Notes on Measure, Probability and Stochastic Processes

João Lopes Dias

Departamento de Matemática, ISEG, Universidade de Lisboa, Rua do Quelhas 6, 1200-781 Lisboa, Portugal

Email address: jldias@iseg.ulisboa.pt

September 24, 2019.

Contents

CHAPTER 1

Introduction

These are the lecture notes for the course "Probability Theory and Stochastic Processes" of the Master in Mathematical Finance (since 2016/2017) at ISEG–University of Lisbon. It is required good knowledge of calculus and basic probability.

1. Classical definitions of probability

Science is about observing given phenomena, recording data, analysing it and explaining particular features and behaviours using theoretical models. This may be a rough description of what really means to make science, but highlights the fact that experimentation is a crucial part of obtaining knowledge.

Many experiments are of random nature. That is, their results are not possible to predict, often due to the huge number of variables that underlie the processes under scrutiny. One needs to repeat the experiment and observe its different outcomes and events (i.e. sets of possible outcomes). Our main goal is to quantify the likelihood of each event.

These general ideas can be illustrated by the experiment of throwing dice. We can get six possible outcomes depending on too many different factors, so that it becomes impossible to predict the result. Consider the event corresponding to an even number of dots, i.e. 2, 4 or 6 dots. How can we measure the probability of this event to occur when we throw the dice once? If the dice are fair (unbiased), intuition tells us that it is equal to $\frac{1}{2}$.

The way one usually thinks of probability is summarised in the following relation:

$$\text{Probability}(\text{“event”}) = \frac{\text{number of favourable cases}}{\text{number of possible cases}}$$

assuming that all cases are equally possible. This is the classical definition of probability, called the *Laplace law*.

Example 1.1. Tossing of a perfect coin in order to get either heads or tails. The number of possible cases is 2. So,

$$\text{Prob}(\text{“heads”}) = 1/2$$
$$\text{Prob}(\text{“heads at least once in two experiments”}) = 3/4.$$

Example 1.2.

$$\text{P}(\text{“winning the Euromillions with one bet”}) = \frac{1}{C_5^{50} C_2^{12}} \simeq 7 \times 10^{-9}.$$

The Laplace law has the following important consequences:

(1) For any event S, $0 \leq P(S) \leq 1$.
(2) If $P(S) = 1$, then S is a safe event. If $P(S) = 0$, S is an impossible event.
(3) $P(\text{not } S) = 1 - P(S)$.
(4) If A and B are disjoint events, then $P(A \text{ or } B) = P(A) + P(B)$.
(5) If A and B are independent, then $P(A \text{ and } B) = P(A)\, P(B)$.

The first mathematical formulation of the probability concept appeared in 17th century France. A gambler called Antoine Gombauld realized empirically that he was able to make money by betting on getting at least a 6 in 4 dice throwings. Later he thought that betting on getting at least a pair of 6's by throwing two dice 24 times was also advantangeous. As that did not turn out to be the case, he wrote to Pascal for help. Pascal and Fermat exchanged letters in 1654 discussing this problem, and that is the first written account of probability theory, later formalized and further expanded by Laplace.

According to Laplace law, Gombauld's first problem can be described mathematically as follows. Since

$$P(\text{not get 6 in one atempt}) = \frac{5}{6}$$

and each dice is independent of the others, then

$$P(\text{not get 6 in 4 attempts}) = \left(\frac{5}{6}\right)^4 \simeq 0.482.$$

Therefore, in the long run Gombauld was able to make a profit[1]:

$$P(\text{get 6 in 4 attempts}) = 1 - \left(\frac{5}{6}\right)^4 \simeq 0.518 > \frac{1}{2}.$$

However, for the second game,

$$P(\text{no pair of 6's out of 2 dice}) = \frac{35}{36}$$

[1]Test your luck at http://random.org/dice/

and

$$P(\text{no pair of 6's out of 2 dice in 24 attempts}) = \left(\frac{35}{36}\right)^{24} \simeq 0,508.$$

