can be analytically approximated. The side lobe level is dependent on the type of window that is

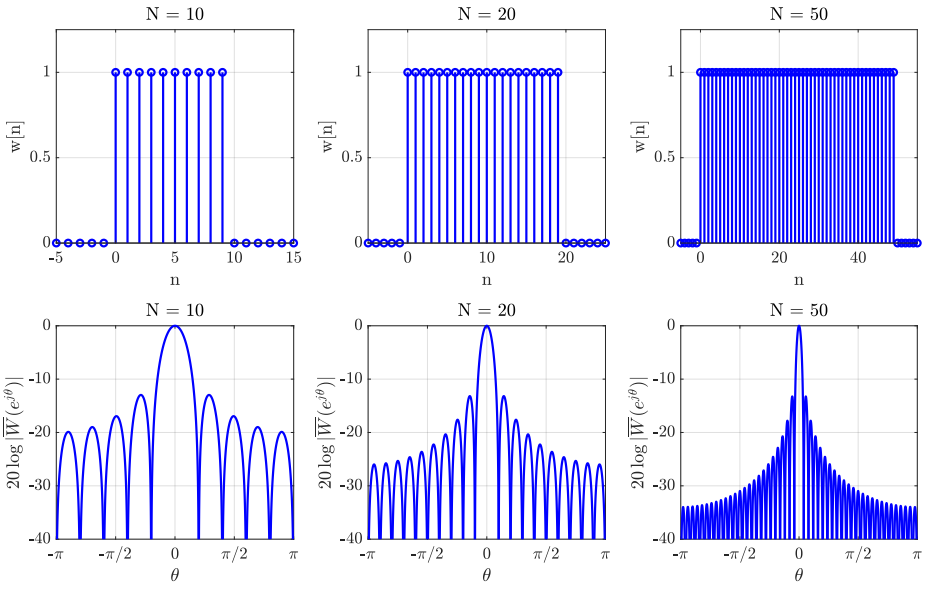


Figure 12.4: The normalized magnitude response of a rectangular window for increasing window lengths N. From the figure it can be noted that the resolution increases for an increasing N. The side lobe level is identical in all three cases.

Window	-3 dB bandwidth	Side lobe level
Rectangular	$1.81\frac{\pi}{N-1}$	-13 dB
Triangular	$5.01\frac{\pi}{N-1}$	-27 dB
Hanning	$6.27\frac{\pi}{N-1}$	-32 dB

Table 12.1: Length-bandwidth trade-off for different types of window.

chosen, but it is not influenced by window length. Tab. 12.1 shows the -3 dB bandwidth and the side lobe levels for various windowing functions.

12.2.3 Loss of resolution and spectral leakage

The resolution, that is the ability to distinguish different spectral peaks, is an important characteristic of a spectral estimation method. When windowing, the resolution depends on the width of the main lobe of the window spectrum. We show this with a practical example. In Figure 12.7, an ideal frequency spectrum of a periodic signal is given. The approximations of this signal due to windowing are shown in all the other plots. Several types and lengths of windows have been used and several observations can be made. First, when two frequency components are close together they may not be distinguishable if the resolution of the window is too low. Secondly, the spectral peak widths are dependent on both the window type and window length. Finally, the influence of the spectral leakage can also be noted. It creates additional spurious peaks, which might overshadow smaller frequency components.

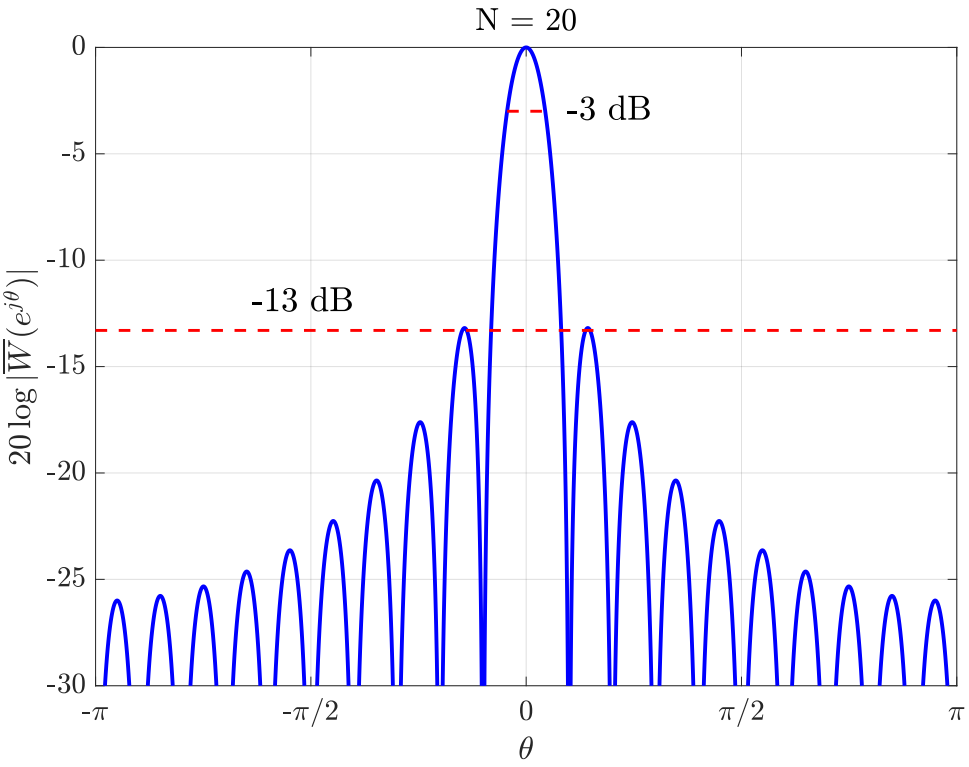


Figure 12.5: Nomenclature of the normalized magnitude response of a window function, where the 3dB bandwidth and the side lob level are indicated.

12.2.4 Zero-padding

If we would like to calculate the spectrum of an infinitely long signal using only a finite number of samples then we will need to window our signal in some way. These windows influence the frequency spectrum that we approximate. The approximated frequency spectra discussed so far were continuous frequency spectra obtained by the Fourier transform for discrete-time signals (FTD). However, in practice, this Fourier transform cannot be obtained by computing software; therefore, the discrete-time Fourier transform (DFT) is usually used, which is commonly implemented as the fast Fourier transform. In short, the discrete-time Fourier transform does not return a continuous frequency spectrum, but it returns samples of the continuous spectrum (see Appendix D).

Thus, besides the consequences of windowing, the estimated spectrum is also a sampled version of the true continuous spectrum. If the estimated continuous spectrum is represented by $\hat{X}(e^{j\theta})$, the sampled spectrum is denoted by $X[k]$, where k is the sample index, ranging from 0 to $N-1$, with N the window length. When it is not possible to extend the window length to obtain more samples, it is also possible to zero-pad the signal. Suppose that a windowed signal $\tilde{x}[n]$ of length N is observed and the spectrum is calculated using the DFT. This DFT samples the FTD of the windowed signal. The number of samples simply equals the length of the windowed signal, thus N. To decrease the step at which we "walk" on the continuous spectrum, the signal can be extended with $L-N$ zeros, such that the signal, as well as the obtained spectrum, have a

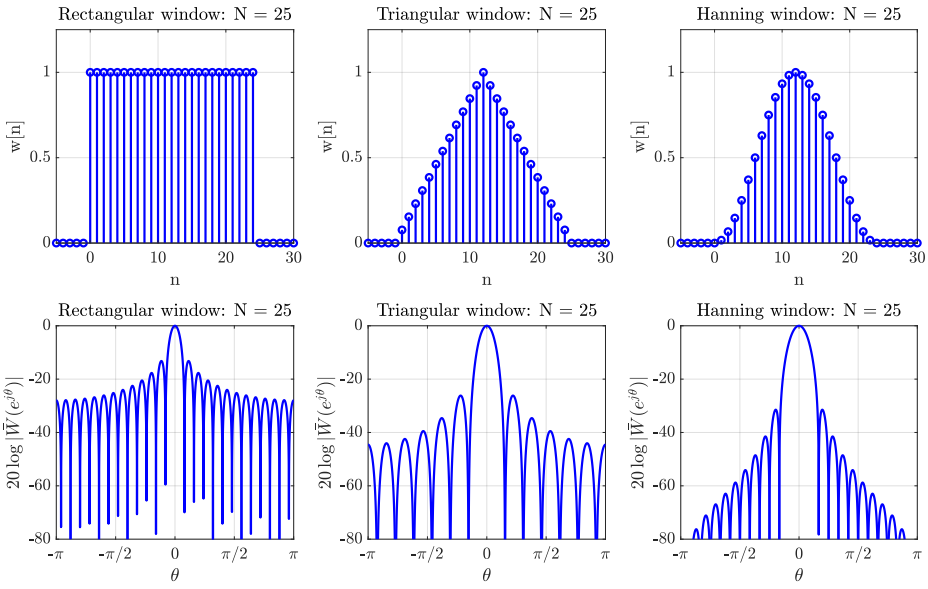


Figure 12.6: Three types of windows, including the rectangular window, the triangular window, and the Hanning window, with their respective normalized frequency spectra.

length of $L > N$. This operation is known as zero-padding. To understand how this work, recall the definition of the DFT, by which the sampled spectrum $\hat{X}[k]$ can be determined as

$$\hat{X}[k] = \sum_{n=0}^{N-1} \tilde{x}[n] e^{-j\frac{2\pi}{N}kn}. \qquad \text{for } k = 0, 1, \ldots, N-1 \tag{12.19}$$

If the signal is now zero-padded to length L, the range of k changes and so does the factor in the complex exponential, which determines the sampling interval. This can be seen by

$$\begin{aligned} \hat{X}[k] &= \sum_{n=0}^{L-1} \tilde{x}[n] e^{-j\frac{2\pi}{L}kn}, \qquad \text{for } k = 0, 1, \ldots, L-1 \\ &= \sum_{n=0}^{N-1} \tilde{x}[n] e^{-j\frac{2\pi}{L}kn}. \qquad \text{for } k = 0, 1, \ldots, L-1 \end{aligned} \tag{12.20}$$

The sampling distance in the frequency domain is now decreased, but the summation is still from $n = 0$ up to $N - 1$, because for larger values of n the signal $\tilde{x}[n]$ is simply zero.

Figure 12.8 shows the approximated continuous spectrum (FTD) of a windowed signal of length N. Besides this, the DFT of the windowed signal is calculated, leading to the sampling of the continuous spectrum. Furthermore, the signal is zero-padded up to several lengths L, which changes the number of samples and the sampling intervals of the DFT. Depending on the length of the signal and the length of the zero-padding, some interesting phenomena can be observed, because in some cases the values at the sampling locations might give an unrealistic view of the frequency spectrum.

Let us go briefly through Figure 12.8. The ideal spectrum $X(e^{j\theta})$ is plotted in blue. This could be obtained only from an infinite-length signal. In red, we see the estimated FTD of

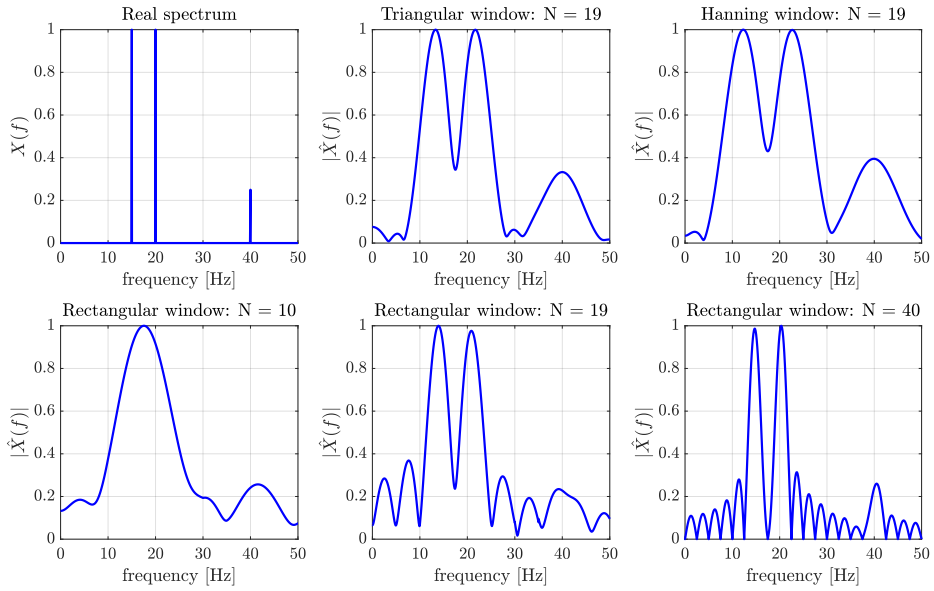


Figure 12.7: A signal with three frequency components is windowed by several windows, differing in type and length, and the final estimated normalized frequency spectrum is determined. It can be seen how the length of the filter and the type of window influence the resolution, i.e. the ability to distinguish between spectral components, and the spectral leakage, i.e. the appearance of spurious peaks.

the signal after applying a window of N=30. This function is continuous and it is calculated analytically by knowledge of the $x[n]$ before windowing and by knowledge of the window function $w[n]$. However, in practice, we do not know $x[n]$ before windowing, so we calculate the FFT from the windowed signal $\tilde{x}[n]$, obtaining a discrete function $\tilde{X}[k]$. If we only use $N = 30$ samples to calculate the FFT (no zero padding), we obtain the function in green, by which we would probably misinterpret the signal as composed of two frequency peaks. If we apply a zero padding of 90 samples, then we obtain the discrete function in pink, from which we can correctly see the main peak. By applying zero padding we have increased the step at which we walk on the underlying red function, which is the theoretical $\tilde{X}(e^{j\theta})$ obtained from the windowed signal. However, although our understanding of the original signal might improve by virtually increasing the number of samples on which we calculate the FFT, we can never recover the true spectrum (in blue) after windowing.

It is important to keep in mind that zero-padding does not improve the resolution of the estimated frequency spectrum, as this is only determined by the window type and length. The zero-padding only influences the sampling interval of the DFT. In other words, it determines the step at which we "walk" on the underlying continuous spectrum. Moreover, since windowing causes loss of information (it is equivalent to setting to zero all the samples outside the window), the continuous spectrum is an approximated version of the real spectrum. To summarize, in practice, we can only compute a sampled, approximated version of the frequency spectrum.

To relate this section to the topic of PSD estimation, it should be noted that there is a direct relation between the spectrum of a signal and the PSD. Since the discrete-time Fourier transform is often difficult to compute analytically, it is usually approximated by the sampled version of

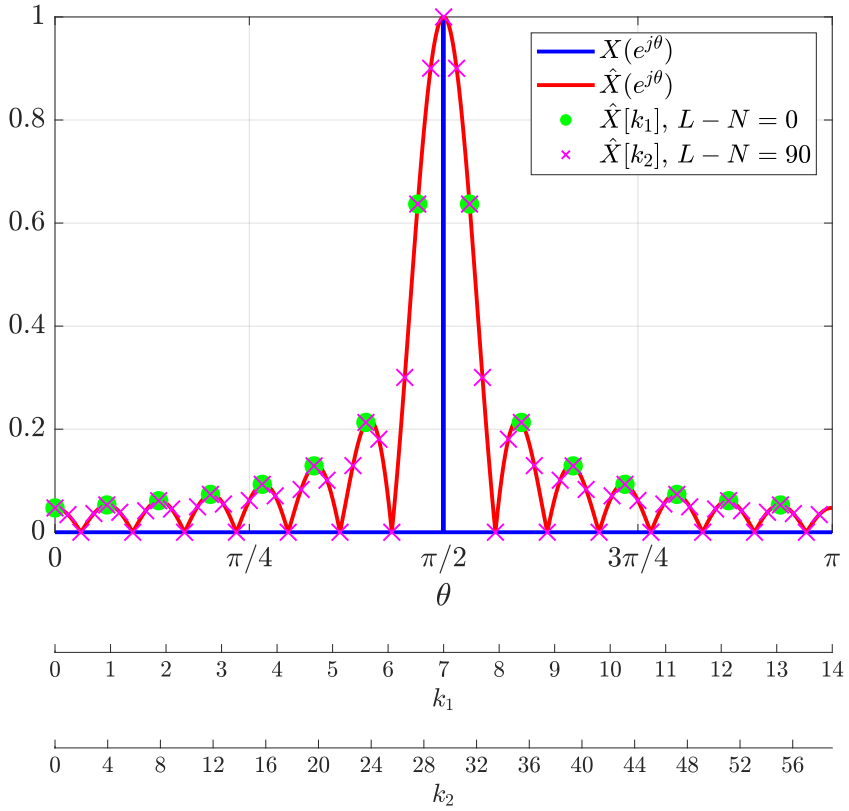


Figure 12.8: Comparison between an ideal frequency spectrum of a sinusoidal signal with frequency $\theta_0 = \pi/2$, an approximated frequency spectrum using a rectangular window of length N, where $N = 30$, and the DFT of the zero-padded signal for different lengths L. Keep in mind that this is a single-sided spectrum and therefore the number of sample points is half of the length of the zero-padded signal.

the spectrum determined by the discrete Fourier transform. As a direct consequence, the PSD is also a sampled, approximated version of the real power spectrum.

13

Non-parametric spectral estimation

This chapter will discuss several techniques to perform PSD estimation by non-parametric methods. In Chapter 12, we have seen the direct and indirect methods for calculating the PSD, which are also referred to as *periodogram* and *correlogram*, respectively. In the following, we will start by showing the performance of these two estimators of the PSD in terms of bias and variance, and then we will describe methods to improve the estimation performance.

13.1 Performance of the periodogram and correlogram

Here we briefly recap the "raw" estimators of the PSD: the direct method, also referred to as periodogram, and the indirect method, also referred to as correlogram.

Periodogram (direct method)

The periodogram estimate of the PSD is obtained as

$$\hat{P}_N(e^{j\theta}) = \frac{1}{N}\left|\sum_{n=0}^{N-1} x[n]e^{-jn\theta}\right|^2 . \tag{13.1}$$

which is the normalized squared magnitude of the spectrum of the windowed signal. The subscript N indicates that the PSD estimate is calculated on N samples of the infinite-length signal $x[n]$.

Correlogram (indirect method)

The correlogram estimate of the PSD is obtained as

$$\hat{P}_N(e^{j\theta}) = \sum_{l=-(N-1)}^{N-1} \hat{r}_x[l]e^{-jl\theta}. \tag{13.2}$$

which, recalling the Wiener-Khintchine relationship

$$P(e^{j\theta}) = \sum_{l=-\infty}^{\infty} r[l]e^{-jl\theta}, \tag{13.3}$$

is interpreted as the Fourier transform of the autocorrelation, but applied to an estimated autocorrelation function. This estimate is obtained for lags $-(N-1) \leq l \leq N-1$ from a signal of length N.

13.1.1 Performance of the "raw" estimators

The periodogram and correlogram are equivalent methods to calculate a PSD estimate $\hat{P}(e^{j\theta})$. Since this is an estimate, we can calculate the expected value and variance to assess the estimator's performance. In the following, we mostly focus on the correlogram, but same conclusions can be found for the periodogram.

Biased and unbiased estimators of the autocorrelation function

Before we dig into the estimation performance, let us take a look at the estimators of the autocorrelation function from a windowed signal. In Section 3.4, we provided an approximate estimator of the autocorrelation function for ergodic signals as

$$\hat{r}_b[l] = \frac{1}{N} \sum_{n=0}^{N-1-|l|} x[n]x^*[n-l]. \tag{13.4}$$

This is actually a biased estimator of the autocorrelation function (hence the subscript b), as can be proven by taking the expected value

$$\begin{aligned} \mathrm{E}\left[\hat{r}_b[l]\right] &= \mathrm{E}\left[\frac{1}{N} \sum_{n=0}^{N-1-|l|} x[n]x^*[n-|l|]\right], \\ &= \frac{1}{N} \sum_{n=0}^{N-1-|l|} \mathrm{E}\left[x[n]x^*[n-|l|]\right], \\ &= \frac{N-|l|}{N} r_x[l]. \end{aligned} \tag{13.5}$$

From (13.5), it is easy to understand that in order to obtain an unbiased estimate of the autocorrelation function, we need to use the following unbiased estimator

$$\hat{r}_{ub}[l] = \frac{1}{N-|l|} \sum_{n=0}^{N-1-|l|} x[n]x^*[n-l]. \tag{13.6}$$

Note that in both (13.4) and (13.6), we assume that the autocorrelation is zero for lags outside of the summation. While there is a large variance for lags l close to N, for both the biased and unbiased estimators the variance goes to zero asymptotically with N. This is proven for the biased estimator by (13.5).

An alternative way to look at this is to rewrite (13.4) and (13.6) as

$$\hat{r}_b[l] = r[l] \cdot w_B[l], \tag{13.7}$$

$$\hat{r}_{ub}[l] = r[l] \cdot w_R[l], \tag{13.8}$$

where $r[l]$ is the true autocorrelation function, while

$$w_R[l] = \begin{cases} 1, & \text{for } |l| < N \\ 0, & \text{otherwise} \end{cases} \tag{13.9}$$

and

$$w_B[l] = \begin{cases} \frac{N-|l|}{N}, & \text{for } |l| < N \\ 0, & \text{otherwise} \end{cases} \tag{13.10}$$

are rectangular and triangular windows, respectively, applied for lags $-N < l < N$ in the correlation domain. The triangular window is commonly referred to as Bartlett window. Equations (13.7) and (13.8) basically show that the unbiased estimate is equivalent to looking at the true autocorrelation function in a limited window, while the biased estimate provides a bias equivalent to multiplying for a triangular window.

Bias of the power spectral density estimator

The expected value of $\hat{P}(e^{j\theta})$ can be found as

$$\begin{aligned} \mathrm{E}\left[\hat{P}(e^{j\theta})\right] &= \mathrm{E}\left[\sum_{l=-(N-1)}^{N-1} \hat{r}x[l]e^{-jl\theta}\right], \\ &= \sum_{l=-(N-1)}^{N-1} \mathrm{E}\left[\hat{r}_x[l]\right] e^{-jl\theta}. \end{aligned} \tag{13.11}$$

Combining (13.11) with (13.7) or (13.8), and since the window is deterministic and the expectation of the true autocorrelation is the autocorrelation itself, we can write

$$\mathrm{E}\left[\hat{P}(e^{j\theta})\right] = \sum_{l=-\infty}^{\infty} \mathrm{E}\left[w[l]r_x[l]\right] e^{-jl\theta} = \sum_{l=-\infty}^{\infty} w[l]\,\mathrm{E}\left[r_x[l]\right] e^{-jl\theta} = \sum_{l=-\infty}^{\infty} w[l]r_x[l]e^{-jl\theta}, \tag{13.12}$$

where $w[l]$ is a generic window function, which is non zero only for $-N < l < N$.