This time he did not have an advantage as

$$P(\text{pair of 6's out of 2 dice in 24 attempts}) \simeq 1 - 0,508 = 0,492 < \frac{1}{2}.$$

Laplace law is far from what one could consider as a useful definition of probability. For instance, we would like to examine also "biased" experiments, that is, with unequally possible outcomes. A way to deal with this question is defining probability by the frequency that some event occurs when repeating the experiment many times under the same conditions. So,

$$\mathrm{P}(\text{"event"}) = \lim_{n\to+\infty} \frac{\text{number of favourable cases in } n \text{ experiments}}{n}.$$

Example 1.3. In 2015 there was 85500 births in Portugal and 43685 were boys. So,

$$\mathrm{P}(\text{"it's a boy!"}) \simeq 0.51.$$

A limitation of this second definition of probability occurs if one considers infinitely many possible outcomes. There might be situations were the probability of every event is zero!

Modern probability is based in measure theory, bringing a fundamental mathematical rigour and an abrangent concept (although very abstract as we will see). This course is an introduction to this subject.

Exercise 1.4. Gamblers A and B throw a dice each. What is the probability of A getting more dots than B? (5/12)

Exercise 1.5. A professor chooses an integer number between 1 and N, where N is the number of students in the lecture room. By alphabetic order each of the N students try to guess the hidden number. The first student at guessing it wins 2 extra points in the final exam. Is this fair for the students named Xavier and Zacarias? What is the probability for each of the students (ordered alphabetically) to win? (all $\frac{1}{N}$, fair).

Exercise 1.6. A gambler bets money on a roulette either even or odd. By winning he receives the same ammount that he bet. Otherwise he looses the betting money. His strategy is to bet $\frac{1}{N}$ of the total money at each moment in time, starting with €M. Is this a good strategy?

2. Mathematical expectation

Knowing the probability of every event concerning some experiment gives us a lot of information. In particular, it gives a way to compute the best prediction, namely the weighted average. Let X be the value of a measurement taken at the outcome of the experiment, a so called random variable. Suppose that X can only attain a finite number of values, say $a_1, \dots, a_n$, and we know that the probability of each event $X = a_i$ is given by $P(X = a_i)$ for all $i = 1, \dots, n$. Then, the weighted average of all possible values of X given their likelihoods of realization is naturally given by

$$E(X) = a_1 P(X = a_1) + \cdots + a_n P(X = a_n).$$

If all results are equally probable, $P(X = a_i) = \frac{1}{n}$, then $E(X)$ is just the arithmetical average.

The weighted average above is better known as the expected value of X. Other names include expectation, mathematical expectation, average, mean value, mean or first moment. It is the best option when making decisions and for that it is a fundamental concept in probability theory.

Example 1.7. Throw dice 1 and dice 2 and count the number of dots denoting them by X_1 and X_2, respectively. Their sum $S_2 = X_1 + X_2$ can be any integer number between 2 and 12. However, their probabilities are not equal. For instance, $S_2 = 2$ corresponds to a unique configuration of one dot in each dice, i.e. $X_1 = X_2 = 1$. On the other hand, $S_2 = 3$ can be achieved by two different configurations: $X_1 = 1, X_2 = 2$ or $X_1 = 2, X_2 = 1$. Since the dice are independent,

$$P(X_1 = a_1, X_2 = a_2) = P(X_1 = a_1)\, P(X_2 = a_2) = \frac{1}{36},$$

one can easily compute that

$$P(S_2 = n) = \begin{cases} \frac{n-1}{36}, & 2 \leq n \leq 7 \\ \frac{12-n+1}{36}, & 8 \leq n \leq 12 \end{cases}$$

and $E(S_2) = 7$.

Example 1.8. Toss two fair coins. If we get two heads we win €4, two tails €1, otherwise we loose €3. Moreover, $P(\text{two heads}) = P(\text{two heads}) = \frac{1}{4}$ and also $P(\text{one head one tail}) = \frac{1}{2}$. Let X be the gain for a given outcome, i.e. $X(\text{two heads}) = 4$, $X(\text{two tails}) = 1$ and $X(\text{one head one tail}) = -3$. The profit expectation for this game is therefore

$$E(X) = 4P(X = 4) + 1P(X = 1) - 3P(X = -3).$$

The probabilities above correspond to the probabilities of the corresponding events so that

$$E(X) = 4P(\text{two heads}) + 1P(\text{two tails}) - 3P(\text{one head one tail}) = -0.25.$$

It is expected that one gets a loss in this game on average, so the decision should be not to play it. This is an unfair game, one would need to have a zero expectation for the game to be fair.

The definition above of mathematical expectation is of course limited to the case X having a finite number of values. As we will see in the next chapters, the way to generalize this notion to infinite sets is by interpreting it as the integral of X with respect to the probability measure.

Exercise 1.9. Consider the throwing of three dice. A gambler wins €3 if all dice are 6's, €2 if two dice are 6's, €1 if only one dice is a 6, and looses €1 otherwise. Is this a fair game?

CHAPTER 2

Measure and probability

1. Algebras

Given an experiment we consider Ω to be the set of all possible outcomes. This is the probabilistic interpretation that we want to associate to Ω, but in the point of view of the more general measure theory, Ω is just any given set.

The collection of all the subsets of Ω is denoted by

$$\mathcal{P}(\Omega) = \{A \colon A \subset \Omega\}.$$

It is also called the set of the parts of Ω. When there is no ambiguity, we will simply write $\mathcal{P}$. We say that $A^c = \Omega \setminus A$ is the complement of $A \in \mathcal{P}$ in Ω.

As we will see later, a proper definition of the measure of a set requires several properties. In some cases, that will restrict the elements of $\mathcal{P}$ that are measurable. It turns out that the measurable ones just need to verify the following conditions.

A collection $\mathcal{A} \subset \mathcal{P}$ is an *algebra* of Ω iff

(1) $\emptyset \in \mathcal{A}$,
(2) If $A \in \mathcal{A}$, then $A^c \in \mathcal{A}$,
(3) If $A_1, A_2 \in \mathcal{A}$, then $A_1 \cup A_2 \in \mathcal{A}$.

An algebra $\mathcal{F}$ of Ω is called a *σ-algebra* of Ω iff given $A_1, A_2, \dots \in \mathcal{F}$ we have

$$\bigcup_{n=1}^{+\infty} A_n \in \mathcal{F}.$$

REMARK 2.1. We can easily verify by induction that any finite union of elements of an algebra is still in the algebra. What makes a σ-algebra different is that the infinite countable union of elements is still in the σ-algebra.

EXAMPLE 2.2. Consider the set $\mathcal{A}$ of all the finite union of intervals in $\mathbb{R}$, including $\mathbb{R}$ and $\emptyset$. Notice that the complementary of an interval is a finite union of intervals. Therefore, $\mathcal{A}$ is an algebra. However, the countable union of the sets $A_n =]n, n+1[\in \mathcal{A}$, $n \in \mathbb{N}$, is no longer finite. That is, $\mathcal{A}$ is not a σ-algebra.

REMARK 2.3. Any finite algebra $\mathcal{A}$ (i.e. it contains only a finite number of subsets of Ω) is immediately a σ-algebra. Indeed, any infinite union of sets is in fact finite.

The elements of a σ-algebra $\mathcal{F}$ of Ω are called *measurable sets*. In probability theory they are also known as *events*. The pair $(\Omega, \mathcal{F})$ is called a *measurable space*.

EXERCISE 2.4. Decide if $\mathcal{F}$ is a σ-algebra of Ω where:

(1) $\mathcal{F} = \{\emptyset, \Omega\}$.
(2) $\mathcal{F} = \mathcal{P}(\Omega)$.
(3) $\mathcal{F} = \{\emptyset, \{1,2\}, \{3,4,5,6\}, \Omega\}$, $\Omega = \{1,2,3,4,5,6\}$.
(4) $\mathcal{F} = \big\{\emptyset, \{0\}, \mathbb{R}^-, \mathbb{R}_0^-, \mathbb{R}^+, \mathbb{R}_0^+, \mathbb{R}\setminus\{0\}, \mathbb{R}\big\}$, $\Omega = \mathbb{R}$.