When this relationship is regarded in the Fourier domain the multiplication-convolution property of the Fourier transform should be taken into account. It states that a multiplication in the time domain results in a convolution in the frequency domain. Therefore, the expected value of the estimated PSD function can be interpreted as the (periodic) convolution in the frequency domain between the true PSD function and the spectrum of a window function. This convolution is given by

$$\mathrm{E}\left[\hat{P}(e^{j\theta})\right] = P(e^{j\theta}) *_{2\pi} W(e^{j\theta}) = \frac{1}{2\pi}\int_{-\pi}^{\pi} P(e^{j\theta})W(e^{j(\theta-\phi)})d\phi \tag{13.13}$$

This shows that the correlogram (or periodogram) provides an estimate of the PSD which is a smoothed version of the true PSD (see Figure 13.1). As $N \to \infty$ the spectrum of the window function will approach a delta pulse and the expected value of the estimated periodogram will converge to the true periodogram. In fact, convolving any function with a delta pulse gives the function itself.

From (13.13) we also notice that the expected value of the PSD estimator is related to the spectrum of the window function. For a rectangular window, this results in

$$[H]W_R(e^{j\theta}) = \left(\frac{\sin(N\theta/2)}{\sin(\theta/2)}\right) e^{-j\frac{N-1}{2}\theta}, \tag{13.14}$$

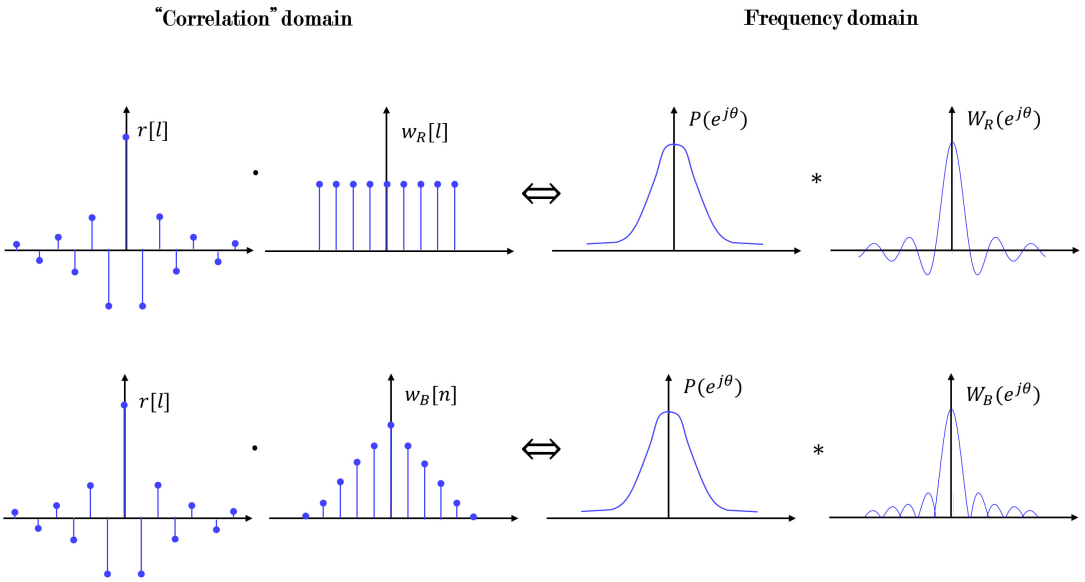


Figure 13.1: Schematic representation of the PSD estimation by the correlogram method using the unbiased (top) and biased (bottom) estimators of the autocorrelation function. The unbiased estimator (top) can lead to a non-negative PSD.

which is a periodic sinc function (diric).

For a triangular (Bartlet) window, we obtain

$$W_B(e^{j\theta}) = \frac{1}{N}\left(\frac{\sin(N\theta/2)}{\sin(\theta/2)}\right)^2, \tag{13.15}$$

which is a squared periodic sinc function (squared diric). The proof of (13.15) is beyond the scope of this reader, but it can easily be found by regarding the Bartlett window as a convolution between two rectangular windows of the same length, that is $w_B[n] = w_R[n] * w_R[n]$. Then, by the multiplication-convolution property of the Fourier transform, in the frequency domain the result is the multiplication of two rectangular window transform, that is $W_B(e^{j\theta}) = W_R(e^{j\theta}) \cdot W_R(e^{j\theta}) = = \text{diric}(\theta) \cdot \text{diric}(\theta) = \text{diric}^2(\theta)$.

Since $W_R(e^{j\theta})$ can have negative values, it may lead to an invalid PSD function, which by definition is always non-negative. For a triangular window, we instead obtain a squared periodic sinc, which is a non-negative function. This explains why we typically use the biased estimate of the autocorrelation function $r_b[l]$ given in (13.4). In fact, although unbiased, using $r_{ub}[l]$ to estimate the PSD by the correlogram method might lead to an invalid PSD.

A key aspect here is that the window is applied directly to the autocorrelation rather than to the signal. Thus, to obtain the PSD estimate, we simply take the transform of the windowed autocorrelation (correlogram), while if we were to calculate the PSD estimate from the signal, we would need to calculate the modulus squared of the transform (periodogram).In the latter case, a rectangular window applied to the signal in the time domain would not lead to an invalid PSD.

Loss of resolution and spectral leakage

In Chapter 12.2, we saw how windowing a signal before calculating the spectrum leads to loss of resolution and spectral leakage. To better understand the contributions of the main and side lobes, let us rewrite the window function as a sum of two components as $W_B(e^{j\theta}) = W_{ML}(e^{j\theta}) + W_{SL}(e^{j\theta})$, with $W_{ML}(e^{j\theta})$ accounting for the main lobe and given by

$$W_{ML}(e^{j\theta}) = \begin{cases} W_B(e^{j\theta}), & \text{for } |\theta| < \frac{2}{\pi} \\ 0, & \text{otherwise} \end{cases}, \tag{13.16}$$

and $W_{SL}(e^{j\theta})$ accounting for the side lobes and given by

$$W_{SL}(e^{j\theta}) = W_B(e^{j\theta}) - W_{ML}(e^{j\theta}). \tag{13.17}$$

Then, we can rewrite equation (13.13) as

$$\hat{P}(e^{j\theta}) = \underbrace{P(e^{j\theta}) *_{2\pi} W_{ML}(e^{j\theta})}_{\text{Loss of resolution}} + \underbrace{P(e^{j\theta}) * 2\pi W_{SL}(e^{j\theta})}_{\text{Spectral leakage}} \tag{13.18}$$

From (13.18), we can easily separate the contribution of the main lobe, which causes loss in spectral resolution, from the contribution of the side lobes, which cause spectral leakage, that is the appearance of spurious spectral peaks at the location of the side lobes.

Variance of the power spectral density estimator

The variance of the "raw" PSD estimator (periodogram/correlogram) is rather challenging to calculate since it has dependencies on the fourth-order moment of the signal. In the simple case of an AR(1) process, it can be calculated as

$$\text{Var}\left[\hat{P}_{AR1}(e^{j\theta})\right] = \sigma_i^2 \left[1 + \left(\frac{1}{N}\frac{\sin(\theta N)}{\sin\theta}\right)^2\right]. \tag{13.19}$$

Another simple case is a normally distributed sequence $x[n]$ input to an LTI, for which we obtain

$$\text{Var}\left[\hat{P}_x(e^{j\theta})\right] = (P(e^{j\theta}))^2 \left[1 + \left(\frac{1}{N}\frac{\sin(\theta N)}{\sin\theta}\right)^2\right]. \tag{13.20}$$

Although the proof is beyond the scope of this lecture notes, a general trend for the variance of the "raw" PSD estimator can be deducted from (13.19) and (13.20): the variance is proportional to the square of the true spectrum. This can be approximately written as

$$\text{Var}\left[\hat{P}_x(e^{j\theta})\right] \approx (P(e^{j\theta}))^2. \tag{13.21}$$

From this, it can be seen that $\hat{P}(e^{j\theta})$ is not a consistent estimator, since the variance does not converge to zero for increasing N.

13.2 Periodogram improvements

When the PSD is estimated using the squared Fourier transform of the signal, it is referred to as a periodogram. In the following, we will discuss two improvements to the "raw" periodogram, namely, Bartlett's method and Welch's method.

13.2.1 Bartlett's method: average periodogram

The periodogram was determined as an asymptotically unbiased, non-consistent estimator of the PSD. If the window length goes to ∞, the bias reduces, however, the variance does not decrease.

Bartlett proposed a procedure by which the variance can be decreased. The main idea is to split the total signal of length N into k segments of length M. In other words, the original signal $x[n]$ of length N is split into k smaller but equally long non-overlapping signals $x_i[n]$ of length M, for which the respective PSD estimators can be written as

$$\hat{P}_i(e^{j\theta}) = \frac{1}{M}\left|\sum_{n=0}^{M-1} x_i[n]e^{-jn\theta}\right|^2. \qquad \text{for } i = 1, 2, \ldots, k \tag{13.22}$$

The final PSD estimator can then be determined by averaging all the sub-periodograms as

$$\hat{P}_B(e^{j\theta}) = \frac{1}{K}\sum_{i=1}^{K}\hat{P}_i(e^{j\theta}). \tag{13.23}$$

Figure 13.2 shows a graphical representation of the described method.

To understand why this method actually decreases the variance, we first take a look at the estimator expected value. The expected value can be determined as

$$\begin{aligned}
\mathrm{E}\left[\hat{P}_B(e^{j\theta})\right] &= \mathrm{E}\left[\frac{1}{K}\sum_{i=1}^{K}\hat{P}_i(e^{j\theta})\right], \\
&= \frac{1}{K}\sum_{i=1}^{K}\mathrm{E}\left[\hat{P}_i(e^{j\theta})\right], \\
&= \mathrm{E}\left[\hat{P}(e^{j\theta})\right], \\
&= \frac{1}{2\pi}P(e^{j\theta}) * W_B(e^{j\theta}),
\end{aligned} \tag{13.24}$$

which indeed shows a similar relationship as obtained previously; however, because the signal segments are now shorter ($M < N$) the bias increases because the window length of each signal is now shorter. On the other hand, the variance can be determined as

$$\begin{aligned}
\mathrm{Var}\left[\hat{P}_B(e^{j\theta})\right] &= \mathrm{Var}\left[\frac{1}{K}\sum_{i=1}^{K}\hat{P}_i(e^{j\theta})\right], \\
&= \frac{1}{K^2}\,\mathrm{Var}\left[\sum_{i=1}^{K}\hat{P}_i(e^{j\theta})\right], \\
&= \frac{1}{K}\,\mathrm{Var}\left[\hat{P}_i(e^{j\theta})\right] \approx \frac{1}{K}P(e^{j\theta})^2,
\end{aligned} \tag{13.25}$$

which proves that the variance indeed decreases for an increase in the number of signal segments. From this, it can be concluded that Bartlett's method reduces the variance of the periodogram at the cost of an increased bias.

13.2.2 Welch's overlapped segment averaging (WOSA) method

A similar method is Welch's method, also known as Welch's overlapped segment averaging (WOSA) spectral density estimation. In this method, the signal is also split into different segments but, this method allows for overlap between the segments (typically with 50% or 75%

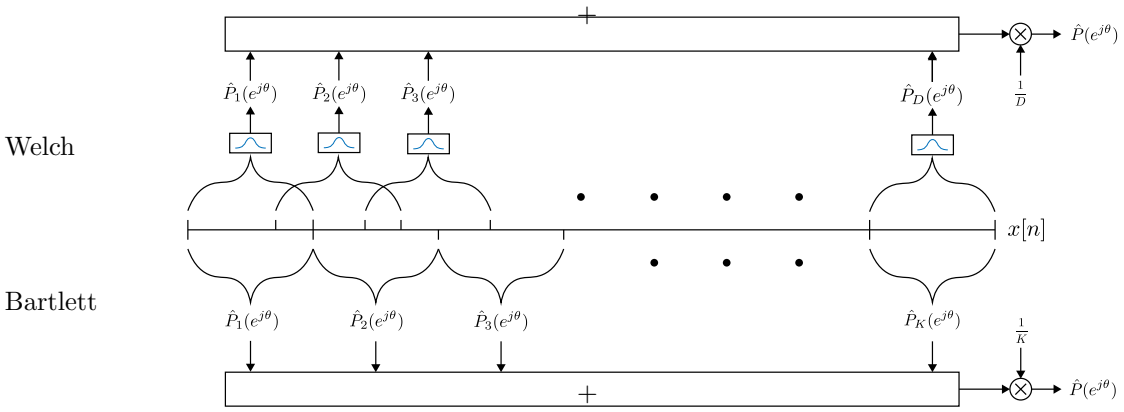


Figure 13.2: Schematic representation of Bartlett's and Welch's method. In Bartlett's method, several periodograms are obtained from non-overlapping signal segments and averaged to obtain the final periodogram. In Welch's method, several periodograms are obtained from overlapping signal segments and averaged to obtain the final periodogram.

overlap); then, these segments are windowed before calculating the individual periodograms, and then averaged, similarly to Bartlett's method. Figure 13.2 shows a graphical visualization of this method. The consequence of this overlap is that the individual segments are now no longer independent, which results in (13.25) not being valid anymore. When applying this method the variance is still decreased, but to a smaller extent compared to Bartlett's method. On the other hand, since the segments are now longer due to the overlap, the bias does not increase as much as with Bartlett's method.

13.3 Correlogram improvements

When the PSD is estimated using the Fourier transform of the estimated autocorrelation function, it is referred to as a correlogram. Hereafter, we describe an improvement of the correlogram known as the Blackman-Tukey correlogram.

13.3.1 Blackman-Tukey method

The basic idea of The Blackman-Tukey correlogram is to apply a symmetric window function $w_M[n]$ of length $2M-1$ to the estimated autocorrelation function, before applying the Fourier transform. The Blackman-Tukey estimate of the PSD $\hat{P}_{BT}(e^{j\theta})$ can be defined as

$$\hat{P}_{BT}(e^{j\theta}) = \sum_{l=-(M-1)}^{M-1} w_M[l]\hat{r}[l]e^{-jl\theta}. \tag{13.26}$$

Here the estimated autocorrelation function is windowed. Note that this window is additional to the triangular window which is inherent in the biased estimation of the autocorrelation function. The reason to include an extra window is that the estimated autocorrelation is very uncertain at the edges of the observation window since at these lags it is calculated using a very limited number of samples. Whereas in the center of the window $\hat{r}[l]$ is computed with almost all samples available. Therefore, by using a suitable window, we can reduce the weight of the autocorrelation lags at the edges of the window, where only a few samples are multiplied and

averaged over. In the frequency domain, the Blackman-Tukey estimator can be interpreted as the convolution between the spectrum of the window and the estimated PSD as

$$\hat{P}_{BT}(e^{j\theta}) = \frac{1}{2\pi}\hat{P}(e^{j\theta}) * W_M(e^{j\theta}). \tag{13.27}$$

Let us now determine the performance of this method. The expected value of the estimator can be determined using (13.13) as

$$\begin{aligned} \mathrm{E}\left[\hat{P}_{BT}(e^{j\theta})\right] &= \frac{1}{2\pi}\mathrm{E}\left[\hat{P}(e^{j\theta})\right] * W_M(e^{j\theta}), \\ &= \frac{1}{2\pi}P(e^{j\theta}) * W_B(e^{j\theta}) * W_M(e^{j\theta}), \\ &\approx \frac{1}{2\pi}P(e^{j\theta}) * W_M(e^{j\theta}), \end{aligned} \tag{13.28}$$

where it is assumed that the Blackman-Tukey window has a length significantly smaller than the length of the autocorrelation function. This assumption also results in the spectrum of the Blackman-Tukey window having wider lobes in comparison to the spectrum of the Bartlett window, which was a direct consequence of the definition of the autocorrelation estimator. Using the definition of the continuous convolution in the frequency domain, the expected value of the estimator can be rewritten as

$$\mathrm{E}\left[\hat{P}_{BT}(e^{j\theta})\right] \approx \frac{1}{2\pi}\int_{-\pi}^{\pi} P(e^{j\phi})W_M(e^{j(\theta-\phi)})d\phi. \tag{13.29}$$

From 13.28, it should be apparent that not all windows can be used. Windows for which $W_M(e^{j\theta})$ have negative values cannot be used, as they could potentially lead to a negative PSD, which is theoretically impossible. Sufficient conditions for a valid autocorrelation window are:

$$W_M(e^{j\theta}) \geq 0, \quad \forall\theta \tag{13.30}$$

or equivalently that the window in the time domain is positive semi-definite, in the sense that

$$\sum_{i=0}^{M-1}\sum_{j=0}^{M-1} a_i w_M[i-j]a_j \geq 0, \tag{13.31}$$

for any vector $\bar{a} = [a_1, a_2, ..., a_M]^T$.

If we consider the window for an increasing length $2M-1$, we would find

$$\lim_{M\to\infty} W_M(e^{j\theta}) = A\delta(\theta), \tag{13.32}$$

which states that the frequency spectrum of the window function converges to a delta pulse. Thus $\hat{P}BT(e^{j\theta})$ is unbiased provided that $A = 1$. This occurs if

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} W_M(e^{j(\theta)})d\theta = \frac{1}{2\pi}\int_{-\pi}^{\pi} A\delta(\theta)d\theta = 1 = w_m[0], \tag{13.33}$$

where the latter passage is due to the basic properties of the Fourier transform.

Combining the previous two equations leads to

$$\mathrm{E}\left[\hat{P}BT(e^{j\theta})\right] \approx P(e^{j\theta})\frac{1}{2\pi}\int -\pi^{\pi} A\delta(\theta)d\theta = w_M[0]P(e^{j\theta}) \tag{13.34}$$

stating that the estimator is asymptotically unbiased if $A = 1$ or equivalently $w_M[0] = 1$. Note that the approximation is only valid in the assumption of very large M. In this case, the main lobe of the window is so narrow, that we can consider $P(e^{j\theta})$ constant within the narrow main lobe.

Although the derivation is complex and beyond the scope of these lecture notes, it can be shown that the variance of the Blackman-Tukey estimator is found as

$$\mathrm{Var}\left[\hat{P}_{BT}(e^{j\theta})\right] \approx \frac{P(e^{j\theta})^2}{N}\left(\sum_{l=-(M-1)}^{M-1} w_m^2[l]\right), \tag{13.35}$$

which shows that the estimator is consistent when $N \to \infty$. There is a compromise made for the length of the Blackman-Tukey window function M. A large value of M will namely decrease the bias of the estimator since the spectrum of the window will approach a delta pulse, but the variance of the estimator will increase, because of the longer summation, including more lags at the edges (more uncertain).

13.4 Summary: "raw" estimators and improvements

The "raw" periodogram and correlogram calculate the PSD by either the square of the absolute windowed signal spectrum or by the Fourier transform of the estimated autocorrelation function, respectively. They produce asymptotically unbiased estimators for an increasing length of the signal. However, there is no way to decrease the variance.

Bartlett's method improves the "raw" periodogram by averaging over multiple PSD estimators, which are calculated for non-overlapping segments of the signal. Using this approach the variance can be decreased by using more segments, but the bias is increased.

Welch's method is similar to Bartlett's method but allows for overlap between the windowed segments. Because the estimators are no longer independent, the variance does not decrease as much, but neither will the bias increase as much.

Blackman-Tukey (BT) method uses a different approach and windows the autocorrelation estimator because this estimator contains a lot of uncertainty at the edges (large lags) due to the limited number of samples used to calculate the autocorrelation at these points. With this operation, the variance now scales with N, i.e., the number of signal samples, and thus the BT is a consistent estimator of the PSD. In the Fourier domain, it can be seen as the convolution of the estimated PSD (which, for the biased estimator, is already a convolution between the true PSD and a triangular window) with the spectrum of the BT window function. A longer window allows for more uncertainty, increasing the variance, but reduces the effects of the convolution, therefore decreasing the bias.

As an example, Figure 13.3 shows the estimated PSD of the signal

$$x[n] = \cos(0.35\pi n) + \cos(0.4\pi n) + 0.25\cos(0.8\pi n) + \epsilon[n], \tag{13.36}$$

where $\epsilon[n]$ is standard Gaussian noise. Several methods are used to estimate the PSD using various design parameters. Please note the important differences and characteristics of each method and see how the parameters affect the spectral estimation. The BT window length is denoted by L instead of M.

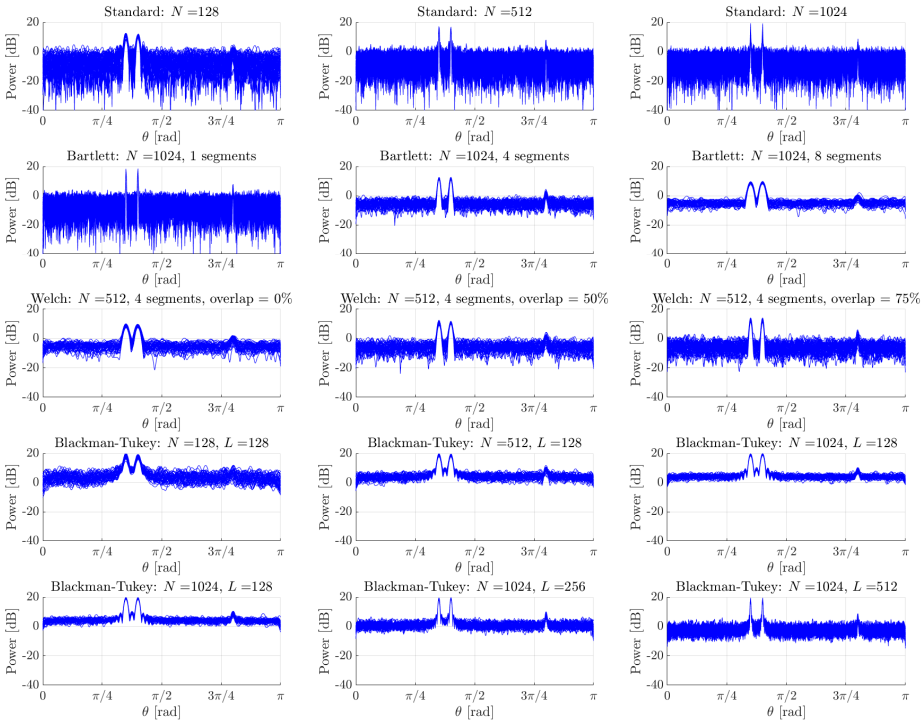


Figure 13.3: The estimated PSD of a signal using multiple methods and design parameters, which are depicted in the plot titles. It should be noted that the Blackman-Tukey method uses a Bartlett (triangular) window of length $2L - 1$.

14

Parametric spectral estimation

Non-parametric spectral estimation methods are based on two major assumptions. First of all, by the calculation of the Fourier transform it is assumed that our finite signal is periodic. The period, in this case, is fixed by the length of the signal N. Because the signal is unknown, it is highly unlikely that this assumption is actually valid. Secondly, the estimated autocorrelation function of the finite signal is zero for absolute lags larger than the signal length. Because the autocorrelation has now also a finite window, the indirect method cannot calculate the exact PSD, limiting the resolution of the estimate.