PROPOSITION 2.5. *Let $\mathcal{F} \subset \mathcal{P}$ such that it contains the complementary set of all its elements. For $A_1, A_2, \dots \in \mathcal{F}$,*

$$\bigcup_{n=1}^{+\infty} A_n \in \mathcal{F} \quad \textit{iff} \quad \bigcap_{n=1}^{+\infty} A_n \in \mathcal{F}.$$

PROOF. ($\Rightarrow$) Using Morgan's laws,

$$\bigcap_{n=1}^{+\infty} A_n = \left(\bigcup_{n=1}^{+\infty} A_n^c\right)^c \in \mathcal{F}$$

because the complements are always in $\mathcal{F}$.

($\Leftarrow$) Same idea. $\square$

Therefore, the definitions of algebra and σ-algebra can be changed to require intersections instead of unions.

EXERCISE 2.6. Let Ω be a finite set with $\#\Omega = n$. Compute $\#\mathcal{P}(\Omega)$. *Hint*: Find a bijection between $\mathcal{P}$ and $\{v \in \mathbb{R}^n \colon v_i \in \{0,1\}\}$.

EXERCISE 2.7. Let Ω be an infinite set, i.e. $\#\Omega = +\infty$. Consider the collection of all finite subsets of Ω:

$$\mathcal{C} = \{A \in \mathcal{P}(\Omega) \colon \#A < +\infty\}.$$

Is $\mathcal{C} \cup \{\Omega\}$ an algebra? Is it a σ-algebra?

EXERCISE 2.8. Let $\Omega = [-1,1] \subset \mathbb{R}$. Determine if the following collection of sets is a σ-algebra:

$$\mathcal{F} = \{A \in \mathcal{B}(\Omega) \colon x \in A \Rightarrow -x \in A\}.$$

EXERCISE 2.9. Let $(\Omega, \mathcal{F})$ be a measurable space. Consider two disjoint sets $A, B \subset \Omega$ and assume that $A \in \mathcal{F}$. Show that $A \cup B \in \mathcal{F}$ is equivalent to $B \in \mathcal{F}$?

1.1. Generation of σ-algebras. In many situations one requires some sets to be measurable due to their relevance to the problem we are studying. If the collection of those sets is not already a σ-algebra, we need to take a larger one that is. That will be called the σ-algebra generated by the original collection, which we define below.

Take I to be any set (of indices).

THEOREM 2.10. *If $\mathcal{F}_\alpha$ is a σ-algebra, $\alpha \in I$, then $\mathcal{F} = \bigcap_{\alpha \in I} \mathcal{F}_\alpha$ is also a σ-algebra.*

PROOF.

(1) As for any α we have $\emptyset \in \mathcal{F}_\alpha$, then $\emptyset \in \mathcal{F}$.
(2) Let $A \in \mathcal{F}$. So, $A \in \mathcal{F}_\alpha$ for any α. Thus, $A^c \in \mathcal{F}_\alpha$ and $A^c \in \mathcal{F}$.
(3) If $A_n \in \mathcal{F}$, we have $A_n \in \mathcal{F}_\alpha$ for any α. So, $\bigcup_n A_n \in \mathcal{F}_\alpha$ and $\bigcup_n A_n \in \mathcal{F}$.

□

EXERCISE 2.11. Is the union of σ-algebras also a σ-algebra?