It would be nice to extrapolate the autocorrelation function for larger lags because this would increase the resolution of the estimated PSD. With extrapolation, we are referring to the process where additional values are estimated or predicted based on the already available information. In other words, the autocorrelation function can be extended or extrapolated by finding a pattern in the available autocorrelation function and by filling the unknown values with the expected values from this pattern.

This pattern or signal model requires us to have some prior information about the signal generating process. Otherwise, it is impossible to determine an accurate signal model that would correspond to the obtained autocorrelation function. After the signal model has been identified, it needs to be fitted to the available autocorrelation function. This means that the unknown parameters of the signal model are estimated according to the available autocorrelation function, and once the parameters are estimated, the signal model can be used to estimate the entire autocorrelation function.

This approach is very different from the non-parametric methods and it is usually referred to as the parametric method, by which parameters of a signal model are estimated such that signal and autocorrelation function can be described by such a model. The parametric approach roughly consists of three steps. First, a signal model needs to be defined. This step is usually the hardest since it would ideally require prior knowledge of the random process. If this knowledge is not available, then several models can be fitted to the data after which the most optimal is chosen. The definition of optimal depends on the evaluation criterion. Once the model is defined, the second step is to estimate the parameters of the model based on the available data (finite-length signal or autocorrelation function). Finally, the last step is to obtain the PSD using the signal model.

Rational signal models

The first step in estimating the PSD is finding a signal model. How do we define such a signal model? One way is to use spectral factorization, i.e., modeling the signal as obtained by filtering a

white Gaussian process $i[n]$ by a linear time-invariant (LTI) filter. A white Gaussian process has the characteristic that the PSD is constant. If we transform this constant spectrum by applying filters (low-pass, high-pass, etc.), it is possible to shape the PSD in a controlled way. In other words, a constant spectrum can be filtered such that the resulting spectrum provides us with a good estimate of the PSD of the observed random signal. Of course, we cannot directly assess the accuracy of the resulting spectrum since this is unknown, but we can compare the estimated with the determined autocorrelation functions. Although in principle there is an infinite number of possibilities for the filter, we typically restrict the choice to finite impulse response (FIR) and infinite impulse response (IIR) filters. This is equivalent to modeling the random signal as an ARMA process.

14.1 AR spectral estimation

As we have seen in Section 4.3, an autoregressive process of order p is defined by the following difference equation

$$x[n] = i[n] - a_1 x[n-1] - a_2 x[n-2] - \ldots - a_p x[n-p], \tag{14.1}$$

where $i[n]$ is the input white noise and a_i are the filter coefficients.

The transfer function can be found as

$$H(z) = \frac{1}{1 + \sum_{p=1}^{P} a_p z^{-p}} = \frac{1}{A(z)}. \tag{14.2}$$

The autocorrelation function is given by

$$r[l] = \begin{cases} \sigma_i^2 - \sum_{k=1}^{p} a_k r_x[|l| - k] & \text{for} l = 0 \\ -\sum_{k=1}^{p} a_k r_x[|l| - k] & \text{for } |l| > 0 \end{cases} \tag{14.3}$$

When a windowed signal is observed and the autocorrelation function is estimated as $\hat{r}_x[l]$, the Yule-Walker equations can be used to estimate parameters $\hat{a}_1$,..., $\hat{a}_p$ and $\hat{\sigma}_i^2$. Since there are $p+1$ unknowns, we require $p+1$ equations to solve the Yule-walker equations, which is equivalent to using $p+1$ estimated correlation lags. The Yule-Walker equations for an AR process are linear, and thus can be written in a matrix form as

$$\begin{bmatrix} \hat{r}_x[0] & \hat{r}_x[1] & \hat{r}_x[2] & \cdots & \hat{r}_x[p] \\ \hat{r}_x[1] & \hat{r}_x[0] & \hat{r}_x[1] & \cdots & \hat{r}_x[p-1] \\ \hat{r}_x[2] & \hat{r}_x[1] & \hat{r}_x[0] & \cdots & \hat{r}_x[p-2] \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \hat{r}_x[p] & \hat{r}_x[p-1] & \hat{r}_x[p-2] & \vdots & \hat{r}_x[0] \end{bmatrix} \begin{bmatrix} 1 \\ \hat{a}_1 \\ \hat{a}_2 \\ \vdots \\ \hat{a}_p \end{bmatrix} = \begin{bmatrix} \hat{\sigma}_i^2 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{14.4}$$

Here we assume the signal $x[n]$ to be real, and thus a symmetric autocorrelation function of the form $\hat{r}_x[l] = \hat{r}_x[-l]$. Solving this system of equations, the unknown filter coefficient $\hat{a}_1, ..., \hat{a}_p$, and the unknown variance of the input noise, $\hat{\sigma}_i^2$, can be estimated by least-square linear estimation. Note that the autocorrelation matrix has a Toeplitz structure, which is a matrix with constant diagonals.

Finally, the PSD of an AR process is given by

$$P_X(e^{j\theta}) = \frac{\sigma_i^2}{|1 + a_1 e^{-j\theta} + a_2 e^{-j2\theta} + \ldots + a_p e^{-jp\theta}|^2}. \tag{14.5}$$

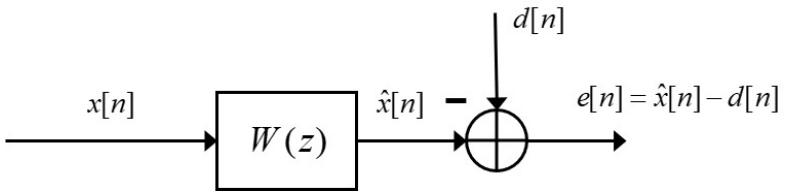


Figure 14.1: General scheme of a Wiener filter.

Estimation of an AR spectrum

In order to estimate the spectrum of an AR(p) process, the first step is to write down the Yule-Walker equations for a certain model order p. This model order can be freely chosen, but its performance can be measured using the metrics that will be discussed later on. To obtain the Yule-Walker equations, first, the autocorrelation function of the signal should be estimated for lags $|l| \leq p$. From these equations, the AR coefficients $\hat{a}_i$ and the innovation variance $\hat{\sigma}_i^2$ can then be calculated. With these parameters, the analytical AR PSD can be estimated using (14.5). For practical implementations, however, the PSD is often calculated using the discrete Fourier transform, zero-padded to length L, through

$$\hat{P}[k] = \frac{\hat{\sigma}_i^2}{\left|\sum_{i=0}^{p} \hat{a}_i e^{-jik\frac{2\pi}{L}}\right|^2}. \tag{14.6}$$

Calculation of AR parameters via 1-step linear predictor

An alternative way to calculate the AR model parameters $a_1, a_2, ..., a_p$ and σ_i^2 is to use a 1-step linear predictor. This can be implemented as an FIR Wiener filter. As shown in Figure 14.1, the general goal of Wiener filters is to filter a signal $x[n]$ such as to minimize the error between the filtered signal $\hat{x}[n]$ and the desired signal $d[n]$ according to the minimum mean square error (MMSE) criterion.

The general FIR solution for this Wiener filter is found by finding the set of parameters that minimizes a cost function defined as:

$$J = \mathrm{E}\left[e^2[n]\right] = \mathrm{E}\left[(d[n] - \hat{x}[n])^2\right] = \mathrm{E}\left[d[n]^2\right] - \mathbf{w}^T\mathbf{r_{dx}} - \mathbf{r_{dx}}^T\mathbf{w} + \mathbf{w}^T\mathbf{R_x}\mathbf{w} \tag{14.7}$$

where $\mathbf{w}$ is the vector composed of the filter coefficients, $\mathbf{r_{dx}}$ is the cross-correlation between the desired and observed signals, and $\mathbf{R_x}$ is the autocorrelation matrix of the observed signal. The optimal filter coefficients $\mathbf{w}_{\mathrm{opt}}$ can be found as the solution to the following minimization problem:

$$\mathbf{w}_{\mathrm{opt}} = \arg\min_{\mathbf{w}} \; J = \arg\min_{\mathbf{w}} \; (\mathrm{E}\left[e^2[n]\right]), \tag{14.8}$$

By setting the gradient of J equal to zero, the so-called normal equations are found as

$$\frac{dJ}{d\mathbf{w}} = 2(\mathbf{r_{dx}} - \mathbf{R_x}\mathbf{w}) = 0. \tag{14.9}$$

From (14.9), the solution of the FIR Wiener filter is found as

$$\mathbf{w}_{\mathrm{opt}} = \mathbf{R_x}^{-1}\mathbf{r_{dx}}. \tag{14.10}$$

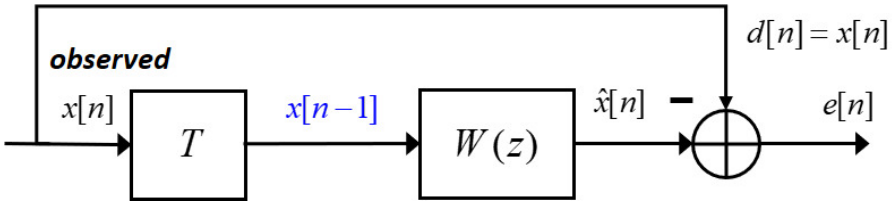


Figure 14.2: Schematic representation of a Wiener filter for linear 1-step prediction.

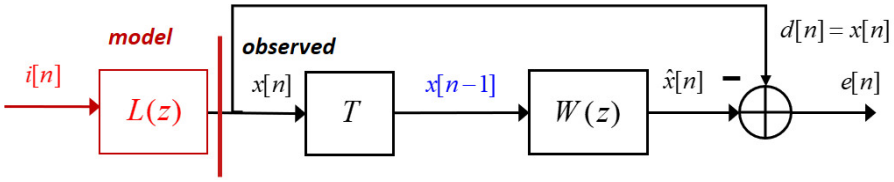


Figure 14.3: Interpretation of a Wiener filter for 1-step linear prediction as the inverse of the innovation filter.

Moreover, the filter error can be calculated as

$$J_{\text{FIR}} = r_d[0] - \sum_{i=0}^{N-1} w_{\text{opt}}[i] r_{dx}[i] = r_d[0] - \mathbf{w}_{\text{opt}}^T \mathbf{r_{dx}}. \tag{14.11}$$

The goal of a 1-step linear predictor is to estimate the next value of the signal based on a linear combination of the past values of the signal. In this case, as schematically shown in Figure 14.2, the input to the filter is a delayed version of the observed signal and the desired signal is the observed signal itself. For a 1-step linear predictor, the delay is $T = 1$.

The solution of this filter is found by using (14.10) and substituting for $\mathbf{R_x}$ a matrix obtained using autocorrelation lags from 0 up to $N-2$, where N is the signal length; this means using the original signal as the desired signal. While for the cross-covariance $\mathbf{r_{dx}}$ we use the autocorrelation of $x[n]$, since in this case the input to the filter is a shifted version of the desired signal, and both are the observed signal $x[n]$; thus $\mathbf{r_{dx}} = \mathbf{r_x}$ using lags from 1 up to $N-1$. This can be described in formulas as

$$\mathbf{R_x} = \text{Toeplitz}[r_x[0], r_x[1], ..., r_x[N-2]] \tag{14.12}$$

$$\mathbf{r_x} = [r[1], r[2], ..., r[n-1]]^T. \tag{14.13}$$

For an AR model, our signal is given by the sum of an unpredictable part and a predictable part composed of filtered past signal samples, that is $x[n] = i[n] + \hat{x}[n] = i[n] - a_1 x[n-1] + ... - a_p x[n-p]$.
Then, since the filter error is given by $x[n] - \hat{x}[n] = i[n]$, we are basically left with white noise. As a result the expected squared error is $E\{(x - x[n])^2\} = \sigma_i^2$.

When we want to use this filter to estimate the AR parameters $a_1, a_2, ..., a_p$ and σ_i^2, we should bare in mind that the observed signal is actually modeled as the output of an LTI with p poles and no zeros driven by white noise, as shown in Figure 14.3.

This means that the optimal filter $W(z)$ is actually the inverse of $L(z)$. In fact, $W(z)$ can be interpreted as a whitening filter that takes as input a random signal constituted by a predictable part and an unpredictable part and outputs white noise, that is the unpredictable part. Because of this, to obtain our AR model parameters, we need to invert the sign of the filter coefficients,

while the input noise variance can be simply found by applying the formula for the filter error, as given below

$$\mathbf{w}_{\text{opt}} = [w_1, w_2, ..., w_{N-1}]^T = [-\hat{a_1}, -\hat{a_2}, ..., -\hat{a_p}]^T. \tag{14.14}$$

$$J = r_x[0] - \sum_{i=0}^{N-1} w_{opt}[i] r_x[i] = r_x[0] - \mathbf{w}_{\text{opt}}^T \mathbf{r}_{\mathbf{x}} = \hat{\sigma_i}^2. \tag{14.15}$$

Although Wiener filtering is beyond the scope of this course, we have discussed here the 1-step linear predictor as a convenient way to find the AR model parameters.

14.2 MA spectral estimation

A moving-average process of order q is defined by the following difference equation

$$x[n] = i[n] + b_1 i[n-1] + b_2 i[n-2] - \ldots - b_q i[n-q]. \tag{14.16}$$

where $i[n]$ is the input white noise and b_i are the filter coefficients.

The transfer function can be found as

$$H(z) = 1 + \sum_{q=1}^{Q} b_q z^{-q} = B(z). \tag{14.17}$$

The autocorrelation function is given by

$$r_x[l] = \begin{cases} \sigma_i^2 \sum_{k=|l|}^{q} b_k b_{k-|l|}, & \text{for } 0 \leq |l| \leq q \\ 0. & \text{otherwise} \end{cases} \tag{14.18}$$

Finally, the PSD of an AR process is given by

$$P_x(e^{j\theta}) = P_I(e^{j\theta}) H(e^{j\theta}) H^*(e^{j\theta}) = \sigma_i^2 |1 + b_1 e^{-j\theta} + \ldots + b_q e^{-jq\theta}|^2. \tag{14.19}$$

Estimation of an MA spectrum

In order to estimate the spectrum of an MA(q) process three methods exist. First, the windowed estimated autocorrelation function can be used to estimate the PSD directly using the Wiener-Khinchin relationship. For MA processes, it is known that the analytical autocorrelation function is non-zero for lags $|l| < q$ where q denotes the process order. Using this fact, the autocorrelation function can be approximated by windowing the estimated autocorrelation function to the expected non-zero lags. From this, the PSD can be estimated as

$$\hat{P}(e^{j\theta}) = \sum_{l=-q}^{q} \hat{r}[l] e^{jl\theta}. \tag{14.20}$$

This description can be compared with the Blackman-Tukey method of the previous section using a rectangular window of length $2q + 1$. It should also be noted that care should be taken whilst performing this estimation because model mismatch can lead to a negative PSD at some relative frequencies, which should not be possible by the definition of the PSD.

The second approach involves estimating the MA model parameters $\hat{b}_i$ and the innovation variance $\hat{\sigma}_i^2$ from the estimated autocorrelation function of the signal. This estimation is a non-linear estimation problem. However, once the parameters have been obtained and the PSD is calculated using (14.19), the PSD is guaranteed to be non-negative.

A third approach is aimed more specifically at the practical implementation of the above-mentioned methods. Here the analytical description of the PSD is approximated using (a zero-padded version of) the DFT or FFT.

Example 14.1

Suppose we have an estimate of the autocorrelation function of a random signal $x[n]$ at lags $0, \pm 1, \pm 2$, as given below

$$r_x[0] = \tfrac{49}{36}, \quad r_x[\pm 1] = \tfrac{1}{3}, \quad r_x[\pm 2] = -\tfrac{1}{3}.$$

- Assuming a MA(1) model, find the parameters σ_i^2, and b_1, calculate the spectrum $P_{MA1}(e^{j\theta})$ and plot it.
- Assuming a MA(2) model, and knowing that $\sigma_i^2 = 1$, find the parameters b_1 and b_2, calculate the spectrum $P_{MA2}(e^{j\theta})$ and plot it.
- Now use the correlagram method, assuming that $r_x[l]$ is zero for $l \geq 3$. How does the obtained spectrum $P_{corr}(e^{j\theta})$ compare with the previous ones?

Solution.

Assuming a MA(1) model, we can find the parameters σ_i^2, and b_1 by using the Yule-Walker equations in (14.18). Since we have 2 unknowns, we only need 2 equations, and thus we only need to use the autocorrelation values at 0 and 1. Because of the constraints on the innovation filter, we assume $b_0 = 1$, thus obtaining

$$\begin{cases} r_x[0] = \frac{49}{36} = \sigma_i^2 + \sigma_i^2 b_1^2 \\ r_x[\pm 1] = \frac{1}{3} = \sigma_i^2 b_1 \end{cases}$$

Solving the above system of equations, we obtained two possible solutions for b_1, which are is $b_1 \approx 0.26$ or $b_1 \approx 3.82$. Since we want the innovation filter to be minimum-phase, we choose $b_1 = 0.26$, for which the zero of the system is within the unit circle. With this choice, we obtain $\sigma_i^2 \approx 1.28$. Using (14.19), we can now find the PSD as

$$\begin{aligned} P_{MA1}(e^{j\theta}) &= 1.28|1 + 0.26e^{-j\theta}|^2 \\ &= 1.28|1 + 0.26\cos(\theta) - j \cdot 0.26\sin(\theta)|^2 \\ &= 1.28\left((1 + 0.26\cos(\theta))^2 + 0.26^2\sin^2(\theta)\right) \\ &= 1.28(1.07 + 0.52\cos(\theta)) \end{aligned} \tag{14.21}$$

For an MA(2) model, we need to estimate 3 parameters and thus we need 3 equations. Similar to the equations above, we can write:

$$\begin{cases} r_x[0] = \frac{49}{36} = \sigma_i^2 + \sigma_i^2 b_1^2 + \sigma_i^2 b_2^2 \\ r_x[\pm 1] = \frac{1}{3} = \sigma_i^2(b_1 + b_2 b_1) \\ r_x[\pm 2] = -\frac{1}{3} = \sigma_i^2 b_2 \end{cases}$$

Solving the above equations with $\sigma_i^2 = 1$, gives $b_1 = \frac{1}{2}$ and $b_2 = -\frac{1}{3}$. The resulting PSD is:

$$P_{MA2}(e^{j\theta}) = \left|1 + \frac{1}{2}e^{-j\theta} - \frac{1}{3}e^{-j2\theta}\right|^2 \tag{14.22}$$

Using the correlogram, we need to take the Fourier transform of the autocorrelation function :

$$\begin{aligned} P_{corr}(e^{j\theta}) &= \sum_{l=-\infty}^{\infty} r_x[l]e^{-jl\theta} = \sum_{l=-2}^{2} r_x[l]e^{-jl\theta} = \\ &= -\frac{1}{3}e^{j2\theta} + \frac{1}{3}e^{j\theta} + \frac{14}{9} + \frac{1}{3}e^{-j\theta} - \frac{1}{3}e^{-j2\theta} = \\ &= \frac{49}{36} + \frac{2}{3}\cos(\theta) - \frac{2}{3}\cos(2\theta) \end{aligned} \tag{14.23}$$

In the following figure, the estimated power spectral densities are plotted in the fundamental interval. Not surprisingly, the obtained PSD by an MA(2) model or by the correlogram are the same. In fact, for an MA(2) process, the autocorrelation function is only non zero for $|l| \leq 2$. When calculating the correlogram, we assumed that the autocorrelation was zero for $|l| \geq 3$. This assumption is equivalent to assuming a MA(2) generating process for $x[n]$. When we model $x[n]$ as a MA(1) process, the peak of the obtained PSD is quite different, but the valley gets close to the other estimates.

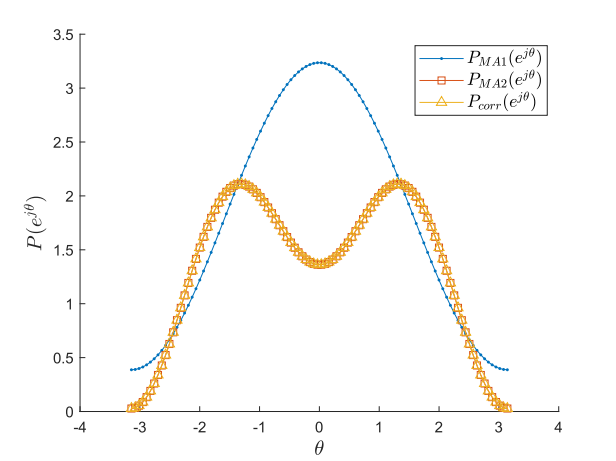


Figure 14.4: Estimates of the PSD of random signal *x[n]* by assuming a MA(1) or MA(2) generating process, and by using the correlogram method.

14.3 ARMA spectral estimation

A general autoregressive moving-average process of order p, q is defined by the following difference equation

$$x[n] = i[n] + b_1 i[n-1] + b_2 i[n-2] + \ldots + b_q i[n-q] \\ - a_1 x[n-1] - a_2 x[n-2] - \ldots - a_p x[n-p], \quad (14.24)$$

where $i[n]$ is the input white noise and a_i, b_i are the filter coefficients.

The transfer function can be found as

$$H(z) = \frac{1 + \sum_{q=1}^{Q} b_q z^{-q}}{1 + \sum_{p=1}^{P} a_p z^{-p}} = \frac{B(z)}{A(z)}. \quad (14.25)$$

The autocorrelation function is given by

$$r_x[l] = \begin{cases} \sigma_i^2 \sum_{k=|l|}^{q} b_k h[k-|l|] - \sum_{k=1}^{p} a_k r_x[|l|-k], & \text{for } 0 \leq |l| \leq q \\ -\sum_{k=1}^{p} a_k r_x[|l|-k]. & \text{for } |l| > q \end{cases} \quad (14.26)$$

Finally, the PSD of an AR process is given by

$$P_x(e^{j\theta}) = P_I(e^{j\theta}) H(e^{j\theta}) H^*(e^{j\theta}) = \sigma_i^2 \frac{|1 + b_1 e^{-j\theta} + \ldots + b_q e^{-jq\theta}|^2}{|1 + a_1 e^{-j\theta} + \ldots + a_p e^{-jp\theta}|^2}. \quad (14.27)$$

Estimation of an ARMA(p,q) spectrum

The ARMA process is a combination of the AR(p) and MA(q) processes. The ARMA coefficients can be estimated as follows.

First, the AR coefficients are estimated. The autocorrelation function of the ARMA process consists of the AR(p) autocorrelation function added to the MA(q) autocorrelation function. The former spans over all lags whereas the influence of the latter is bounded by its model order $|l| \leq q$. For lags $|l| > q$ only the influence of the AR(p) process is visible and from this region of the autocorrelation function, the AR coefficients can be determined using the methods described previously.

Secondly, the goal is to remove the influence of the AR(p) process such that only the MA(q) process remains. Several methods exist to perform this inverse filtering and are often denoted as deconvolution or spectral subtraction, however, this is beyond the scope of these lecture notes.

Once the MA(q) process has been separated, the MA parameters can be determined using the methods described earlier. After this, both the AR and MA parameters have been estimated and the analytical PSD can be determined using (14.27) or by approximating it with the DFT or FFT.

However, any spectrum can be approximated by an AR model, provided that a sufficiently high filter order is chosen. Thus, since it is practically and computationally easier, an AR model is typically assumed, unless we have a priori knowledge of the signal generating process, for which we know that the process is MA or ARMA.

14.4 Model selection

As mentioned at the beginning of this section, choosing the right model is the hardest design choice in parametric estimation. There is an infinite number of options to explore each with

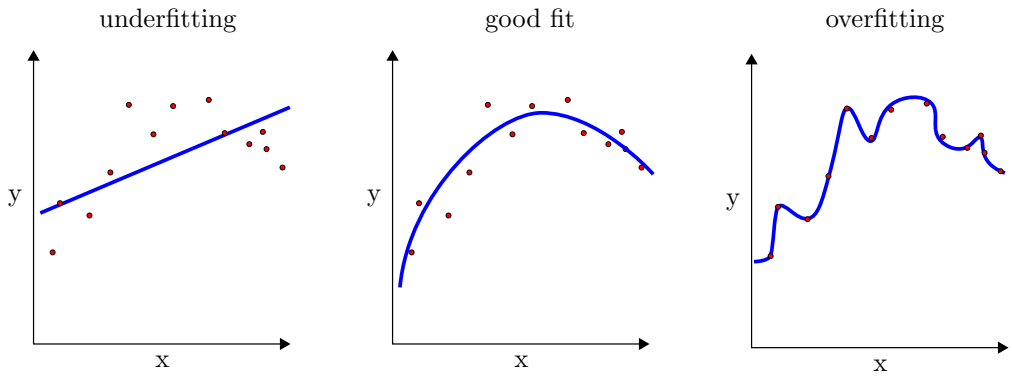


Figure 14.5: Three different fits of data using a polynomial, differentiating between the underfitting, overfitting, and having an appropriate fit.

different complexity. But how can we know which model is the best? As an analogy, we present a well-known regression example, where the goal is to describe a set of points as well as possible using a function. Figure 14.5 shows three lines that were fitted to some data. Intuitively, the middle plot seems to represent the best fit; the left plot does not capture the essence of the samples (large error), while the right plot tunes to the specific data points, and hence might tune to the noise rather than the actual model generating the data. If there would be new data, which is similarly distributed to the already present data, then the fit of the right plot would be inadequate. Thus, when choosing a model it can either be too simplistic or too complex. These phenomena are respectively called underfitting and overfitting.

In machine learning, underfitting and overfitting are major problems. Nonetheless, several methodologies have been developed to limit this undesired behavior. One of these includes a more realistic optimization scheme. Instead of optimizing only the fit of the function, now the complexity is also considered. This is called regularization. The optimization algorithm needs to find a balance between a very complex model with a perfect fit and a very simple model with a bad fit.

With the spectral estimation techniques presented above, similar reasoning is used. As the performance criterion of the best model, not only the error should be considered, but also the model complexity. This criterion can be implemented in several ways, some of which are discussed hereafter.

14.4.1 Residual error

The most basic model selection criterion is the residual error $\hat{\sigma}_r^2$. This error is the variance of the residual signal, defined as $\hat{x} - x$, which is the difference between the estimated and true signal. By the definition of the variance, this is also commonly known as the mean-squared error.

14.4.2 Coefficient of determination

A performance metric that is commonly used in the field of statistics and which uses the residual error is the coefficient of determination, denoted by R^2. This coefficient is defined as

$$R^2 = 1 - \frac{\sum_i (x_i - \hat{x}_i)^2}{\sum_i (x_i - \mu_x)^2}. \tag{14.28}$$

The coefficient can be understood as one minus the ratio between the variance of the residuals and the variance of the data. This coefficient is upper-bounded by the value of 1 since the variance cannot be negative. For perfect models, the variance of the error equals 0 and therefore the coefficient of determination attains its upper bound. A lower value of R^2 denotes a worse model fit. Although the coefficient of determination is denoted using the square operator, it should be noted that its value can actually be negative, denoting a very poor fit of the model.

14.4.3 Final prediction error

The final prediction error (FPE) is a metric that adds an additional factor to the variance of the residual error $\hat{\sigma}_P^2$. This additional factor depends on the signal length N and the number of parameters in the model P. For example, for an AR model $P = p + 1$, where p is the AR model order and 1 is given by the input noise variance. The FPE provides some compensation between the error (which decreases as the model order increases) and the number of model parameters, by applying a penalty for increasing P. The FPE is defined as

$$\mathrm{FPE}(P) = \frac{N + (P+1)}{N - (P+1)} \hat{\sigma}_P^2. \tag{14.29}$$

The best model is chosen as the one that yields the lowest FPE.

14.4.4 Akaike's information criterion

Another model selection criterion is Akaike's information criterion (AIC). This criterion is based on the Kullback-Leibler divergence and is defined as

$$\mathrm{AIC}(P) = N \cdot \ln(\hat{\sigma}_P^2) + 2P. \tag{14.30}$$

This criterion, however, is very likely to lead to overfitting when the number of samples is limited. Therefore a corrected version of Akaike's information criterion is introduced as

$$\mathrm{AICc}(P) = N \cdot \ln(\hat{\sigma}_P^2) + 2P + \frac{2P(P+1)}{N - P - 1}. \tag{14.31}$$

The best model order is chosen as the one that yields the lowest AIC (or AICc).

Although these criteria are useful to compare similar models of different orders, they are not suited to compare different model structures. This is a much more complex task that is beyond the scope of these lecture notes.

Example 14.2

Suppose we have some data that can ideally be modeled as

$$y[n] = ax^2[n] + bx[n] + c + w[n], \tag{14.32}$$

where a, b and c denote the true model parameters and $w[n]$ is Gaussian distributed noise.

The following figure shows the observed data, the true underlying model, and estimates of the underlying model. These estimates correspond to linear equations where the order equals the largest power of the input signal in the model. Order 0 corresponds to only a constant signal.

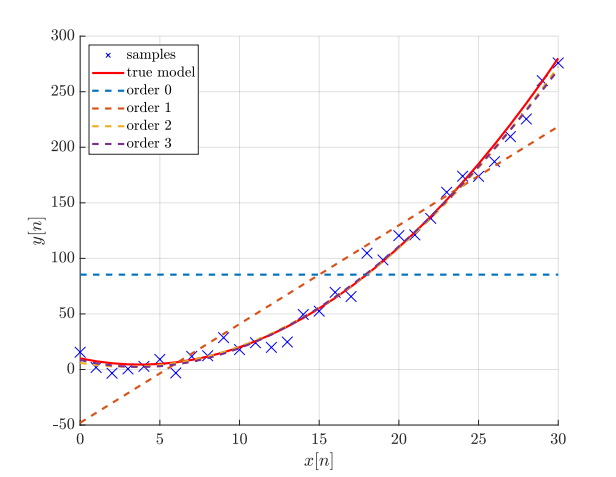


Figure 14.6: Observed data with the true underlying data generation model. Furthermore, several orders of polynomials that try to fit the data are plotted.

To infer the correct model order, we can calculate different model performance criteria for different model orders, as shown in the following table. The residual error and the coefficient of determination do not take the number of parameters into account and therefore prefer the most complex models. All other performance metrics correct for the number of parameters and also prefer the order of the true underlying model.

Order	P	$\hat{\sigma}_P^2$	R^2	FPE(P)	$AIC(P)$	$AICc(P)$
0	1	7321	0	7808	277.9	278.0
1	2	791	0.8919	900	210.9	211.3
2	3	64.1	0.9912	**77.9**	**135.0**	**135.8**
3	4	**62.7**	**0.9914**	81.2	136.3	137.7

Part IV

Detection Theory

15

Detection theory

15.1 Introduction

In Part. II, we addressed estimation of an unknown and *continuous* parameter from a set of observations $\{x_0, x_1, \ldots, x_{N-1}\}$. In detection theory, on the other hand, we assume that the unknown parameter is a discrete variable. The individual values of the discrete variable, which can take M distinct values are called hypotheses and labeled $\mathcal{H}_0, \mathcal{H}_1, \ldots, \mathcal{H}_{M-1}$. As in estimation theory, the observations are obtained via a probabilistic mapping $p(\mathbf{x}|\mathcal{H}_i)$, depending on which hypothesis is true. Based on the observations, we wish to hypothesize which of the M hypothesis caused our observations.

In the following, the focus will be on binary hypothesis testing. The two hypotheses, $\mathcal{H}_0$ and $\mathcal{H}_1$, are referred to as the *null* and *alternative* hypotheses, respectively. Typical examples of binary hypothesis problems are communication systems where a receiver has to decide whether the received bit was 0 or 1 based on some noisy observations or in radars where the presence or absence of a target echo in a noisy signal has to be hypothesized. Depending on whether $\mathcal{H}_0$ or $\mathcal{H}_1$ is true, the observations have the following distributions:

$$\begin{aligned} \mathcal{H}_0: &\quad \mathbf{x} \sim p(\mathbf{x}|\mathcal{H}_0); \\ \mathcal{H}_1: &\quad \mathbf{x} \sim p(\mathbf{x}|\mathcal{H}_1). \end{aligned} \tag{15.1}$$

By assessing the hypotheses, we partition the observation space into two disjoint parts, $\mathcal{X}_0$ and $\mathcal{X}_1$, with $\mathcal{X}_0 \cup \mathcal{X}_1 = \mathbb{R}^N$. Depending on the part of the observation space in which the observation is located in, we will decide on $\mathcal{H}_0$ or $\mathcal{H}_1$. Figure 15.1 depicts the mapping from a hypothesis to an observation in the observation space and forming a decision based on the partitioning of the observation space. When the hypotheses map into a joint set in the observation space, we face the problem that both hypotheses can cause the same observation. By deciding on which hypothesis is true, we run the risk of making a wrong decision, i.e., we decide for $\mathcal{H}_1$ even $\mathcal{H}_0$ was true or vice versa.

In general, there are many ways to partition the observation space. In detection theory, we seek an optimal decision rule, i.e., a partitioning of the observation space, by defining various optimal criteria. Since the two hypothesis can cause the same observations, four different events can happen when making a decision. The four possible events are summarized in Table 15.1.

To evaluate the performance of a binary hypothesis test, we define two quantities: the *probability of detection* P_D, i.e., the probability of choosing $\mathcal{H}_1$ when $\mathcal{H}_1$ is true, and the *probability of false alarm* P_F, which is the probability choosing $\mathcal{H}_1$ even though $\mathcal{H}_0$ is true. The two quantities

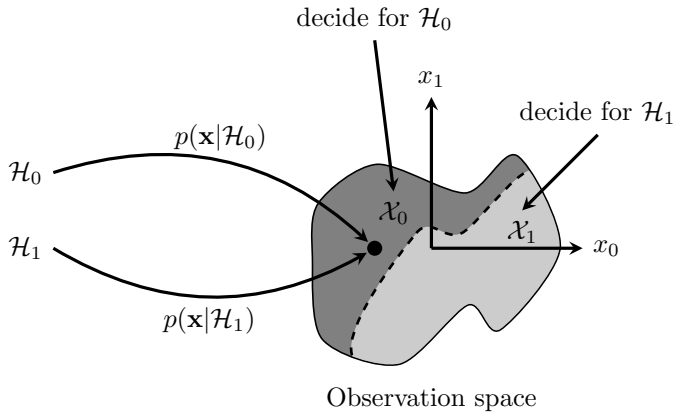


Figure 15.1: Binary hypothesis testing problem

	True hypothesis	
Decision	$\mathcal{H}_0$	$\mathcal{H}_1$
$\mathcal{H}_0$	true negative	false negative/miss/type II error
$\mathcal{H}_1$	false positive/false alarm/type I error	true positive/detection

Table 15.1: Four possible scenarios of binary hypothesis testing.

are given by

$$P_{\mathrm{D}} = \int_{\mathcal{X}_1} p(\mathbf{x}|\mathcal{H}_1) d\mathbf{x}, \tag{15.2}$$

and

$$P_{\mathrm{F}} = \int_{\mathcal{X}_1} p(\mathbf{x}|\mathcal{H}_0) d\mathbf{x}. \tag{15.3}$$

Typically these two quantities are plotted against each other, as shown in Figure 15.2. The interior contains a feasible region for pairs $(P_{\mathrm{F}}, P_{\mathrm{D}})$ that can be achieved. The feasible region always contains the points $(0, 0)$ and $(1, 1)$. These two points are achieved by assigning the entire observation space to either $\mathcal{X}_0$ or $\mathcal{X}_1$. The top-left and the bottom-right corner are only feasible if the values of $\mathbf{x}$ under $\mathcal{H}_0$ and $\mathcal{H}_1$ do not overlap. The top-left corner $(P_{\mathrm{F}} = 0, P_{\mathrm{D}} = 1)$ is the utopia point we wish to achieve by our test since it corresponds to an error-free performance. Conversely, the bottom right corner $(P_{\mathrm{F}} = 1, P_{\mathrm{D}} = 0)$ correspond to poor performance since it always decides for the wrong hypothesis. The feasible region below the line $P_{\mathrm{F}} = P_{\mathrm{D}}$ (dashed line) is characterized by the fact that P_{F} is larger than P_{D}. However, swapping the decision regions $\mathcal{X}_0$ and $\mathcal{X}_1$ turns this poor performance into good performance.

15.2 Neyman-Pearson Test

In Figure 15.2, the feasible region for pairs $(P_{\mathrm{F}}, P_{\mathrm{D}})$ is shown. The upper boundary of this region is of particular interest since from all feasible pairs $(P_{\mathrm{F}} = \alpha, P_{\mathrm{D}})$, it has the largest P_{D}. This upper boundary is achieved by the Neyman-Pearson test, which test partitions the observation

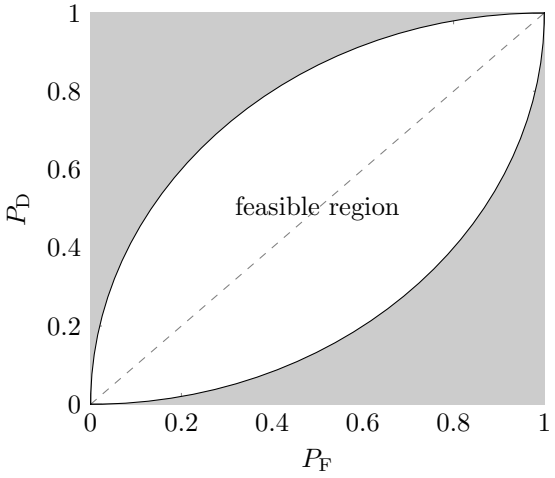


Figure 15.2: Feasible regions for binary hypothesis test. The bisect

space such that for an upper bounded $P_F \leq \alpha$, P_D is maximized. The Neyman-Pearson test has the form of a likelihood test, which is given as

$$L(\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{H}_1)}{p(\mathbf{x}|\mathcal{H}_0)} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \lambda. \tag{15.4}$$

In the likelihood test, we compare the ratio of the PDFs for the observation $\mathbf{x}$ against a threshold and decide that $\mathcal{H}_1$ is true if the likelihood ratio exceeds this threshold. Otherwise, we decide that $\mathcal{H}_0$ holds. From the definition of the likelihood test, it follows that the decision regions $\mathcal{X}_1(\lambda)$ and $\mathcal{X}_0(\lambda)$ are given by

$$\mathcal{X}_1(\lambda) = \{\mathbf{x} : L(\mathbf{x}) \geq \lambda\}, \tag{15.5}$$

and

$$\mathcal{X}_0(\lambda) = \{\mathbf{x} : L(\mathbf{x}) < \lambda\}. \tag{15.6}$$

We show that the likelihood ratio test maximizes P_D for an upper-bounded P_F by looking at the case for $\mathbf{x} \in \mathbb{R}^2$. In Figure 15.3, the two decision regions $\mathcal{X}_0(\lambda)$ and $\mathcal{X}_1(\lambda)$ are shown. Let D denote the boundary between the regions for which $P_F \leq \alpha$ is satisfied. Now let us introduce another decision boundary D' with decision regions $\mathcal{X}_1'$ and $\mathcal{X}_0'$. The new decision regions are obtained by assigning the region $\mathcal{X}^+$ (red shaded) from $\mathcal{X}_0(\lambda)$ to $\mathcal{X}_1(\lambda)$ and $\mathcal{X}^-$ (blue shaded) from $\mathcal{X}_1(\lambda)$ to $\mathcal{X}_0(\lambda)$. The region $\mathcal{X}^+$ and $\mathcal{X}^-$ are chosen such that they have the same probability under $\mathcal{H}_0$ or equivalently

$$\int_{\mathcal{X}^+} p(\mathbf{x}|\mathcal{H}_0) d\mathbf{x} = \int_{\mathcal{X}^-} p(\mathbf{x}|\mathcal{H}_0) d\mathbf{x}. \tag{15.7}$$

Thus, the probability of false alarm for the decision regions $\mathcal{X}_1'$, denoted by P_F', is identical to P_F. To show that P_D', the probability of detection for decision region $\mathcal{X}_1(\lambda)$, has decreased compared to P_D with boundary D, we express the P_D' as

$$P_D' = \int_{\mathcal{X}_1'} p(\mathbf{x}|\mathcal{H}_1) d\mathbf{x} = \int_{\mathcal{X}_1'} p(\mathbf{x}|\mathcal{H}_1) \frac{p(\mathbf{x}|\mathcal{H}_0)}{p(\mathbf{x}|\mathcal{H}_0)} d\mathbf{x} = \int_{\mathcal{X}_1'} L(\mathbf{x}) p(\mathbf{x}|\mathcal{H}_0) d\mathbf{x}. \tag{15.8}$$

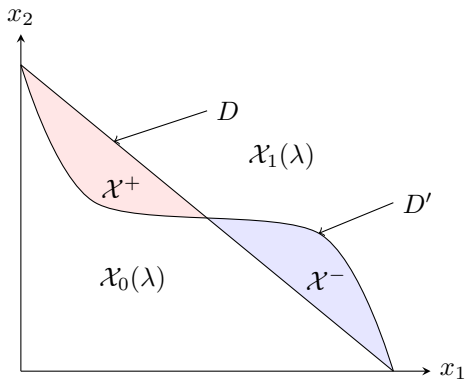


Figure 15.3: Optimality of the Neyman Pearson test. The decision boundary D is chosen such that $L(\mathbf{x}) = \lambda$. Any decision region D' different from D that has the same P_{F} decreases P_{D}.

Since $\mathcal{X}_1' = \mathcal{X}_1(\lambda) \cup \mathcal{X}^+ \setminus \mathcal{X}^-$ and the individual sets are disjoint, we can express P_{D}' as

$$\begin{aligned} P_{\mathrm{D}}' &= \int_{\mathcal{X}_1(\lambda)} L(\mathbf{x}) p(\mathbf{x}|\mathcal{H}_0) d\mathbf{x} + \int_{\mathcal{X}^+} L(\mathbf{x}) p(\mathbf{x}|\mathcal{H}_0) d\mathbf{x} - \int_{\mathcal{X}^-} L(\mathbf{x}) p(\mathbf{x}|\mathcal{H}_0) d\mathbf{x} \\ &= P_{\mathrm{D}} + \int_{\mathcal{X}^+} L(\mathbf{x}) p(\mathbf{x}|\mathcal{H}_0) d\mathbf{x} - \int_{\mathcal{X}^-} L(\mathbf{x}) p(\mathbf{x}|\mathcal{H}_0) d\mathbf{x}. \end{aligned} \tag{15.9}$$

Note that $L(\mathbf{x}) < \lambda$ for $\mathbf{x} \in \mathcal{X}^+$ and $L(\mathbf{x}) > \lambda$ for $\mathbf{x} \in \mathcal{X}^-$. Consequently, we have that

$$\int_{\mathcal{X}^+} L(\mathbf{x}) p(\mathbf{x}|\mathcal{H}_0) d\mathbf{x} - \int_{\mathcal{X}^-} L(\mathbf{x}) p(\mathbf{x}|\mathcal{H}_0) d\mathbf{x} < 0. \tag{15.10}$$

Using the above result in (15.9) yields $P_{\mathrm{D}}' < P_{\mathrm{D}}$ while $P_{\mathrm{F}}' = P_{\mathrm{F}}$. Thus, by choosing a decision boundary different than D for a fixed P_{F}, we decrease P_{D}. Conversely, we have the likelihood test in (15.4) maximizes P_{D} for a fixed P_{F}.

The likelihood ratio is a one dimensional quantity, regardless of the dimensionality of $\mathbf{x}$. This has some practical implication as shown in Example 15.1. Furthermore, the likelihood ratio is a function of the observation $\mathbf{x}$, which is a random vector. As such, it is also a random variable and will possess different distributions under $\mathcal{H}_0$ and $\mathcal{H}_1$ as $\mathbf{x}$ will have a different distribution under $\mathcal{H}_0$ and $\mathcal{H}_1$. Let ℓ be the value of the likelihood function in (15.4) and $p_L(\ell|\mathcal{H}_0)$ and $p_L(\ell|\mathcal{H}_1)$ denote the PDF of $L(\mathbf{x})$ under $\mathcal{H}_0$ and $\mathcal{H}_1$, respectively. Then, P_{F} can be expressed in terms of ℓ as

$$P_{\mathrm{F}}(\lambda) = \Pr[L(\mathbf{x}) > \lambda|\mathcal{H}_0] = \int_{\lambda}^{\infty} p_L(\ell|\mathcal{H}_0) d\ell. \tag{15.11}$$

Equivalently, P_{D} can be expressed as

$$P_{\mathrm{D}}(\lambda) = \Pr[L(\mathbf{x}) > \lambda|\mathcal{H}_1] = \int_{\lambda}^{\infty} p_L(\ell|\mathcal{H}_1) d\ell. \tag{15.12}$$

By decreasing λ we assign more parts of the observation space to $\mathcal{X}_1(\lambda)$, which increases both P_{D} and P_{F}. Therefore, we decrease λ until our constraint $P_{\mathrm{F}} \leq \alpha$ is met. Here we focus on the case that λ is a continuous variable for which $P_{\mathrm{F}} \leq \alpha$ holds with equality. Note that there exists cases for which this assumption does not hold.

Example 15.1

Assume that we want to discriminate between two Gaussian distributions. Both distributions have the same variance but different means. Under $\mathcal{H}_0$, the mean is zero while under $\mathcal{H}_1$ the mean is A. Furthermore, assume that multiple observations are available and that these observations are independent. Hence, the observations obey the following distributions under the two hypothesis:

$$\begin{aligned} p(\mathbf{x}|\mathcal{H}_0) &= \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1} x_n^2\right); \\ p(\mathbf{x}|\mathcal{H}_1) &= \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1} (x_n - A)^2\right). \end{aligned} \tag{15.13}$$

Such an example can be encountered when we want to detect the presence of a DC voltage in noise from some observations.