Consider now the collection of all σ-algebras:

$$\Sigma = \{\text{all } \sigma\text{-algebras of } \Omega\}.$$

So, e.g. $\mathcal{P} \in \Sigma$ and $\{\emptyset, \Omega\} \in \Sigma$. In addition, let $\mathcal{I} \subset \mathcal{P}$ be a collection of subsets of Ω, i.e. $\mathcal{I} \subset \mathcal{P}$, not necessarily a σ-algebra. Define the subset of Σ given by the σ-algebras that contain $\mathcal{I}$:

$$\Sigma_\mathcal{I} = \{\mathcal{F} \in \Sigma \colon \mathcal{I} \subset \mathcal{F}\}.$$

The *σ-algebra generated by $\mathcal{I}$* is the intersection of all σ-algebras containing $\mathcal{I}$,

$$\sigma(\mathcal{I}) = \bigcap_{\mathcal{F} \in \Sigma_\mathcal{I}} \mathcal{F}.$$

Hence, $\sigma(\mathcal{I})$ is the smallest σ-algebra containing $\mathcal{I}$ (i.e. it is a subset of any σ-algebra containing $\mathcal{I}$).

EXAMPLE 2.12.

(1) Let $A \subset \Omega$ and $\mathcal{I} = \{A\}$. Any σ-algebra containing $\mathcal{I}$ has to include the sets $\emptyset$, Ω, A and A^c. Since these sets form already a σ-algebra, we have
$$\sigma(\mathcal{I}) = \{\emptyset, \Omega, A, A^c\}.$$
(2) Consider two disjoint sets $A, B \subset \Omega$ and $\mathcal{I} = \{A, B\}$. The generated σ-algebra is
$$\sigma(\mathcal{I}) = \{\emptyset, \Omega, A, B, A^c, B^c, A \cup B, (A \cup B)^c\}.$$

(3) Consider now two different sets $A, B \subset \Omega$ such that $A \cap B \neq \emptyset$, and $\mathcal{I} = \{A, B\}$. Then,

$$\begin{aligned}\sigma(\mathcal{I}) = &\{\emptyset, \Omega, A, B, A^c, B^c,\\ &A \cup B, A \cup B^c, A^c \cup B, (A \cup B)^c, (A \cup B^c)^c, (A^c \cup B)^c,\\ &B^c \cup (A \cup B^c)^c, (B^c \cup (A \cup B^c)^c)^c,\\ &((A \cup B^c)^c) \cup ((A^c \cup B)^c), (((A \cup B^c)^c) \cup ((A^c \cup B)^c))^c\}\\ = &\{\emptyset, \Omega, A, B, A^c, B^c,\\ &A \cup B, A \cup B^c, A^c \cup B, A^c \cap B^c, B \setminus A, A \setminus B,\\ &(A \cap B)^c, A \cap B,\\ &(A \cup B) \setminus (A \cap B), (A^c \cap B^c) \cup (A \cap B)\}\,.\end{aligned}$$

Exercise 2.13. Show that

(1) If $\mathcal{I}_1 \subset \mathcal{I}_2 \subset \mathcal{P}$, then $\sigma(\mathcal{I}_1) \subset \sigma(\mathcal{I}_2)$.
(2) $\sigma(\sigma(\mathcal{I})) = \sigma(\mathcal{I})$ for any $\mathcal{I} \subset \mathcal{P}$.

Exercise 2.14. Consider a finite set $\Omega = \{\omega_1, \dots, \omega_n\}$. Prove that $\mathcal{I} = \{\{\omega_1\}, \dots, \{\omega_n\}\}$ generates $\mathcal{P}(\Omega)$.

Exercise 2.15. Determine $\sigma(\mathcal{C})$, where

$$\mathcal{C} = \{\{x\} \colon x \in \Omega\}\,.$$

What is the smallest algebra that contains $\mathcal{C}$.