First, let us focus on the likelihood ratio, which is

$$\begin{aligned} L(\mathbf{x}) &= \frac{\frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x_n - A)^2\right)}{\frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left(-\frac{1}{\sigma^2}\sum_{n=0}^{N-1} x_n^2\right)} \\ &= \exp\left(-\frac{1}{2\sigma^2}\left(\sum_{n=0}^{N-1}(x_n - A)^2 - \sum_{n=0}^{N-1} x_n^2\right)\right) \\ &= \exp\left(-\frac{1}{2\sigma^2}\left(NA^2 - 2NA\bar{x}\right)\right), \end{aligned} \tag{15.14}$$

where $\bar{x} = (1/N)\sum_{n=0}^{N-1} x_n$ is the sample mean. The likelihood ratio is compared against a threshold

$$\exp\left(-\frac{1}{2\sigma^2}\left(NA^2 - 2AN\bar{x}\right)\right) \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \lambda. \tag{15.15}$$

We can apply the logarithm without changing the inequality, which simplifies the likelihood ratio test to

$$-\frac{1}{2\sigma^2}\left(NA^2 - 2NA\bar{x}\right) \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \ln\lambda. \tag{15.16}$$

We can bring all known variables on the right hand side, leaving us with

$$\bar{x} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \frac{\sigma^2}{NA}\ln\lambda + \frac{A}{2} = \lambda'. \tag{15.17}$$

We see that we can make a decision only by comparing the sample mean against the threshold λ' in (15.17).

To determine the new threshold λ', we look at the distribution of $\bar{x}$ under $\mathcal{H}_0$. Under $\mathcal{H}_0$, the sample mean is a sum of IID zero-mean Gaussian random variables with variance σ^2. Thus, the sample mean is a Gaussian random variable with zero mean and variance σ^2/N. P_F is now the probability that $\bar{x}$ exceeds λ' which is the complementary cumulative distribution function (tail distribution), as shown on the left in Figure 15.4. λ' is chosen such that the tail probability is equal to the desired P_F. Typically, the inverse Q-function

is used for this. On the other hand, P_{D} is the probability that $\bar{x}$ exceeds λ' under $\mathcal{H}_1$, which is shown on the right of Figure 15.4. Under $\mathcal{H}_1$, $\bar{x}$ is also Gaussian distributed with mean A and variance σ^2/N.

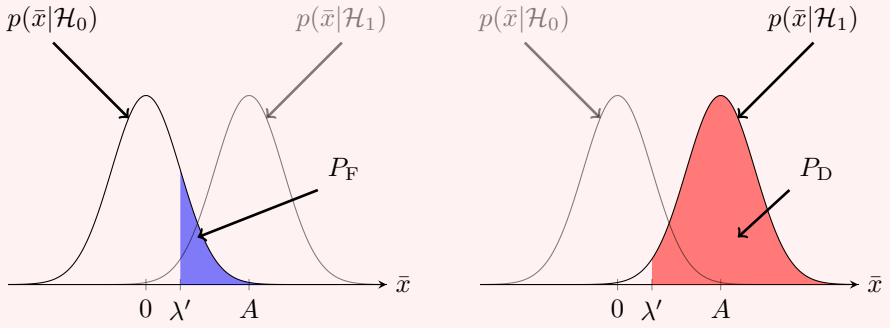


Figure 15.4: Distribution of $\bar{x}$ under $\mathcal{H}_0$ and $\mathcal{H}_1$

For the Neyman-Pearson test, we look at the upper boundary of the feasible region in Figure 15.2. This upper boundary is referred to as *receiver operating characteristic* (ROC) and is a concave curve for continuous likelihood tests. An example of the ROC is illustrated in Figure 15.5. The likelihood ratio is a non-negative quantity, implying $0 \leq \lambda < \infty$. For $\lambda = 0$, we assign the entire observation space to $\mathcal{X}_1(\lambda)$ and we have that $P_{\mathrm{D}} = 1$ and $P_{\mathrm{F}} = 1$. When we increase λ, we increase the decision region $\mathcal{X}_0$ and thereby simultaneously decrease P_{D} and P_{F}. Hence, P_{D} and P_{F} go to zero as λ increases. Both extrema are indicated in Figure 15.5. Furthermore, the slope of the tangent at a particular point $(P_{\mathrm{F}}(\lambda), P_{\mathrm{D}}(\lambda))$ of the ROC is the value of λ that is required to achieve these probabilities. To show this, we express P_{D} as

$$P_{\mathrm{D}}(\lambda) = \int_{\mathcal{X}_1(\lambda)} p(\mathbf{x}|\mathcal{H}_1) d\mathbf{x} = \int_{\mathcal{X}_1(\lambda)} L(\mathbf{x}) p(\mathbf{x}|\mathcal{H}_0) d\mathbf{x}. \tag{15.18}$$

Using the definition of the decision region $\mathcal{X}_1(\lambda)$ in (15.5), we can express P_{D} equivalently as

$$P_{\mathrm{D}}(\lambda) = \int_{\lambda}^{\infty} \ell p_L(\ell|\mathcal{H}_0) d\ell. \tag{15.19}$$

Applying the fundamental theorem of calculus and the chain rule yields

$$\frac{dP_{\mathrm{D}}(\lambda)}{d\lambda} = -\lambda p(\lambda|\mathcal{H}_0). \tag{15.20}$$

Similarly, differentiating (15.11) with respect to λ, we obtain

$$\frac{dP_{\mathrm{F}}(\lambda)}{d\lambda} = -p(\lambda|\mathcal{H}_0). \tag{15.21}$$

The resulting quotient is

$$\frac{\frac{dP_{\mathrm{D}}(\lambda)}{d\lambda}}{\frac{dP_{\mathrm{F}}(\lambda)}{d\lambda}} = \frac{dP_{\mathrm{D}}(\lambda)}{dP_{\mathrm{F}}(\lambda)} = \lambda, \tag{15.22}$$

which completes the proof.

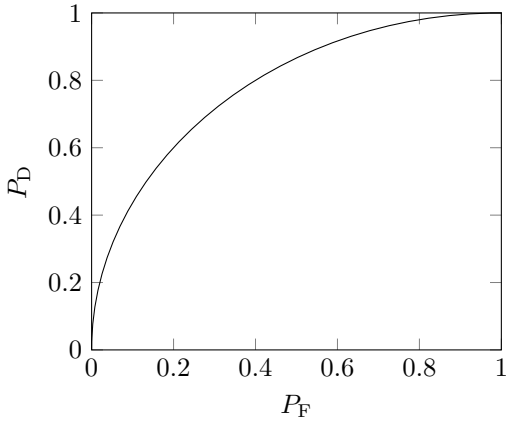


Figure 15.5: Receiver Operating Characteristic

15.3 Bayesian Test

In Chapter 10, we already applied the Bayesian framework to estimation problems. If prior probabilities of the hypotheses, denoted by P_0 and P_1, are available, we can also apply the Bayesian framework to hypothesis testing. To get started with Bayesian hypothesis testing, we introduce a cost function $C_{i,j}$. The cost function assigns a cost for each decision $\mathcal{H}_i$, given that hypothesis $\mathcal{H}_j$ is true. The expected cost, also called Bayesian risk, is

$$\mathcal{R} = \int_{\mathcal{X}_0} C_{0,0}P_0 p(\mathbf{x}|\mathcal{H}_0)d\mathbf{x} + \int_{\mathcal{X}_1} C_{1,0}P_0 p(\mathbf{x}|\mathcal{H}_0)d\mathbf{x} + \int_{\mathcal{X}_0} C_{0,1}P_1 p(\mathbf{x}|\mathcal{H}_1)d\mathbf{x} + \int_{\mathcal{X}_1} C_{1,1}P_1 p(\mathbf{x}|\mathcal{H}_1)d\mathbf{x}. \tag{15.23}$$

In Bayesian hypothesis testing, we make a decision ($\mathcal{H}_0$ or $\mathcal{H}_1$) such that the Bayesian risk for a given cost function is minimized. $\mathcal{X}_0$ and $\mathcal{X}_1$ are disjoint parts of the observation space. Therefore, we have that

$$\int_{\mathcal{X}_0} p(\mathbf{x}|\mathcal{H}_i)d\mathbf{x} = 1 - \int_{\mathcal{X}_1} p(\mathbf{x}|\mathcal{H}_i)d\mathbf{x}, \tag{15.24}$$

for $i = 0, 1$. Substituting (15.24) in (15.23) and rearranging terms allows us to express the Bayesian risk solely as a function of $\mathcal{X}_1$, which is

$$\mathcal{R} = C_{0,0}P_0 + C_{0,1}P_1 + \int_{\mathcal{X}_1} P_0(C_{1,0} - C_{0,0})p(\mathbf{x}|\mathcal{H}_0) + P_1(C_{1,1} - C_{0,1})p(\mathbf{x}|\mathcal{H}_1)P_1 d\mathbf{x}. \tag{15.25}$$

The first two terms are a fixed cost and are independent of the choice of the decision region $\mathcal{X}_1$. The Bayesian risk is minimized by choosing $\mathcal{X}_1$ such that the integrand to be either zero or negative, i.e.,

$$\mathcal{X}_1 = \{\mathbf{x} : P_0(C_{1,0} - C_{0,0})p(\mathbf{x}|\mathcal{H}_0) + P_1(C_{1,1} - C_{0,1})p(\mathbf{x}|\mathcal{H}_1) \leq 0\} \tag{15.26}$$

In general, we want to penalize a wrong decision for a given hypothesis with a higher cost than for correct decision, i.e., $C_{0,1} > C_{1,1}$ and $C_{1,0} > C_{0,0}$. Therefore, we include $\mathbf{x}$ to $\mathcal{X}_1$ if $P_1(C_{1,1} - C_{0,1})p(\mathbf{x}|\mathcal{H}_1) \geq P_0(C_{1,0} - C_{0,0})p(\mathbf{x}|\mathcal{H}_0)$, or equivalently

$$\frac{p(\mathbf{x}|\mathcal{H}_1)}{p(\mathbf{x}|\mathcal{H}_0)} \geq \frac{(C_{1,0} - C_{0,0})P_0}{(C_{0,1} - C_{1,1})P_1} = \lambda. \tag{15.27}$$

The quantity on the left hand side is the likelihood ratio, which we already encountered in Chapter 15.2.

The decision threshold simplifies if we consider the costs $C_{1,1} = C_{0,0} = 0$ and $C_{0,1} = C_{0,1} = 1$, or in other words, we do not penalize correct decisions at all. For this particular choice of cost function, the Bayesian risks in (10.5) becomes

$$\mathcal{R} = P_0 \int_{\mathcal{X}_1} p(\mathbf{x}|\mathcal{H}_0) d\mathbf{x} + P_1 \int_{\mathcal{X}_0} p(\mathbf{x}|\mathcal{H}_1) d\mathbf{x}. \tag{15.28}$$

Thus, this cost assignment minimizes the total probability of error.

15.4 Matched Filter

In application such as radars or communication system, the task at hand is to detect the presence or absence of a known signal s_n of length N from some noisy observations. Therefore, we define the two hypothesis as

$$\begin{aligned} \mathcal{H}_0 : &\quad x_n = w_n; \\ \mathcal{H}_1 : &\quad x_n = s_n + w_n. \end{aligned} \tag{15.29}$$

where w_n is IID zero-mean Gaussian noise with variance σ^2. The corresponding PDFs are

$$\begin{aligned} p(\mathbf{x}|\mathcal{H}_0) &= \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} x_n^2 \right); \\ p(\mathbf{x}|\mathcal{H}_1) &= \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} (x_n - s_n)^2 \right). \end{aligned} \tag{15.30}$$

Using the likelihood ratio test as described in Chapter 15.2 yields

$$L(\mathbf{x}) = \exp\left(-\frac{1}{2\sigma^2} \left(\sum_{n=0}^{N-1} (x_n - s_n)^2 - \sum_{n=0}^{N-1} x_n^2 \right) \right) \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \lambda. \tag{15.31}$$

As in Example 15.1, we can take the natural logarithm, which results in

$$\begin{aligned} -\frac{1}{2\sigma^2} \left(\sum_{n=0}^{N-1} (x_n - s_n)^2 - \sum_{n=0}^{N-1} x_n^2 \right) &\underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \ln \lambda \\ \frac{1}{\sigma^2} \sum_{n=0}^{N-1} x_n s_n - \frac{1}{2\sigma^2} \sum_{n=0}^{N-1} s_n^2 &\underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \ln \lambda. \end{aligned} \tag{15.32}$$

Since s_n is known and does not dependent on the data x_n, we can incorporate it in the threshold, which yields

$$\sum_{n=0}^{N-1} x_n s_n \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \sigma^2 \ln \lambda + \frac{1}{2} \sum_{n=0}^{N-1} s_n^2 = \lambda'. \tag{15.33}$$

We recognize the left hand side as the inner product between the observed data x_n and the known signal x_n. A detector which calculates the inner product is also called *correlator*.

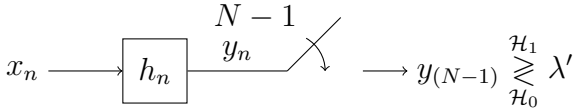


Figure 15.6: The output of the matched filter is sampled after all samples have been observed. The sampled output is then compared against a threshold.

The correlator can efficiently implemented using finite impulse response filter (FIR). Therefore, consider the output y_n of a filter that is given by the convolution of the input sequence x_n and the finite impulse response h_n of the filter

$$y_n = \sum_{l=0}^{N-1} x_l h_{n-l}. \tag{15.34}$$

Comparing (8.2) for $n = N-1$ and (15.33), we can see that (15.33) can be obtained by choosing $h_n = s_{N-1-n}$:

$$\begin{aligned} y_{N-1} &= \sum_{l=0}^{N-1} x_l s_{(N-1+n-N+1)} \\ &= \sum_{l=0}^{N-1} x_l s_l \end{aligned} \tag{15.35}$$

Another important property of the matched filter is that from all possible filters h it has the largest signal-to-noise ratio (SNR). To prove this statement, we first define the SNR as

$$\text{SNR} = \frac{\text{E}^2\left[y_{(N-1)}; \mathcal{H}_1\right]}{\text{Var}\left[y_{(N-1)}; \mathcal{H}_0\right]} \tag{15.36}$$

The sample $y_{(N-1)}$ under $\mathcal{H}_1$ is

$$y_{(N-1)} = \sum_{l=0}^{N-1} (s_l + w_l) h_{N-1-l}, \tag{15.37}$$

which can equivalently be expressed, using vector notation, as

$$y_{(N-1)} = \mathbf{h}^T(\mathbf{s} + \mathbf{w}), \tag{15.38}$$

where $\mathbf{s} = [s_0, s_1, \ldots, s_{(N-1)}]^T$, $\mathbf{w} = [w_0, w_1, \ldots, w_{N-1}]^T$, and $\mathbf{h} = [h_{(N-1)}, h_{(N-2)}, \ldots, h_0]^T$. Similarly, $y_{(N-1)}$ under $\mathcal{H}_0$ is

$$y_{(N-1)} = \mathbf{h}^T\mathbf{w}. \tag{15.39}$$

Using the above introduce vector notation, we can express the SNR as

$$\begin{aligned}
\text{SNR} &= \frac{\mathrm{E}^2\left[\mathbf{h}^T(\mathbf{s}+\mathbf{w})\right]}{\mathrm{Var}\left[\mathbf{h}^T\mathbf{w}\right]} \\
&= \frac{\mathrm{E}^2\left[\mathbf{h}^T\mathbf{s}+\mathbf{h}^T\mathbf{w}\right]}{\mathrm{E}\left[(\mathbf{h}^T\mathbf{w})^2\right]-\mathrm{E}^2\left(\mathbf{h}^T\mathbf{w}\right]} \\
&= \frac{(\mathbf{h}^T\mathbf{s})^2}{\mathbf{h}^T\,\mathrm{E}\left[\mathbf{w}\mathbf{w}^T\right]\mathbf{h}} \\
&= \frac{(\mathbf{h}^T\mathbf{s})^2}{\mathbf{h}^T\sigma^2\mathbf{I}\mathbf{h}} \\
&= \frac{(\mathbf{h}^T\mathbf{s})^2}{\sigma^2\mathbf{h}^T\mathbf{h}}.
\end{aligned} \tag{15.40}$$

Using the Cauchy-Schwarz inequality, we can upper-bound the numerator as follows:

$$(\mathbf{h}^T\mathbf{s})^2 \leq (\mathbf{h}^T\mathbf{h})(\mathbf{s}^T\mathbf{s}). \tag{15.41}$$

Equality is achieved for the case that $\mathbf{h} = c\mathbf{s}$ where c is an arbitrary constant.

Part V

Appendices

Appendix A

Families of discrete random variables

A.1 Bernoulli distribution

The Bernoulli distribution is a discrete probability distribution that models an experiment where only 2 outcomes are possible. The probability distribution of flipping a coin is an example of a Bernoulli distribution. These outcomes are mapped to 0 and 1, whose probabilities are $1-p$ and p respectively. The distribution is fully characterized by the parameter p, which is the probability of success ($\Pr[X = 1]$).

A.1.1 PMF

The PMF of the discrete Bernoulli(p) distribution is given as

$$p_X(x) = \begin{cases} 1-p, & \text{for } x = 0 \\ p, & \text{for } x = 1 \\ 0, & \text{otherwise} \end{cases} \tag{A.1}$$

where p is in the range $0 < p < 1$.

A.1.2 Cumulative distribution function

The CDF of the discrete Bernoulli(p) distribution can be determined as

$$P_X(x) = \begin{cases} 0, & \text{for } x < 0, \\ 1-p, & \text{for } x = 0, \\ 1, & \text{for } x > 0. \end{cases} \tag{A.2}$$

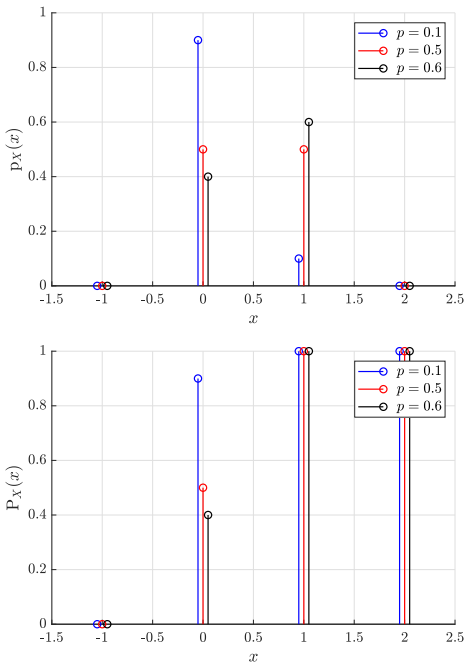


Figure A.1: Example plot of the (a) PMF and (b) cumulative density function of the Bernoulli(p) distribution.

Proof.

$$
\begin{aligned}
P_X(x) &= \sum_{n=-\infty}^{x} p_X(n) \\
&= \begin{cases} 0, & \text{for } x < 0 \\ \sum_{n=-\infty}^{0} p_X(n), & \text{for } x = 0 \\ 1, & \text{for } x > 0 \end{cases} \\
&= \begin{cases} 0, & \text{for } x < 0, \\ 1-p, & \text{for } x = 0, \\ 1, & \text{for } x > 0. \end{cases}
\end{aligned} \tag{A.3}
$$

□

A.1.3 Expected value

The expected value of the discrete Bernoulli(p) distribution can be determined as

$$\mathrm{E}[X] = p. \tag{A.4}$$

Proof.

$$\begin{aligned} \mathrm{E}[X] &= \sum_{n=-\infty}^{\infty} n \cdot p_X(n), \\ &= \sum_{n=0}^{1} n \cdot p_X(n), \\ &= 0 \cdot (1-p) + 1 \cdot p, \\ &= p. \end{aligned} \tag{A.5}$$

□

A.1.4 Variance

The variance of the discrete Bernoulli(p) distribution can be determined as

$$\mathrm{Var}[X] = p(1-p) \tag{A.6}$$

Proof.

$$\begin{aligned} \mathrm{Var}[X] &= \mathrm{E}[(X - \mathrm{E}[X])^2], \\ &= \mathrm{E}[X^2] - 2\,\mathrm{E}[X]\,\mathrm{E}[X] + \mathrm{E}[X]^2, \\ &= \mathrm{E}[X^2] - \mathrm{E}[X]^2, \\ &= \sum_{n=-\infty}^{\infty} n^2 \cdot p_X(n) - p^2, \\ &= \sum_{n=0}^{1} n^2 \cdot p_X(n) - p^2, \\ &= 0^2 \cdot (1-p) + 1^2 \cdot p - p^2. \\ &= p(1-p) \end{aligned} \tag{A.7}$$

□

A.2 Geometric distribution

The Geometric distribution is a discrete probability distribution that models an experiment with probability of success p. The Geometric distribution gives the probability that the first success is observed at the x^{th} independent trial. The distribution is fully characterized by the parameter p, which is the probability of success.

A.2.1 Probability mass function

The PMF of the discrete Geometric(p) distribution is given as

$$p_X(x) = \begin{cases} p(1-p)^{x-1}, & \text{for } x = 1, 2, \dots \\ 0, & \text{otherwise} \end{cases} \tag{A.8}$$

where p is in the range $0 < p < 1$.

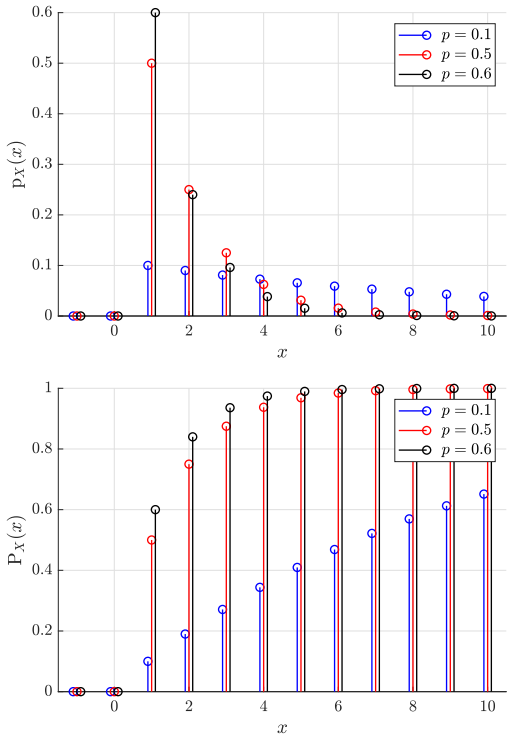


Figure A.2: Example plot of the (a) PMF and (b) cumulative density function of the Geometric(p) distribution.

A.2.2 Cumulative distribution function

The CDF of the discrete Geometric(p) distribution can be determined as

$$P_X(x) = \begin{cases} 0, & \text{for } x < 1 \\ 1 - (1-p)^x. & \text{for } x \geq 1 \end{cases} \tag{A.9}$$

Proof.

$$\begin{aligned} P_X(x) &= \sum_{n=-\infty}^{x} p_X(n), \\ &= \begin{cases} 0, & \text{for } x < 1 \\ \sum_{n=1}^{x} p(1-p)^{n-1}, & \text{for } x \geq 1 \end{cases} \\ &= \begin{cases} 0, & \text{for } x < 1 \\ p\sum_{m=0}^{x-1}(1-p)^{m}, & \text{for } x \geq 1 \end{cases} \\ &= \begin{cases} 0, & \text{for } x < 1 \\ p\frac{1-(1-p)^x}{1-(1-p)}, & \text{for } x \geq 1 \end{cases} \\ &= \begin{cases} 0, & \text{for } x < 1 \\ 1 - (1-p)^x. & \text{for } x \geq 1 \end{cases} \end{aligned} \tag{A.10}$$

□

A.2.3 Expected Value

The expected value of the discrete Geometric(p) distribution can be determined as

$$\mathrm{E}[X] = \frac{1}{p}. \tag{A.11}$$

Proof.

$$\begin{aligned}
\mathrm{E}[X] &= \sum_{n=-\infty}^{\infty} n \cdot p_X(n), \\
&= \sum_{n=1}^{\infty} n \cdot p(1-p)^{n-1}, \\
&= p \sum_{n=1}^{\infty} n(1-p)^{n-1}, \\
&= \frac{p}{(1-(1-p))^2} = \frac{1}{p}.
\end{aligned} \tag{A.12}$$

□

A.2.4 Variance

The variance of the discrete Geometric(p) distribution can be determined as

$$\mathrm{Var}[X] = \frac{1-p}{p^2}. \tag{A.13}$$

Proof.

$$\begin{aligned}
\mathrm{Var}[X] &= \mathrm{E}[(X - \mathrm{E}[X])^2], \\
&= \mathrm{E}[X^2] - 2\,\mathrm{E}[X]\,\mathrm{E}[X] + \mathrm{E}[X]^2, \\
&= \mathrm{E}[X^2] - \mathrm{E}[X]^2, \\
&= \sum_{n=-\infty}^{\infty} n^2 \cdot p(1-p)^{n-1} - \frac{1}{p^2}, \\
&= p \sum_{n=1}^{\infty} n^2 \cdot (1-p)^{n-1} - \frac{1}{p^2}, \\
&= -\frac{p((1-p)+1)}{((1-p)-1)^3} - \frac{1}{p^2}, \\
&= \frac{2p-p^2}{p^3} - \frac{p}{p^3} = \frac{p-p^2}{p^3} = \frac{1-p}{p^2}.
\end{aligned} \tag{A.14}$$

□

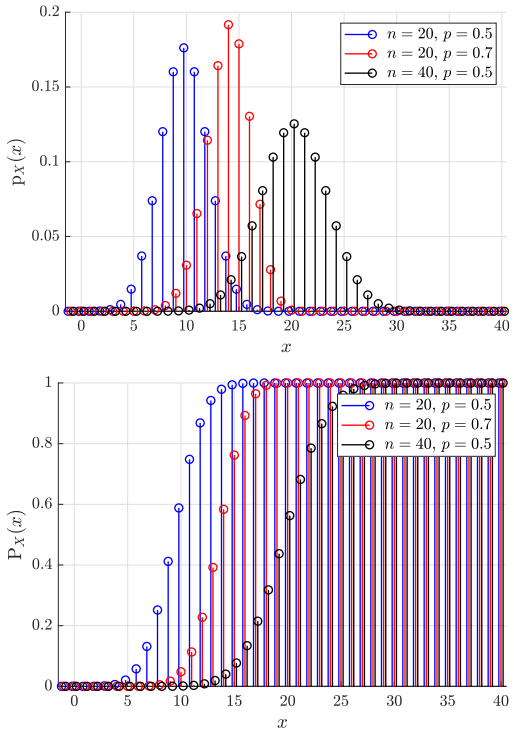


Figure A.3: Example plot of the (a) PMF and (b) cumulative density function of the Binomial(p) distribution.

A.3 Binomial distribution

The Binomial distribution is a discrete probability distribution that models an experiment with probability of success p. The Binomial distribution gives the probability of observing x successes in n independent trials. The distribution is fully characterized by the parameters n and p. The parameter n denotes the number of independent trials and the parameter p denotes the probability of observing a success per trial.

A.3.1 Probability mass function

The PMF of the discrete Binomial(n,p) distribution is given as

$$p_X(x) = \binom{n}{x} p^x (1-p)^{n-x}, \tag{A.15}$$

where $0 < p < 1$ and n is an integer such that $n \geq 1$.

A.3.2 Cumulative distribution function

The CDF of the discrete Binomial(n,p) distribution can be determined as

$$P_X(x) = \sum_{m=0}^{x} \binom{n}{m} p^m (1-p)^{n-m}. \tag{A.16}$$

Proof.

$$\begin{aligned} P_X(x) &= \sum_{m=0}^{x} p_X(m), \\ &= \sum_{m=0}^{x} \binom{n}{m} p^m (1-p)^{n-m}. \end{aligned} \tag{A.17}$$

□

A.3.3 Expected value

The expected value of the discrete Binomial(n,p) distribution can be determined as

$$\mathrm{E}[X] = np \tag{A.18}$$

Proof.

$$\begin{aligned} \mathrm{E}[X] &= \sum_{m=-\infty}^{\infty} m \cdot p_X(m), \\ &= \sum_{m=0}^{n} m \cdot \binom{n}{m} p^m (1-p)^{n-m}, \\ &= \sum_{m=1}^{n} m \cdot \binom{n}{m} p^m (1-p)^{n-m}, \\ &= \sum_{m=1}^{n} n \cdot \binom{n-1}{m-1} p^m (1-p)^{n-m}, \\ &= np \sum_{m=1}^{n} \binom{n-1}{m-1} p^{m-1} (1-p)^{(n-1)-(m-1)}, \\ &= np \sum_{k=0}^{l} \binom{l}{k} p^k (1-p)^{l-k}, \\ &= np \cdot (p + (1-p))^l = np. \end{aligned} \tag{A.19}$$

□

A.3.4 Variance

The variance of the discrete Binomial(n,p) distribution can be determined as

$$\mathrm{Var}[X] = np(1-p). \tag{A.20}$$

Proof.

$$
\begin{aligned}
\mathrm{Var}[X] &= \mathrm{E}[(X - \mathrm{E}[X])^2], \\
&= \mathrm{E}[X^2] - \mathrm{E}[X]^2, \\
&= -(np)^2 + \sum_{m=-\infty}^{\infty} m^2 \cdot p_X(m), \\
&= -(np)^2 + \sum_{m=0}^{n} m^2 \cdot \binom{n}{m} p^m (1-p)^{n-m}, \\
&= -(np)^2 + \sum_{m=1}^{n} nm \cdot \binom{n-1}{m-1} p^m (1-p)^{n-m}, \\
&= -(np)^2 + np \sum_{m=1}^{n} m \binom{n-1}{m-1} p^{m-1} (1-p)^{(n-1)-(m-1)}, \\
&= -(np)^2 + np \sum_{k=0}^{l} (k+1) \binom{l}{k} p^k (1-p)^{l-k}, \\
&= -(np)^2 + np \sum_{k=1}^{l} k \binom{l}{k} p^k (1-p)^{l-k} + np \sum_{k=0}^{l} \binom{l}{k} p^k (1-p)^{l-k}, \\
&= -(np)^2 + np \sum_{k=1}^{l} l \binom{l-1}{k-1} p^k (1-p)^{l-k} + np \sum_{k=0}^{l} \binom{l}{k} p^k (1-p)^{l-k}, \\
&= -(np)^2 + nlp^2 \sum_{k=1}^{l} \binom{l-1}{k-1} p^{k-1} (1-p)^{(l-1)-(k-1)} + np \sum_{k=0}^{l} \binom{l}{k} p^k (1-p)^{l-k}, \\
&= -(np)^2 + nlp^2 \sum_{j=0}^{i} \binom{i}{j} p^j (1-p)^{i-j} + np \sum_{k=0}^{l} \binom{l}{k} p^k (1-p)^{l-k}, \\
&= -(np)^2 + nlp^2 (p + (1-p))^i + np(p + (1-p))^l, \\
&= -n^2p^2 + n(n-1)p^2 + np = -np^2 + np = np(1-p).
\end{aligned}
\tag{A.21}
$$

□

A.4 Discrete uniform distribution

The discrete uniform distribution is a discrete probability distribution that models an experiment where the outcomes are mapped only to discrete points on the interval from k up to and including l. The distribution is fully characterized by the parameters k and l, which are the discrete lower and upper bound of the interval respectively.

A.4.1 Probability mass function

The PMF of the discrete Uniform(k,l) distribution is given as

$$
p_X(x) = \begin{cases} \frac{1}{l-k+1}, & \text{for } x = k, k+1, k+2, \ldots, l \\ 0, & \text{otherwise} \end{cases}
\tag{A.22}
$$

where k and l are integers such that $k < l$.

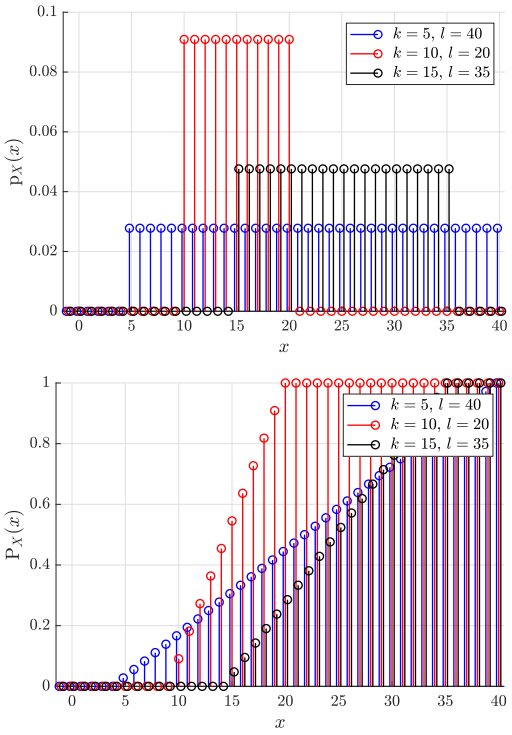


Figure A.4: Example plot of the (a) PMF and (b) cumulative density function of the discrete uniform (k,l) distribution.

A.4.2 Cumulative distribution function

The CDF of the discrete Uniform(k,l) distribution can be determined as

$$P_X(x) = \begin{cases} 0 & \text{for } x < k \\ \frac{x-k+1}{l-k+1} & \text{for } k \leq x < l \\ 1 & \text{for } x \geq l \end{cases} \tag{A.23}$$

Proof.

$$\begin{aligned} P_X(x) &= \sum_{m=-\infty}^{x} p_X(m), \\ &= \begin{cases} 0, & \text{for } x < k \\ \sum_{m=k}^{x} \frac{1}{l-k+1}, & \text{for } k \leq x < l \\ 1, & \text{for } x \geq l \end{cases} \\ &= \begin{cases} 0, & \text{for } x < k \\ \frac{x-k+1}{l-k+1}, & \text{for } k \leq x < l \\ 1. & \text{for } x \geq l \end{cases} \end{aligned} \tag{A.24}$$

□

A.4.3 Expected value

The expected value of the discrete Uniform(k,l) distribution can be determined as

$$\mathrm{E}[X] = \frac{k+l}{2} \tag{A.25}$$

Proof.

$$\begin{aligned}
\mathrm{E}[X] &= \sum_{n=-\infty}^{\infty} n \cdot p_X(n), \\
&= \sum_{n=k}^{l} n \cdot \frac{1}{l-k+1}, \\
&= \frac{1}{l-k+1} \sum_{n=k}^{l} n, \\
&= \frac{1}{l-k+1} \sum_{n=k}^{l} n, \\
&= \frac{1}{l-k+1} \cdot \frac{1}{2}(k+l)(l-k+1), \\
&= \frac{k+l}{2}.
\end{aligned} \tag{A.26}$$

□

A.4.4 Variance

The variance of the discrete Uniform(k,l) distribution can be determined as

$$\mathrm{Var}[X] = \frac{(l-k+1)^2 - 1}{12} \tag{A.27}$$

Proof. The variance of the discrete Uniform(k,l) distribution can be determined as

$$
\begin{aligned}
\mathrm{Var}[X] &= \mathrm{E}[X^2] - \mathrm{E}[X]^2, \\
&= -\frac{(k+l)^2}{4} + \sum_{n=k}^{l} n^2 \cdot \frac{1}{l-k+1}, \\
&= -\frac{(k+l)^2}{4} + \frac{1}{l-k+1} \sum_{n=k}^{l} n^2, \\
&= -\frac{(k+l)^2}{4} + \frac{1}{l-k+1} \sum_{m=1}^{l-k+1} (m+k-1)^2, \\
&= -\frac{(k+l)^2}{4} + \frac{1}{l-k+1} \sum_{m=1}^{l-k+1} (m^2 + 2mk + k^2 + 1 - 2m - 2k), \\
&= -\frac{(k+l)^2}{4} + \frac{1}{l-k+1} \left(\sum_{m=1}^{l-k+1} m^2 + (2k-2) \sum_{m=1}^{l-k+1} m + (k^2+1-2k) \sum_{m=1}^{l-k+1} 1 \right), \\
&= -\frac{(k+l)^2}{4} + \frac{1}{l-k+1} \Bigg(\frac{(l-k+1)(l-k+2)(2l-2k+3)}{6} \\
&\qquad + (2k-2)\frac{1}{2}(l-k+1)(l-k+2) + (k^2+1-2k)(l-k+1) \Bigg), \\
&= \frac{-3k^2 - 3l^2 - 6kl}{12} + \frac{4l^2 - 4kl + 6l - 4kl + 4k^2 - 6k + 8l - 8k + 12}{12} \\
&\qquad + \frac{12kl - 12k^2 + 24k - 12l + 12k - 24}{12} + \frac{12k^2 + 12 - 24k}{12}, \\
&= \frac{k^2 + l^2 - 2kl - 2k + 2l}{12}, \\
&= \frac{(l-k+1)^2 - 1}{12}
\end{aligned}
\tag{A.28}
$$

□

A.5 Poisson distribution

The Poisson distribution is a discrete probability distribution that models the number of events occurring within a certain interval of time, in which the events occur independently from each other at a constant rate. The exact moments at which the events occur are unknown, however, the average number of events occurring within the interval is known and is denoted by the parameter α. An example of a process, where the number of events within an interval can be described as a Poisson distribution, is the number of phone calls over a network. For optimal allocation of resources, a service provider needs to know the chance that the allocated capacity is insufficient in order to limit the number of dropped calls. The inhabitants can be described as independent entities (i.e. everyone makes a phone call whenever it suits him or her), whilst they usually have their own habit of making phone calls.

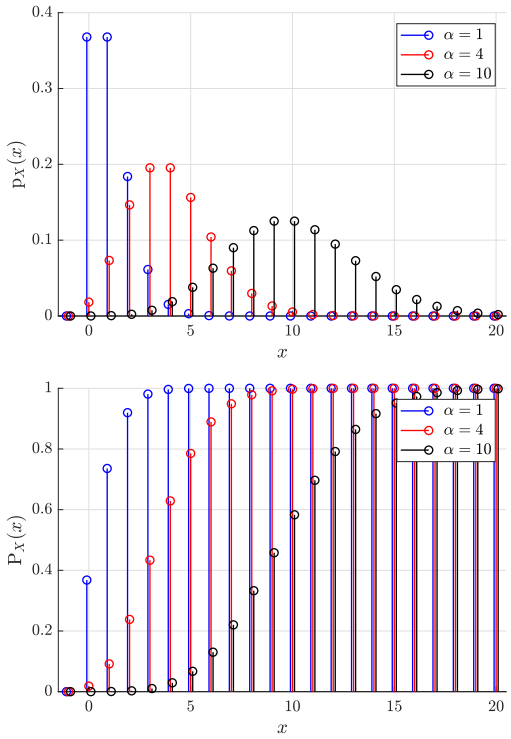


Figure A.5: Example plot of the (a) PMF and (b) cumulative density function of the Poisson(α) distribution.

A.5.1 Probability mass function

The probability mass function of the discrete Poisson(α) distribution is given as

$$p_X(x) = \begin{cases} \frac{\alpha^x e^{-\alpha}}{x!}, & \text{for } x = 0, 1, 2, \dots \\ 0, & \text{otherwise} \end{cases} \tag{A.29}$$

where α is in the range $\alpha > 0$.

A.5.2 Cumulative distribution function

The CDF of the discrete Poisson(α) distribution can be determined as

$$P_X(x) = \begin{cases} 0, & \text{for } x < 0 \\ e^{-\alpha} \sum_{n=0}^{x} \frac{\alpha^n}{n!}. & \text{for } x \geq 0 \end{cases} \tag{A.30}$$

Proof.

$$\begin{aligned} P_X(x) &= \sum_{n=-\infty}^{x} p_X(n), \\ &= \begin{cases} 0, & \text{for } x < 0 \\ \sum_{n=0}^{x} \frac{\alpha^n e^{-\alpha}}{n!}, & \text{for } x \geq 0 \end{cases} \\ &= \begin{cases} 0, & \text{for } x < 0 \\ e^{-\alpha} \sum_{n=0}^{x} \frac{\alpha^n}{n!}. & \text{for } x \geq 0 \end{cases} \end{aligned} \tag{A.31}$$

□

A.5.3 Expected value

The expected value of the discrete Poisson(α) distribution can be determined as

$$\mathrm{E}[X] = \alpha \tag{A.32}$$

Proof.

$$\begin{aligned} \mathrm{E}[X] &= \sum_{n=-\infty}^{\infty} n \cdot p_X(n), \\ &= \sum_{n=0}^{\infty} n \cdot \frac{\alpha^n e^{-\alpha}}{n!}, \\ &= \sum_{n=1}^{\infty} n \cdot \frac{\alpha^n e^{-\alpha}}{n!}, \\ &= \alpha \sum_{n=1}^{\infty} \frac{\alpha^{n-1} e^{-\alpha}}{(n-1)!}, \\ &= \alpha \sum_{l=0}^{\infty} \frac{\alpha^l e^{-\alpha}}{l!}, \\ &= \alpha \end{aligned} \tag{A.33}$$

□

A.5.4 Variance

The variance of the discrete Poisson(α) distribution can be determined as

$$\mathrm{Var}[X] = \alpha \tag{A.34}$$

Proof. The variance of the discrete Poisson(α) distribution can be determined as

$$\begin{aligned}
\mathrm{Var}[X] &= \mathrm{E}[X^2] - \mathrm{E}[X]^2, \\
&= -\alpha^2 + \sum_{n=-\infty}^{\infty} n^2 \cdot p_X(n), \\
&= -\alpha^2 + \sum_{n=0}^{\infty} n^2 \cdot \frac{\alpha^n e^{-\alpha}}{n!}, \\
&= -\alpha^2 + \sum_{n=1}^{\infty} n^2 \cdot \frac{\alpha^n e^{-\alpha}}{n!}, \\
&= -\alpha^2 + \alpha \sum_{n=1}^{\infty} n \cdot \frac{\alpha^{n-1} e^{-\alpha}}{(n-1)!}, \\
&= -\alpha^2 + \alpha \sum_{l=0}^{\infty} (l+1) \cdot \frac{\alpha^l e^{-\alpha}}{l!}, \\
&= -\alpha^2 + \alpha \sum_{l=0}^{\infty} \frac{\alpha^l e^{-\alpha}}{l!} + \alpha \sum_{l=0}^{\infty} l \cdot \frac{\alpha^l e^{-\alpha}}{l!}, \\
&= -\alpha^2 + \alpha \sum_{l=0}^{\infty} \frac{\alpha^l e^{-\alpha}}{l!} + \alpha \sum_{l=1}^{\infty} l \cdot \frac{\alpha^l e^{-\alpha}}{l!}, \\
&= -\alpha^2 + \alpha \sum_{l=0}^{\infty} \frac{\alpha^l e^{-\alpha}}{l!} + \alpha^2 \sum_{l=1}^{\infty} \frac{\alpha^{l-1} e^{-\alpha}}{(l-1)!}, \\
&= -\alpha^2 + \alpha \sum_{l=0}^{\infty} \frac{\alpha^l e^{-\alpha}}{l!} + \alpha^2 \sum_{i=0}^{\infty} \frac{\alpha^i e^{-\alpha}}{i!}, \\
&= -\alpha^2 + \alpha + \alpha^2 = \alpha
\end{aligned} \tag{A.35}$$

□

Appendix B

Families of continuous random variables

B.1 Exponential distribution

The exponential distribution is a continuous probability distribution that follows an exponential curve. The curve is fully characterized by the rate parameter λ.

B.1.1 Probability density function

The PDF of the continuous Exponential(λ) distribution is given as

$$p_X(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{for } x \geq 0 \\ 0, & \text{for } x < 0 \end{cases} \tag{B.1}$$

where $\lambda > 0$.

B.1.2 Cumulative distribution function

The CDF of the continuous Exponential(λ) distribution can be determined as

$$P_X(x) = \begin{cases} 1 - e^{-\lambda x} & \text{for } x \geq 0 \\ 0 & \text{for } x < 0 \end{cases} \tag{B.2}$$

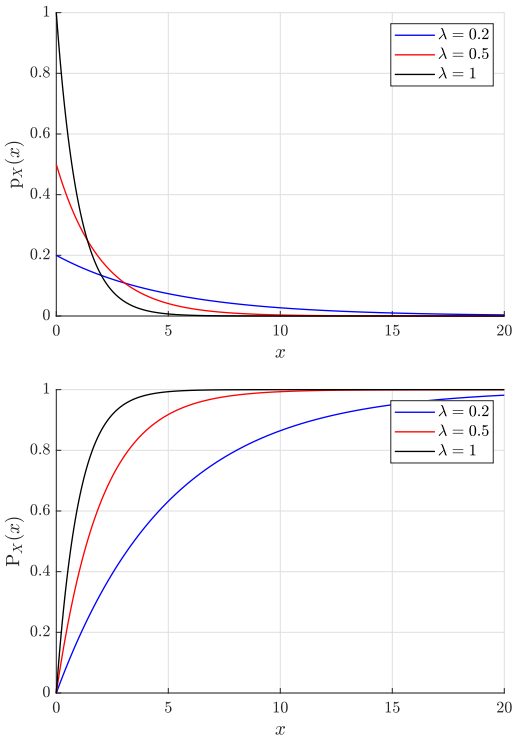


Figure B.1: Example plot of the (a) probability mass function and (b) cumulative density function of the Exponential(λ) distribution.