1.2. Borel sets. A specially important collection of subsets of $\mathbb{R}$ in applications is

$$\mathcal{I} = \{]-\infty, x] \subset \mathbb{R} \colon x \in \mathbb{R}\}.$$

It is not an algebra since it does not contain even the emptyset. Another collection could be obtained by considering complements and intersections of pairs of sets in $\mathcal{I}$. That is,

$$\mathcal{I}' = \{]a, b] \subset \mathbb{R} \colon -\infty \leq a \leq b \leq +\infty\}.$$

Here we are using the following conventions

$$]a, +\infty] =]a, +\infty[\quad \text{and} \quad]a, a] = \emptyset$$

so that $\emptyset$ and $\mathbb{R}$ are also in the collection. The complement of $]a, b] \in \mathcal{I}'$ is still not in $\mathcal{I}'$, but is the union of two sets there:

$$]a, b]^c =]-\infty, a] \cup]b, +\infty].$$

So, the smallest algebra that contains $\mathcal{I}$ corresponds to the collection of finite unions of sets in $\mathcal{I}'$,

$$\mathcal{A}(\mathbb{R}) = \left\{ \bigcup_{n=1}^{N} I_n \subset \mathbb{R} \colon I_1, \dots, I_N \in \mathcal{I}', N \in \mathbb{N} \right\},$$

called the Borel algebra of $\mathbb{R}$. Clearly, $\mathcal{I} \subset \mathcal{I}' \subset \mathcal{A}(\mathbb{R})$.

We define the **Borel σ-algebra** as

$$\mathcal{B}(\mathbb{R}) = \sigma(\mathcal{I}) = \sigma(\mathcal{I}') = \sigma(\mathcal{A}(\mathbb{R})).$$

The elements of $\mathcal{B}(\mathbb{R})$ are called the **Borel sets**. We will often simplify the notation by writing $\mathcal{B}$.

When Ω is a subset of $\mathbb{R}$ we can also define the Borel algebra and the σ-algebra on Ω. It is enough to take

$$\mathcal{A}(\Omega) = \{A \cap \Omega \colon A \in \mathcal{A}(\mathbb{R})\} \quad \text{and} \quad \mathcal{B}(\Omega) = \{A \cap \Omega \colon A \in \mathcal{B}(\mathbb{R})\}.$$

EXERCISE 2.16. Check that $\mathcal{A}(\Omega)$ and $\mathcal{B}(\Omega)$ are an algebra and a σ-algebra of Ω, respectively.

EXERCISE 2.17. Show that:

(1) $\mathcal{B}(\mathbb{R}) \neq \mathcal{A}(\mathbb{R})$.
(2) Any singular set $\{a\}$ with $a \in \mathbb{R}$, is a Borel set.
(3) Any countable set is a Borel set.
(4) Any open set is a Borel set. *Hint*: Any open set can be written as a countable union of pairwise disjoint open intervals.

2. Monotone classes

We write $A_n \uparrow A$ to represent a sequence of sets $A_1, A_2, \dots$ that is increasing, i.e.

$$A_1 \subset A_2 \subset \dots,$$

and converges to the set

$$A = \bigcup_{n=1}^{+\infty} A_n.$$

Similarly, $A_n \downarrow A$ corresponds to a sequence of sets $A_1, A_2, \dots$ that is decreasing, i.e.

$$\dots \subset A_2 \subset A_1,$$

and converging to

$$A = \bigcap_{n=1}^{+\infty} A_n.$$

Notice that in both cases, if the sets A_n are measurable, then A is also measurable.

A collection $\mathcal{A} \subset \mathcal{P}$ is a *monotone class* iff

(1) if $A_1, A_2, \dots \in \mathcal{A}$ such that $A_n \uparrow A$, then $A \in \mathcal{A}$,
(2) if $A_1, A_2, \dots \in \mathcal{A}$ such that $A_n \downarrow A$, then $A \in \mathcal{A}$.

THEOREM 2.18. *Suppose that $\mathcal{A}$ is an algebra. Then, $\mathcal{A}$ is a σ-algebra iff it is a monotone class.*

PROOF.

(⇒) If $A_1, A_2, \dots \in \mathcal{A}$ such that $A_n \uparrow A$ or $A_n \downarrow A$, then $A \in \mathcal{A}$ by the properties of a σ-algebra.

(⇐) Let $A_1, A_2, \dots \in \mathcal{A}$. Take

$$B_n = \bigcup_{i=1}^{n} A_i, \quad n \in \mathbb{N}.$$

Hence, $B_n \in \mathcal{A}$ for all n since $\mathcal{A}$ is an algebra. Moreover, $B_n \subset B_{n+1}$ and $B_n \uparrow \cup_n A_n \in \mathcal{A}$ because $\mathcal{A}$ is a monotone class.

□

THEOREM 2.19. *If $\mathcal{A}$ is an algebra, then the smallest monotone class containing $\mathcal{A}$ is $\sigma(\mathcal{A})$.*

EXERCISE 2.20. Prove it.

3. Product algebras

Let $(\Omega_1, \mathcal{F}_1)$ and $(\Omega_2, \mathcal{F}_2)$ be two measurable spaces. We want to find a natural algebra and σ-algebra of the product space

$$\Omega = \Omega_1 \times \Omega_2.$$

A particular type of subsets of Ω, called *measurable rectangles*, is given by the product of a set $A \in \mathcal{F}_1$ by another $B \in \mathcal{F}_2$, i.e.

$$\begin{aligned} A \times B &= \{(x_1, x_2) \in \Omega \colon x_1 \in A, x_2 \in B\} \\ &= \{x_1 \in A\} \cap \{x_2 \in B\} \\ &= (A \times \Omega_2) \cap (\Omega_1 \times B), \end{aligned}$$

where we have simplified notation in the obvious way. Consider the following collection of finite unions of measurable rectangles

$$\mathcal{A} = \left\{ \bigcup_{i=1}^{N} \bigcup_{j=1}^{M} A_i \times B_j \subset \Omega \colon A_i \in \mathcal{F}_1, B_j \in \mathcal{F}_2, N, M \in \mathbb{N} \right\}. \tag{2.1}$$

We denote it by $\mathcal{A} = \mathcal{F}_1 \times \mathcal{F}_2$.

PROPOSITION 2.21. *$\mathcal{A}$ is an algebra (called the product algebra).*

PROOF. Notice that $\emptyset \times \emptyset$ is the empty set of Ω and is in $\mathcal{A}$.

The complement of $A \times B$ in Ω is

$$\begin{aligned} (A \times B)^c &= \{x_1 \notin A \text{ or } x_2 \notin B\} \\ &= \{x_1 \notin A\} \cup \{x_2 \notin B\} \\ &= (A^c \times \Omega_2) \cup (\Omega_1 \times B^c) \end{aligned}$$

which is in $\mathcal{A}$. Moreover, the intersection between two measurable rectangles is given by

$$\begin{aligned}(A_1 \times B_1) \cap (A_2 \times B_2) &= \{x_1 \in A_1, x_2 \in B_1, x_1 \in A_2, x_2 \in B_2\} \\ &= \{x_1 \in A_1 \cap A_2, x_2 \in B_1 \cap B_2\} \\ &= (A_1 \cap A_2) \times (B_1 \cap B_2),\end{aligned}$$

again in $\mathcal{A}$. So, the complement of a finite union of measurable rectangles is the intersection of the complements, which is thus in $\mathcal{A}$. □

EXERCISE 2.22. Show that any element in $\mathcal{A}$ can be written as a finite union of disjoint measurable rectangles.

The *product σ-algebra* is defined as

$$\mathcal{F} = \sigma(\mathcal{A}).$$

4. Measures

Consider an algebra $\mathcal{A}$ of a set Ω and a function

$$\mu\colon \mathcal{A} \to \bar{\mathbb{R}}$$

that for each set in $\mathcal{A}$ attributes a real number or $\pm\infty$, i.e. in

$$\bar{\mathbb{R}} = \mathbb{R} \cup \{-\infty, +\infty\}.$$

We say that μ is *σ-additive* if for any sequence of pairwise disjoint sets $A_1, A_2, \dots \in \mathcal{A}$ such that $\bigcup_{n=1}^{+\infty} A_n \in \mathcal{A}$ [1] we have

$$\mu\left(\bigcup_{n=1}^{+\infty} A_n\right) = \sum_{n=1}^{+\infty} \mu(A_n).$$

In case it is only possible to prove the inequality $\leq$ instead of the equality, μ is said to be σ-subadditive.