Proof. The CDF of the continuous Exponential(λ) distribution can be determined as

$$\begin{aligned} P_X(x) &= \int_{-\infty}^{x} p_X(n)dn, \\ &= \begin{cases} \int_0^x \lambda e^{-\lambda n} dn, & \text{for } x \geq 0 \\ 0, & \text{for } x < 0 \end{cases} \\ &= \begin{cases} \lambda \left[\frac{-1}{\lambda} e^{-\lambda n}\right]_0^x, & \text{for } x \geq 0 \\ 0, & \text{for } x < 0 \end{cases} \\ &= \begin{cases} \left[-e^{-\lambda n}\right]_0^x, & \text{for } x \geq 0 \\ 0, & \text{for } x < 0 \end{cases} \\ &= \begin{cases} 1 - e^{-\lambda x}, & \text{for } x \geq 0 \\ 0 & \text{for } x < 0 \end{cases} \end{aligned} \tag{B.3}$$

□

B.1.3 Expected value

The expected value of the continuous Exponential(λ) distribution can be determined as

$$\mathrm{E}[X] = \frac{1}{\lambda}. \tag{B.4}$$

Proof.

$$\begin{aligned}
\mathrm{E}[X] &= \int_{-\infty}^{\infty} x \cdot p_X(x) dx, \\
&= \lambda \int_0^{\infty} x e^{-\lambda x} dx, \\
&= \lambda \left[\frac{-1}{\lambda} x e^{-\lambda x} \right]_0^{\infty} - \lambda \int_0^{\infty} \frac{-1}{\lambda} e^{-\lambda x} dx, \\
&= - \left[x e^{-\lambda x} \right]_0^{\infty} - \frac{1}{\lambda} \left[e^{-\lambda x} \right]_0^{\infty}, \\
&= -(0-0) - \frac{1}{\lambda}(0-1) = \frac{1}{\lambda}.
\end{aligned} \tag{B.5}$$

□

B.1.4 Variance

The variance of the continuous Exponential(λ) distribution can be determined as

$$\mathrm{Var}[X] = \frac{1}{\lambda^2} \tag{B.6}$$

Proof. The variance of the continuous Exponential(λ) distribution can be determined as

$$\begin{aligned}
\mathrm{Var}[X] &= \mathrm{E}[X^2] - \mathrm{E}[X]^2, \\
&= \int_{-\infty}^{\infty} x^2 \cdot p_X(x) dx - \frac{1}{\lambda^2}, \\
&= -\frac{1}{\lambda^2} + \lambda \int_0^{\infty} x^2 e^{-\lambda x} dx, \\
&= -\frac{1}{\lambda^2} + \lambda \left[\frac{-1}{\lambda} x^2 e^{-\lambda x} \right]_0^{\infty} - \lambda \int_0^{\infty} \frac{-2}{\lambda} x e^{-\lambda x} dx, \\
&= -\frac{1}{\lambda^2} - \left[x^2 e^{-\lambda x} \right]_0^{\infty} + 2 \int_0^{\infty} x e^{-\lambda x} dx, \\
&= -\frac{1}{\lambda^2} - (0-0) + 2 \left[\frac{-1}{\lambda} x e^{-\lambda x} \right]_0^{\infty} - 2 \int_0^{\infty} \frac{-1}{\lambda} e^{-\lambda x} dx, \\
&= -\frac{1}{\lambda^2} - \frac{2}{\lambda} \left[x e^{-\lambda x} \right]_0^{\infty} + \frac{2}{\lambda} \int_0^{\infty} e^{-\lambda x} dx, \\
&= -\frac{1}{\lambda^2} - \frac{2}{\lambda}(0-0) + \frac{2}{\lambda} \left[\frac{-1}{\lambda} e^{-\lambda x} \right]_0^{\infty}, \\
&= -\frac{1}{\lambda^2} - \frac{2}{\lambda^2}(0-1) = \frac{1}{\lambda^2}
\end{aligned} \tag{B.7}$$

□

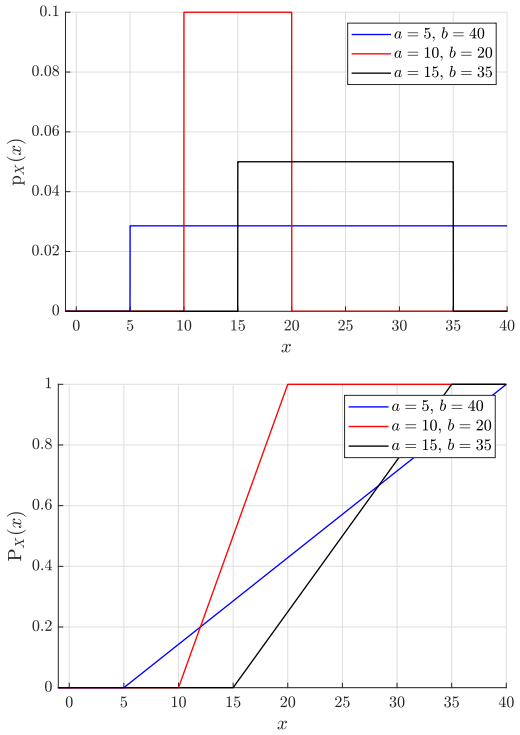


Figure B.2: Example plot of the (a) probability mass function and (b) cumulative density function of the continuous Uniform(a,b) distribution.

B.2 Continuous uniform distribution

The continuous Uniform distribution is a continuous probability distribution that models an experiment where the outcomes are mapped only to the interval from a up to and including b, with the same probability all over this range. The distribution is fully characterized by the parameters a and b, which are the continuous lower and upper bound of the interval respectively.

B.2.1 Probability density function

The PDF of the continuous Uniform(a,b) distribution is given as

$$p_X(x) = \begin{cases} \frac{1}{b-a}, & \text{for } a \leq x \leq b \\ 0, & \text{otherwise} \end{cases}, \tag{B.8}$$

where $b > a$.

B.2.2 Cumulative distribution function

The CDF of the continuous Uniform(a,b) distribution can be determined as

$$P_X(x) = \begin{cases} 0 & \text{for } x \leq a \\ \frac{x-a}{b-a} & \text{for } a < x < b \\ 1 & \text{for } x \geq b \end{cases} \tag{B.9}$$

Proof.

$$\begin{aligned} P_X(x) &= \int_{-\infty}^{x} p_X(n)dn, \\ &= \begin{cases} 0, & \text{for } x \leq a \\ \int_a^x \frac{1}{b-a}dn, & \text{for } a < x < b \\ 1, & \text{for } x \geq \end{cases} \\ &= \begin{cases} 0 & \text{for } x \leq a \\ \frac{x-a}{b-a} & \text{for } a < x < b \\ 1 & \text{for } x \geq b \end{cases} \end{aligned} \tag{B.10}$$

□

B.2.3 Expected value

The expected value of the continuous Uniform(a,b) distribution can be determined as

$$\mathrm{E}[X] = \frac{a+b}{2}. \tag{B.11}$$

Proof.

$$\begin{aligned} \mathrm{E}[X] &= \int_{-\infty}^{\infty} x \cdot p_X(x)dx, \\ &= \int_a^b x\frac{1}{b-a}dx, \\ &= \frac{1}{b-a}\left[\frac{1}{2}x^2\right]_a^b, \\ &= \frac{1}{b-a}\cdot\left(\frac{1}{2}b^2 - \frac{1}{2}a^2\right) = \frac{(b-a)(b+a)}{2(b-a)} = \frac{a+b}{2}. \end{aligned} \tag{B.12}$$

□

B.2.4 Variance

The variance of the continuous Uniform(a,b) distribution can be determined as

$$\mathrm{Var}[X] = \frac{1}{12}(b-a)^2. \tag{B.13}$$

Proof.

$$\begin{aligned}
\mathrm{Var}[X] &= \mathrm{E}[X^2] - \mathrm{E}[X]^2, \\
&= \int_{-\infty}^{\infty} x^2 \cdot p_X(x)dx - \frac{(a+b)^2}{4}, \\
&= \int_a^b x^2 \frac{1}{b-a} dx - \frac{(a+b)^2}{4}, \\
&= \frac{1}{b-a}\left[\frac{1}{3}x^3\right]_a^b - \frac{(a+b)^2}{4}, \\
&= \frac{1}{b-a}\left(\frac{1}{3}b^3 - \frac{1}{3}a^3\right) - \frac{(a+b)^2}{4}, \\
&= \frac{b^3-a^3}{3(b-a)} - \frac{(a+b)^2}{4}, \\
&= \frac{4b^3-4a^3}{12(b-a)} - \frac{3(a^2+b^2+2ab)(b-a)}{12(b-a)}, \\
&= \frac{4b^3-4a^3-3a^2b+3a^3-3b^3+3b^2a-6ab^2+6a^2b}{12(b-a)}, \\
&= \frac{b^3-a^3+3a^2b-3ab^2}{12(b-a)} = \frac{(b-a)^3}{12(b-a)} = \frac{1}{12}(b-a)^2.
\end{aligned} \tag{B.14}$$

□

B.3 Normal or Gaussian distribution

The Normal or Gaussian distribution is probably the most commonly used continuous probability distribution. The distribution is bell-shaped and symmetric. The function is characterized by its mean μ and its variance σ^2.

B.3.1 Standard normal distribution

The Standard normal distribution is a specific case of the Normal or Gaussian distribution, where the mean equals $\mu = 0$ and the variance equals $\sigma^2 = 1$. This function can be regarded as the normalized Gaussian distribution. Any random variable $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$ can be transformed to a random variable X under the Standard normal distribution by subtracting its mean and dividing by the standard deviation as $X = \frac{Y-\mu_Y}{\sigma_Y}$.

B.3.2 Q-function

The Q-function is a commonly used function in statistics, which calculates the probability of a Standard normal distributed random variable X exceeding a certain threshold x. It is also known as the right-tail probability of the Gaussian distribution, since it is calculated by integrating the right side of the Gaussian PDF from the threshold x up to ∞. The Q-function is defined as

$$Q(x) = \Pr[X > x] = \frac{1}{\sqrt{2\pi}} \int_x^{\infty} e^{-\frac{u^2}{2}} du. \tag{B.15}$$

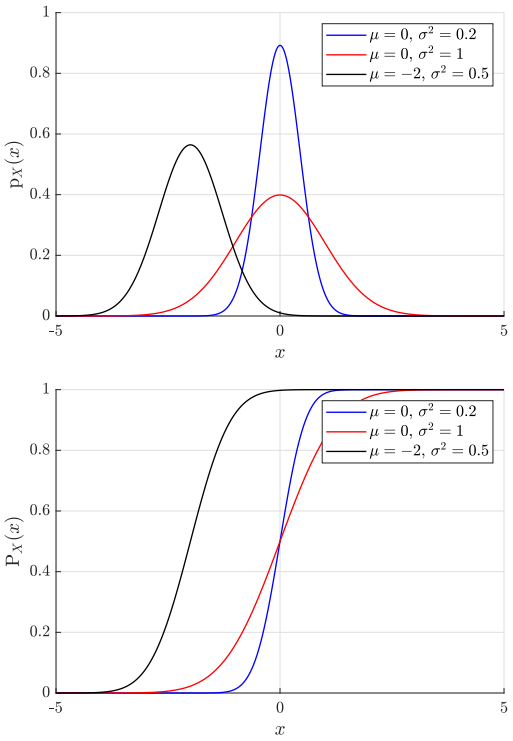


Figure B.3: Example plot of the (a) probability mass function and (b) cumulative density function of the Gaussian(μ, σ^2) distribution.

The function can be used for all Gaussian distributed random variables, however, the random variable and the corresponding threshold should be normalized first. Additionally, through symmetry follows that $Q(x) = 1 - Q(-x)$, where $Q(-x)$ is equal to the cumulative density function $P_X(x)$.

B.3.3 Probability density function

The PDF of the continuous Gaussian $\mathcal{N}(\mu, \sigma^2)$ distribution is given as

$$p_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \tag{B.16}$$

where $\sigma > 0$.

B.3.4 Cumulative distribution function

The CDF of the continuous Gaussian $\mathcal{N}(\mu, \sigma^2)$ distribution can be determined as

$$P_X(x) = Q(-\frac{x-\mu}{\sigma}) \tag{B.17}$$

Proof.

$$\begin{aligned} P_X(x) &= \int_{-\infty}^{x} p_X(n)dn, \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{x} e^{-\frac{(n-\mu)^2}{2\sigma^2}} dn, \\ &= Q(-\frac{x-\mu}{\sigma}) \end{aligned} \tag{B.18}$$

□

B.3.5 Expected value

The expected value of the continuous Gaussian $\mathcal{N}(\mu, \sigma^2)$ distribution can be determined as

$$\mathrm{E}[X] = \mu. \tag{B.19}$$

Proof. The expected value of the continuous Gaussian $\mathcal{N}(\mu, \sigma^2)$ distribution can be determined as

$$\begin{aligned} \mathrm{E}[X] &= \int_{-\infty}^{\infty} x \cdot p_X(x)dx, \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} xe^{-\frac{(x-\mu)^2}{2\sigma^2}} dx, \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} ((x-\mu)+\mu)e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx, \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} (x-\mu)e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx + \mu \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx, \\ &= \frac{-\sigma^2}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \frac{-2(x-\mu)}{2\sigma^2} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx + \mu, \\ &= \frac{-\sigma^2}{\sqrt{2\pi\sigma^2}} \left[e^{-\frac{(x-\mu)^2}{2\sigma^2}} \right]_{-\infty}^{\infty} + \mu, \\ &= \frac{-\sigma^2}{\sqrt{2\pi\sigma^2}} (0-0) + \mu = \mu. \end{aligned} \tag{B.20}$$

□

B.3.6 Variance

The variance of the continuous Gaussian $\mathcal{N}(\mu, \sigma^2)$ distribution can be determined as

$$\mathrm{Var}[X] = \sigma^2. \tag{B.21}$$

Proof.

$$\begin{aligned}
\mathrm{Var}[X] &= \mathrm{E}[(X-\mu)^2], \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} (x-\mu)^2 e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx, \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} (x-\mu)^2 e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx, \\
&= \frac{\sigma^2}{\sqrt{\pi}} \int_{-\infty}^{\infty} w \cdot 2we^{-w^2} dw, \\
&= \frac{\sigma^2}{\sqrt{\pi}} \left[-we^{-w^2}\right]_{-\infty}^{\infty} - \frac{\sigma^2}{\sqrt{\pi}} \int_{-\infty}^{\infty} -e^{-w^2} dw, \\
&= \frac{\sigma^2}{\sqrt{\pi}}(0-0) + \frac{\sigma^2}{\sqrt{\pi}}\sqrt{\pi} = \sigma^2.
\end{aligned} \tag{B.22}$$

□

As is implicated by the central limit theorem (explained in the section Function and pairs of random variables), the Gaussian distribution is extremely important. The Gaussian distribution is often used to model measurements in practice and, thanks to the CLT, its use can often be extended to other distributions. A Gaussian distribution is also often used to model the thermal noise of a band-limited system. This section will generalize the definition of the Gaussian distribution given in the previous reader and extend it to the multivariate case.

Appendix C

Useful functions

C.1 Dirac delta pulse function

The Dirac delta pulse function is a function that only exists for one single value on its domain. The value at this point is undefined, but it guarantees that the integral over its domain equals 1. The dirac delta pulse is therefore defined as

$$\delta(x) = \begin{cases} +\infty, & \text{for } x = 0 \\ 0, & \text{for } x \neq 0 \end{cases} \tag{C.1}$$

under the constraint

$$\int_{-\infty}^{\infty} \delta(x)\mathrm{d}x = \int_{0^-}^{0^+} \delta(x)\mathrm{d}x = 1, \tag{C.2}$$

where 0^- and 0^+ are the limits towards 0 from below and above respectively.

C.1.1 Sifting property

The definition of the Dirac delta pulse function allows us to extract values from a function by simply multiplying with the delta pulse and integrating over the domain. This delta pulse can be shifted on its domain, allowing for the extraction at any point on the function. This process can be regarded as sampling. This is known as the sifting property and is defined as

$$\int_{-\infty}^{\infty} g(x)\delta(x - x_0)\mathrm{d}x = \int_{x_0^-}^{x_0^+} g(x)\delta(x - x_0)\mathrm{d}x = g(x_0). \tag{C.3}$$

C.2 Unit step function

The unit step function is defined as

$$u(x) = \begin{cases} 1, & \text{for } x \geq 0, \\ 0, & \text{for } x < 0. \end{cases} \tag{C.4}$$

This function allows us to put a known probability density function $p_X(x)$ to zero up to a certain threshold x_0, whilst retaining the other part of the function. A generalization of such a

new probability density function $p_{X1}(x)$ can be written as

$$p_{X1}(x) = c \cdot p_X(x) \cdot u(x - x_0), \tag{C.5}$$

where c is a constant that is required to satisfy the total probability axiom ($\int_{-\infty}^{\infty} p_{X1}(x)\mathrm{d}x = 1$).

Appendix D

Fourier transform

A brief summary of the different versions of the Fourier transform is presented hereafter.

D.1 Fourier series

The Fourier series expansion is a method in which a periodical continuous-time signal is decomposed in sinusoidal signals with harmonically related frequencies. The signal is split into complex phasors with frequency $f = kF_0$, where k is an integer and F_0 is the fundamental frequency. The Fourier series coefficients α_k contain the amplitude and phase information of the k^{th} harmonic. The Fourier series and the inverse operation are defined as:

$$\alpha_k = \frac{1}{T_0}\int_0^{T_0} x(t)e^{-j2\pi F_0 kt}dt \Longleftrightarrow x(t) = \sum_{k=-\infty}^{\infty} \alpha_k e^{j2\pi F_0 kt} \tag{D.1}$$

D.2 Fourier transform

Whereas the Fourier series applies to periodical signals, the Fourier transform applies to non-periodic signals. The frequency variable f is a continuous variable. The Fourier transform for continuous-time signals is defined as:

$$X(f) = \int_{-\infty}^{\infty} x(t)e^{-j2\pi ft}dt \Longleftrightarrow x(t) = \int_{-\infty}^{\infty} X(f)e^{j2\pi ft}df \tag{D.2}$$

D.3 Discrete-time Fourier transform

The application of the Fourier transform is usually limited, since continuous signals cannot be directly recorded on a computer. Instead, these signals are sampled to the discrete-time domain. The Fourier transform that is capable of dealing with discrete-time signals is called the discrete-time Fourier transform and is defined as:

$$X(e^{j\theta}) = \sum_{n=-\infty}^{\infty} x[n]e^{-jn\theta} \Longleftrightarrow x[n] = \frac{1}{2\pi}\int_{-\pi}^{\pi} X(e^{j\theta})e^{jn\theta}d\theta \tag{D.3}$$

Note that the frequency variable θ is 2π periodic.

D.4 Discrete Fourier transform

The discrete-time Fourier transform is defined for signals with infinite length. On a computer, however, we are dealing with finite records of a signal. The Fourier transform for this finite length discrete-time signal is the discrete Fourier transform. The discrete Fourier transform calculates an equidistantly sampled version of the discrete-time Fourier transform. The discrete-time Fourier transform is defined as:

$$X_p[k] = \sum_{n=0}^{N-1} x_p[n] e^{-j\frac{2\pi}{N}kn} \Longleftrightarrow x_p[n] = \frac{1}{N} \sum_{k=0}^{N-1} X_p[k] e^{j\frac{2\pi}{N}kn} \tag{D.4}$$

Appendix E

Linear time-invariant systems

This appendix provides a brief recap of the general form and properties of linear time-invariant systems.

E.1 Linearity and time-invariance

A linear time-invariant system is a system having both the properties of linearity and time-invariance, which are described hereafter.

E.2 Linearity

A system is regarded to be linear if it is both additive and homogeneous. Suppose that we have a system, which we provide separately with two distinct input signals $x_1[n]$ and $x_2[n]$. These signals are individually transformed by the system into two output signals, $y_1[n]$ and $y_2[n]$ respectively. This can be represented intuitively as

$$x_1[n] \to y_1[n] \quad \text{and} \quad x_2[n] \to y_2[n]. \tag{E.1}$$

This system is called additive if and only if the output of the system driven by the sum of the distinct inputs equals the sum of the individual outputs. In other words, when the individual input signals are simultaneously applied to the system, the respective outputs are also observed simultaneously. This can be represented as

$$\text{Additive:} \quad x_1[n] + x_2[n] \to y_1[n] + y_2[n]. \tag{E.2}$$

Furthermore, this system is called homogeneous if and only if the output of this system produces an identically scaled output when driven by a scaled input. In other words, when the input signal is scaled with some scaling coefficient c, the output is also scaled with the same scaling coefficient. This can be represented as

$$\text{Homogeneous:} \quad c \cdot x_1[n] \to c \cdot y_1[n]. \tag{E.3}$$

When a system is both additive and homogeneous, the system is linear. This is represented as

$$\text{Linear:} \quad \alpha \cdot x_1[n] + \beta \cdot x_2[n] \to \alpha \cdot y_1[n] + \beta \cdot y_2[n], \tag{E.4}$$

where α and β are scaling coefficients.

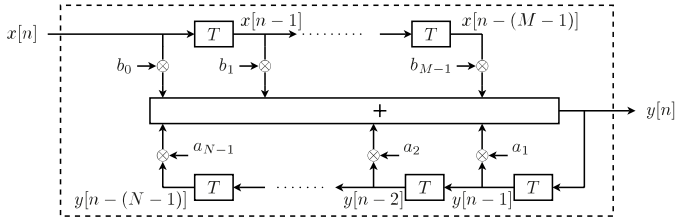


Figure E.1: Signal flow diagram of an LTI system.

E.3 Time-invariance

Another system property is time invariance. This property concerns signal delays or temporal shifts. A system is called time invariant if a temporal shift in the input signal results in the same temporal shift in the respective output signal. This can be represented as

$$\text{Time-invariance:} \quad x[n - n_0] \rightarrow y[n - n_0], \tag{E.5}$$

where n_0 represents the temporal shift.

E.4 System architecture

A general LTI system can be represented using a so-called signal flow diagram as in Figure E.1. The system is driven by an input signal $x[n]$ and outputs the signal $y[n]$. The output is obtained by the addition of two main branches. The upper branch, which is connected to the input signal, transforms the input and delayed inputs into an output. This branch is also called the moving-average part of the system. The lower branch, which is connected to the output signal, transforms delayed outputs into a new output. This branch is also called the autoregressive part of the system.

E.5 Difference equation

The operations of the system can be represented mathematically by the weighted sum of the current input and delayed versions of the input and outputs. This mathematical description is called the difference equation. The output $y[n]$ at sample index n can be calculated as

$$y[n] = \underbrace{\sum_{k=0}^{M-1} b_k x[n-k]}_{\text{moving-average}} + \underbrace{\sum k = 1^{N-1} a_k y[n-k]}_{\text{autoregressive}}, \tag{E.6}$$

where two distinct terms can be distinguished. First, there is the moving-average part, which represents the upper branch of Figure E.1, and then it follows the autoregressive part, which represents the lower branch of Figure E.1. The weights are represented by the coefficients b_k and a_k, respectively.

E.6 Impulse response

Besides the difference equation, there is another way of describing the functioning of a system. This representation is called the impulse response and will prove useful in the following sections.

The impulse response is the output, or response, of a system when driven by a short impulse $\delta[n]$ and is denoted by $h[n]$. This can be described intuitively as

$$\delta[n] \rightarrow h[n], \tag{E.7}$$

where $\delta[n]$ represents the Dirac delta pulse, which is defined as

$$\delta[n] = \begin{cases} 1, & \text{for } n = 0, \\ 0, & \text{elsewhere.} \end{cases} \tag{E.8}$$

Another way of representing the definition of the impulse response is

$$h[n] = y[n]\Big|_{x[n]=\delta[n]}, \tag{E.9}$$

which simply states that the impulse response is the output of a system when the input of that system is given by a Dirac delta pulse.

The impulse response can be metaphorically compared with a tuning fork. A tuning fork is used to tune an instrument. The fork is hit against another object, after which it will resonate at a specific frequency. This frequency is used as a guideline for the tuning of an instrument. This can be compared to the discussion of the impulse response. The system (the tuning fork) is excited with an impulse (hitting it against a solid object) and will output a signal, also called the impulse response (the sound signal at the resonant frequency).

E.7 FIR and IIR filters

From the representation of the general LTI system, two different types of systems can be distinguished. These two types are finite impulse response (FIR) filters and infinite impulse response (IIR) filters. As the name indicates, the distinction is based on the length of the impulse response.

FIR filters are discussed in more detail here. FIR filters are the class of filters in which all autoregressive coefficients a_k are equal to zero. These filters effectively only contain the upper branch of the signal flow diagram in Figure E.1. When a FIR filter is excited by an impulse, this passes through the upper branch and is delayed a total of $M-1$ times; the delayed samples are weighted and summed to produced the output of the system. Mathematically, the impulse response of an FIR filter can be determined by substituting $\delta[n]$ for $x[n]$ in (E.6) and by setting all autoregressive coefficients a_k to zero as

$$\begin{aligned} h[n] &= \sum_{k=0}^{M-1} b_k \cdot \delta[n-k] + \sum_{k=1}^{N-1} 0 \cdot y[n-k], \\ &= \sum_{k=0}^{M-1} b_k \cdot \delta[n-k] = b_0\delta[n] + b_1\delta[n-1] + ... + b_n\delta[n-M-1]. \end{aligned} \tag{E.10}$$

The last step in this equation can be understood by noting that $\delta[n-k]$ only equals 1 when $n = k$ holds. From this it can be noted that the length of the impulse response depends on the number of moving-average weights b_k. The number of moving-average weights b_k depends on the length of the upper branch of the filter, which should be finite to allow for a practical implementation. Therefore, the number of moving-average weights is also finite, leading to finite impulse response given by (E.10).

IIR filters on the other hand, contain non-zero autoregressive coefficients. Intuitively one could understand the consequence of this by paying close attention to the created feedback loop in the system, corresponding to the lower branch in the signal flow diagram in Figure E.1. Suppose an impulse is applied to the input of an IIR filter. This signal will propagate through the moving-average branch (or directly, when only b_0 is non-zero) to the output. Since at least one of the autoregressive coefficients is non-zero, this output will pass through the autoregressive branch back to the output and the signal will therefore keep propagating through the feedback loop and will keep generating outputs. By applying a single impulse at the input of the system, the system will keep generating outputs, leading to an infinitely long impulse response.

In the following example we will derive the impulse response of a simple system with a non-zero autoregressive coefficient.

Example E.1

Calculate the analytical impulse response of a first-order autoregressive filter.

A first-order autoregressive filter is defined as a system where the moving-average coefficients are all zero, except for b_0, which equals 1, and where all autoregressive coefficients are zero, except for a_1. Without the coefficient b_0, no signal would propagate to the output of the system and therefore there would be no influence of the autoregressive portion. This definition gives rise to the corresponding difference equation

$$y[n] = x[n] + a_1 y[n-1].$$

Let us first calculate the first outputs of the system when driven by an input $x[n]$ to get some intuition of an autoregressive process. If an arbitrary input signal $x[n]$ would be applied to the system, the first output of the system would be calculated as

$$y[0] = x[0] + a_1 y[-1] = x[0] + a_1 0 = x[0],$$

where the signal output is assumed to be 0 for $n < 0$ From this first output of the system, the second output can be determined as

$$y[1] = x[1] + a_1 y[0] = x[1] + a_1 x[0]$$

and the third output as

$$y[2] = x[2] + a_1 y[1] = x[2] + a_1 \left(x[1] + a_1 x[0]\right) = x[2] + a_1 x[1] + a_1^2 x[0].$$

Through these iterations the reader should be able to see some structure in the output signal. Using this structure an equation for the output $y[n]$ can be found as

$$y[n] = \begin{cases} \sum^{n} k = 0 a_1^k x[n-k], & \text{for } n \geq 0, \\ 0, & \text{for } n < 0. \end{cases} \tag{E.11}$$

By using the definition of the unit step function $u[n]$, this can also be written as

$$y[n] = u[n] \sum^{n} k = 0 a_1^k x[n-k].$$

Since the impulse response of the system is defined as the output of the system when driven by a Dirac delta pulse, i.e. $x[n] = \delta[n]$, the impulse response can be found as

$$h[n] = u[n] \cdot a_1^n = \begin{cases} a_1^n, & \text{for } n \geq 0, \\ 0, & \text{for } n < 0. \end{cases} \tag{E.12}$$

Based on the value of a_1^n the impulse response will have a different shape. The figure below shows several examples of possible impulse response for different values of a_1^n.

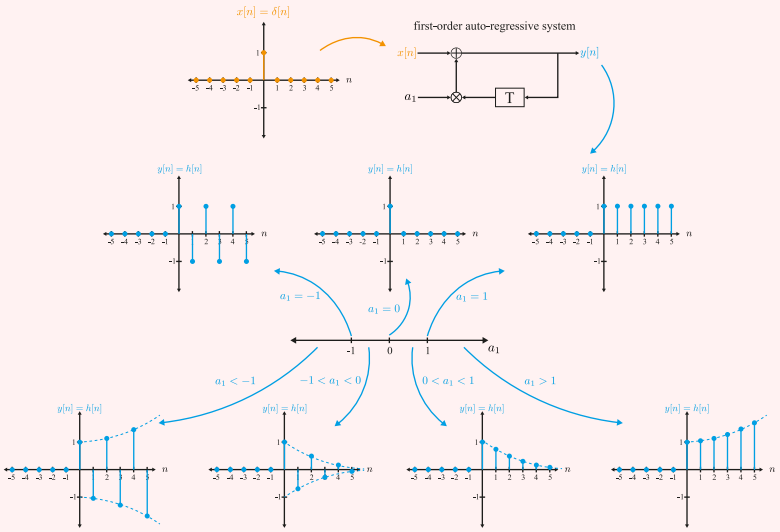


Figure E.2: Different impulse responses for a first-order autoregressive filter, depending on the value of the autoregressive coefficient a_1.

E.8 System invertibility

A system $H(z)$ is invertible if the output $y[n](n = (-\infty, \infty))$ can be used to uniquely define the input $x[n](n = (-\infty, \infty))$. This is only possible if every unique value in $x[n]$ maps to a unique value in $y[n]$. This is called one-to-one mapping. Obtaining the inverse for any system is very difficult. However, if the system is linear time-invariant with an impulse response $h[n]$, the inverse $h_{inv}[n]$ can be defined as follows

$$\begin{aligned} x[n] * h[n] &= y[n] \\ y[n] * h_{inv}[n] &= x[n] \\ x[n] &= (x[n] * h[n]) * h_{inv}[n] \end{aligned}$$

This can be simplified by utilizing the z-transform:

$$\begin{aligned} X(z)H(z) &= Y(z) \\ Y(z)H_{inv}(z) &= X(z) \\ X(z) &= X(z)H(z)H_{inv}(z) \\ H_{inv}(z) &= \frac{1}{H(z)} \end{aligned}$$

If the systems consists of poles and zeros, than it can be split into its denominator and numerator:

$$\begin{aligned} H(z) &= \frac{N(z)}{D(z)} \\ H_{inv}(z) &= \frac{D(z)}{N(z)} \end{aligned}$$

Thus, the poles of the system become zeros in its inverse and vice versa. However, the inverse is not always uniquely defined for every system, as shown by the following example.

Example E.2

Consider a system with the impulse response $h[n] = \delta[n] - \frac{1}{4}\delta[n-1]$. By performing the z-transform we end up with $H(z) = 1 - \frac{1}{4}z^{-1}$. Therefore, the inverse is equal to $H_{inv}(z) = \frac{1}{1-\frac{1}{4}z^{-1}}$, which has one pole at $z = \frac{1}{4}$. If we choose the region of convergence (ROC) as $|z| > \frac{1}{4}$, the inverse system is causal and stable, and

$$h_{inv}[n] = (\frac{1}{4})^n u[n].$$

However, if we choose the ROC as $|z| < \frac{1}{4}$, the inverse system is noncausal and unstable

$$h_{inv}[n] = -(\frac{1}{4})^n u[-n-1]$$

E.9 All-pass filter

An all-pass filter let all frequencies pass with the same magnification factor, that is the magnitude of the frequency response is constant

$$|H_{ap}(e^{j\theta})| = 1 \quad \forall\theta$$

Constraining the system gain to be unitary implies that the poles and zeros of a rational system function to occur in mirrored pairs. Thus if $H(z)$ has a pole $z = \alpha_k$, $H(z)$ must also have a zero at the mirrored location $z = 1/\alpha_k^*$.

$$\text{Complex } h[n] \quad : \quad H_{ap}(z) = \prod_{k=1}^{p} \frac{z^{-1} - \alpha_k^*}{1 - \alpha_k z^{-1}}$$

$$\text{Real } h[n] \quad : \quad H_{ap}(z) = \prod_{k=1}^{N_s} \frac{|\alpha_k|^2 - 2\Re e\{\alpha_k\} z^{-1} + z^{-2}}{1 - 2\Re e\{\alpha_k\} z^{-1} + |\alpha_k|^2 z^{-2}}$$

The poles of a stable and causal all-pass filter $H(z)$ lie inside the unit circle, that is all $|\alpha_k| < 1$. If the impulse response $h[n]$ is real-valued, the complex roots occur in conjugate pairs, and these conjugate pairs can be combined to form second-order factors from which all coefficients are real, as shown in the equation above.

Example E.3

Show that the following system is all pass:

$$H(e^{j\theta}) = \frac{\frac{1}{2} - e^{-j\theta} + e^{-j2\theta}}{1 - e^{-j\theta} + \frac{1}{2}e^{-j2\theta}}$$

Give the pole zero plot and a rough sketch of the magnitude and phase response plots.

By replacing $e^{-j\theta}$ with the complex variable z^{-1} we obtain the following system

function:

$$H(z) = \frac{\frac{1}{2} - z^{-1} + z^{-2}}{1 - z^{-1} + \frac{1}{2}z^{-2}}$$

Because of the special structure of this equation, the absolute value results in:

$$\begin{aligned} |H(z)| &= \sqrt{H(z) \cdot (H(z))^*} = \sqrt{H(z) \cdot H^*(z^{-1})} \\ &= \sqrt{\left(\frac{\frac{1}{2} - z^{-1} + z^{-2}}{1 - z^{-1} + \frac{1}{2}z^{-2}}\right) \cdot \left(\frac{\frac{1}{2} - z + z^2}{1 - z + \frac{1}{2}z^2}\right)} \\ &= \sqrt{\left(\frac{z^{-2}\left(\frac{1}{2}z^2 - z + 1\right)}{z^{-2}\left(z^2 - z + \frac{1}{2}\right)}\right) \cdot \left(\frac{\frac{1}{2} - z + z^2}{1 - z + \frac{1}{2}z^2}\right)} = 1 \\ \Rightarrow \quad & |H(e^{j\theta})| = |H(z)|_{|z|=1} = 1 \end{aligned}$$

The pole and zeros are:

$$H(z) = \frac{(\frac{1}{2}\sqrt{2} - e^{j\frac{\pi}{4}}z^{-1})(\frac{1}{2}\sqrt{2} - e^{-j\frac{\pi}{4}}z^{-1})}{(1 - \frac{1}{2}\sqrt{2}e^{j\frac{\pi}{4}}z^{-1})(1 - \frac{1}{2}\sqrt{2}e^{j\frac{\pi}{4}}z^{-1})}$$

as shown at the left hand side of the following plot:

E.10 Minimum-phase systems

The simple example above illustrates that the knowledge of the impulse response of a LTI system does not uniquely specify its inverse. Additional information such as causality and stability would be helpful in many cases. This leads us to the concept of minimum-phase systems.

A minimum-phase system is a system in which the system and its inverse are both stable and causal. To understand what this means we first need to know the conditions for stability and causality.

- In a stable system the ROC contains the unit circle.
- In a causal system the ROC is defined as going outward from a circle, with a radius larger than the largest pole in the system (excluding poles at infinity).
- In a anti-causal system the ROC is defined as going inward from a circle, with a radius smaller than the smallest pole (excluding poles at zero).

Therefore, if a systems is minimum-phase, its poles and zeros have the following properties. The ROC extends outwards from a circle with a radius larger than the largest pole, since the system needs to be causal. The system also needs to be stable. Therefore, the ROC includes the unit circle. Thus, the largest pole, and therefore all poles, need to be within the unit circle. The zeros need to be within the unit circle as well, since these rules also apply to the inverse.

A stable and causal LTI system has all its poles inside the unit circle. The zeros, however may lie anywhere in the z-plane. In order to obtain a minimum-phase system, it is necessary to constraint the system $H(z)$ so that its inverse $G(z) = 1/H(z)$ is also stable and causal. This requires that the zeros of $H(z)$ lie also inside the unit circle.

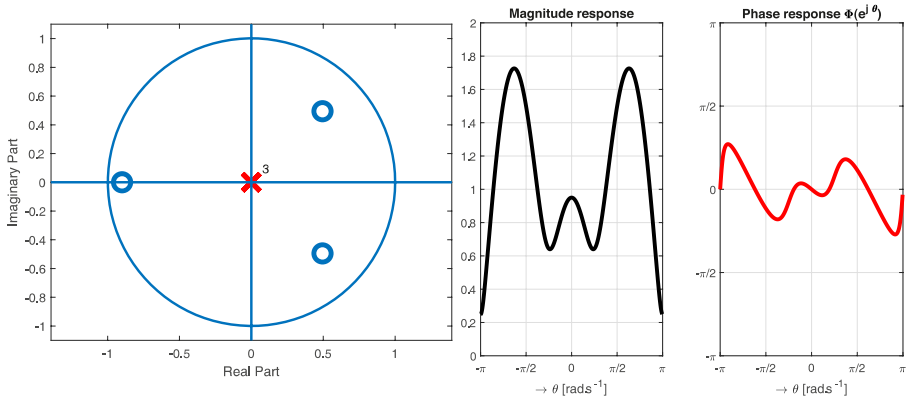


Figure E.4: Pole-zero plot and magnitude- and phase-response of minimum-phase system.

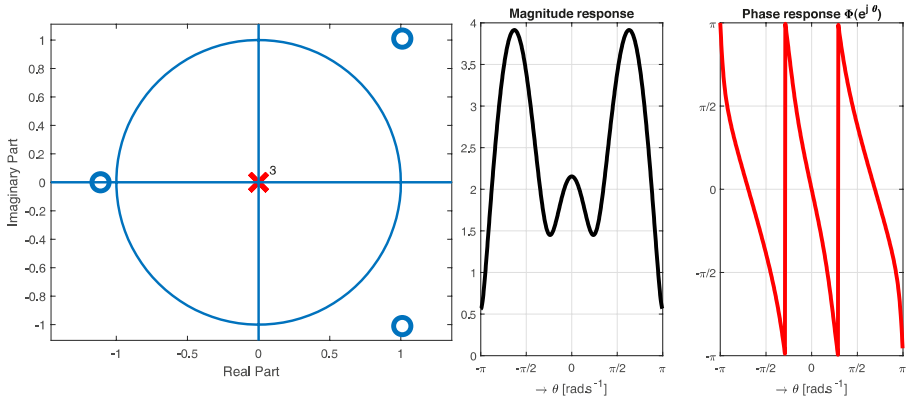


Figure E.5: Pole-zero plot and magnitude- and phase-response of maximum phase system.

In other words, a stable and causal filter that has a stable and causal inverse is said to have minimum-phase and so a minimum-phase system has all of its poles and zeros inside the unit circle, as shown in Figure E.4.

The magnitude and phase response of a minimum-phase system with three zeros inside the unit circle is depicted in Figure E.4. When mirroring all zeros and poles of a stable and causal minimum-phase system we obtain a maximum phase system; thus, a maximum phase system has all of its poles and zeros outside the unit circle, as shown in Figure E.5.

Figure E.5. shows the magnitude and phase response of a maximum phase system with three zeros outside the unit circle as depicted in the pole-zero plot. These zeros are obtained by mirroring all three zeros of the previous minimum-phase system.

- Mirroring a zero with respect to the unit circle does not change the shape of the magnitude response. Thus both minimum and maximum phase systems have, besides a constant factor, the same shape of the magnitude response characteristic.

- The phase of a minimum-phase system has less variations compared to the phase variations of a maximum phase system.

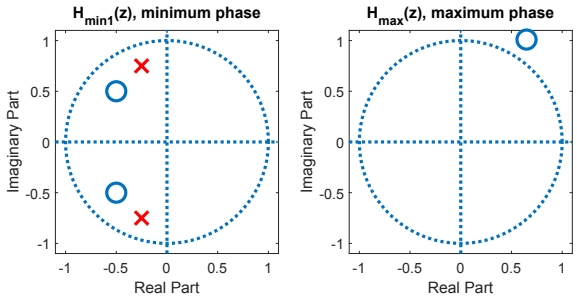


Figure E.6: Pole-zero plot of H(z). Pole zero plot of $H(z)$, decomposed as a cascade of a minimum-phase and a maximum phase systems.

- The group delay is defined as:

$$\tau(\theta) = -\frac{\mathrm{d}\varphi\{H(e^{j\theta})\}}{\mathrm{d}\theta},$$

  The group delay of a minimum-phase system is minimal.

E.11 minimum-phase and all-pass decomposition

With the knowledge that we have about minimum-phase and all-pass systems, we are able to show that any causal pole-zero system with system function $H(z)$ (without poles or zeros on the unit circle) can be decomposed as the product of an all-pass and a minimum-phase system.

Let $H(z)$ be a nonminimum-phase system with one zero $z = \frac{1}{a}, |a| < 1$, outside the unit circle and all other poles and zeros on the inside of the unit circle. To decompose the system the steps to follow are:

- Factorize out all poles and zeros outside of the unit circle to create a minimum-phase and maximum phase system. For our example $H(z)$ can be rewritten as:

$$H(z) = H_{min1}(z)(a - z^{-1}) = H_{min1}(z)H_{max}(z),$$

  where H_1 is minimum-phase.

- Create all the conjugate reciprocals of all poles and zeros that were factorized out. However, you cannot just add poles and zeros to a system for free. You will also need to add a zero on the same spot for every pole you add and vice versa. In our example, this becomes

$$H(z) = H_{min1}(z)(a - z^{-1})\frac{1 - a^* z^{-1}}{1 - a^* z^{-1}} = H_{min1}(z)H_{mix}(z)$$

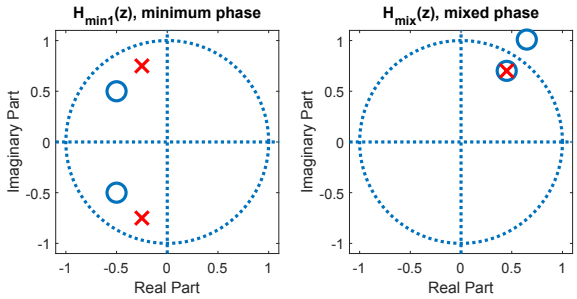


Figure E.7: Pole-zero plot of H(z). Pole zero plot of $H(z)$, decomposed as a cascade of a minimum-phase and a mixed phase systems.

- Match all the factored out poles and zeros with their conjugate reciprocal to obtain the all-pass part, and add the remaining poles and zeros the minimum-phase system part. Finally, we obtain

$$H(z) = [H_{min1}(z)(1 - a^* z^{-1})]\frac{a - z^{-1}}{1 - a^* z^{-1}}$$
$$H(z) = H_{min}(z)H_{ap}(z)$$

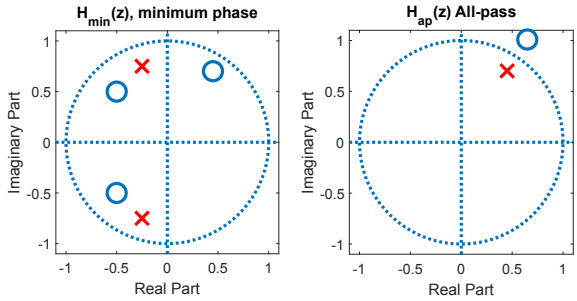


Figure E.8: Pole-zero plot of H(z). Pole zero plot of $H(z)$, decomposed as a cascade of a minimum-phase and an all-pass systems.

In the following exercise, you can try for yourself to apply these steps and obtain a minimum-phase all-pass decomposition.

Example E.4

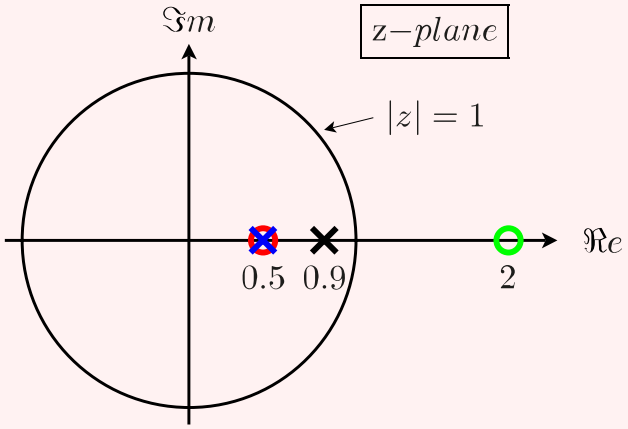
Factor the following system function as the product of a minimum-phase system and an all-pass system: $H(z) = \frac{1-2z^{-1}}{1-0.9z^{-1}}$

Solution.

This can be shown as follows:

$$\begin{aligned} H(z) &= \frac{1-2z^{-1}}{1-0.9z^{-1}} \cdot \frac{1-\frac{1}{2}z^{-1}}{1-\frac{1}{2}z^{-1}} \\ &= \frac{1-\frac{1}{2}z^{-1}}{1-0.9z^{-1}} \cdot \frac{1-2z^{-1}}{1-\frac{1}{2}z^{-1}} \\ &= H_{min}(z) \cdot H_{ap}(z) \end{aligned}$$

First create a zero by mirroring the zero, which results in a new zero at $z = 0.5$. This new zero has to be compensated by a new pole at the same position $z = 0.5$. Then rearrange all old and new poles and zeros of $H(z)$ into two factors. The first factor contains the old pole and the new zero, which are both inside the unit circle and so this factor is minimum-phase. The second factor contains the old zero and new pole which are each others mirrored versions and so this factor is all pass.



These steps can be applied to any filter $H(z)$.