The function μ is called a *measure* on $\mathcal{A}$ iff

(1) $\mu(A) \geq 0$ or $\mu(A) = +\infty$ for any $A \in \mathcal{A}$,
(2) $\mu(\emptyset) = 0$,
(3) μ is σ-additive[2].

REMARK 2.23. We use the arithmetic in $\bar{\mathbb{R}}$ by setting

$$(+\infty) + (+\infty) = +\infty \quad \text{and} \quad a + \infty = +\infty$$

for any $a \in \mathbb{R}$. Moreover, we write $a < +\infty$ to mean that a is a finite number.

We say that $P\colon \mathcal{A} \to \mathbb{R}$ is a *probability measure* iff

[1]Notice that this condition is always satisfied if $\mathcal{A}$ is a σ-algebra.
[2]μ is called an outer measure in case it is σ-subadditive.

(1) P is a measure,
(2) $P(\Omega) = 1$.

REMARK 2.24. A (non-trivial) finite measure, i.e. satisfying $0 < \mu(\Omega) < +\infty$, can be made into a probability measure P by a normalization:

$$P(A) = \frac{\mu(A)}{\mu(\Omega)}, \quad A \in \mathcal{A}.$$

Given a measure μ on an algebra $\mathcal{A}$, a set $A \in \mathcal{A}$ is said to have *full measure* if $\mu(A^c) = 0$. In the case of probability measure we also say that this set (event) has *full probability*.

EXERCISE 2.25. (Counting measure) Show that the function that counts the number of elements of a set $A \in \mathcal{P}(\Omega)$:

$$\mu(A) = \begin{cases} \#A, & \#A < +\infty \\ +\infty, & \text{c.c,} \end{cases}$$

is a measure. Find the sets with full measure μ.

EXERCISE 2.26. Let $\mu \colon \mathcal{P}(\Omega) \to \mathbb{R}$ that satisfies

$$\mu(\emptyset) = 0, \quad \mu(\Omega) = 2, \quad \mu(A) = 1, \quad A \in \mathcal{P}(\Omega) \setminus \{\emptyset, \Omega\}.$$

Determine if μ is σ-additive.

EXERCISE 2.27. Let μ_n be a measure and $a_n \geq 0$ for all $n \in \mathbb{N}$. Prove that

$$\mu = \sum_{n=1}^{+\infty} a_n \mu_n,$$

is also a measure. Furthermore, show that if μ_n is a probability measure for all n and $\sum_n a_n = 1$, then μ is also a probability measure.

4.1. Properties. If $\mathcal{F}$ is a σ-algebra of Ω, μ a measure on $\mathcal{F}$ and P a probability measure on $\mathcal{F}$, we say that $(\Omega, \mathcal{F}, \mu)$ is a *measure space* and $(\Omega, \mathcal{F}, P)$ is a *probability space*.

PROPOSITION 2.28. *Consider a measure space* $(\Omega, \mathcal{F}, \mu)$ *and* $A, B \in \mathcal{F}$. *Then,*

(1) $\mu(A \cup B) + \mu(A \cap B) = \mu(A) + \mu(B)$.
(2) *If* $A \subset B$, *then* $\mu(A) \leq \mu(B)$.
(3) *If* $A \subset B$ *and* $\mu(A) < +\infty$, *then* $\mu(B \setminus A) = \mu(B) - \mu(A)$.

PROOF. Notice that

$$A \cup B = (A \setminus B) \cup (A \cap B) \cup (B \setminus A)$$

is the union of disjoint sets. Moreover, $A = (A \setminus B) \cup (A \cap B)$ and $B = (B \setminus A) \cup (A \cap B)$.

(1) We have then $\mu(A \cup B) + \mu(A \cap B) = \mu(A \setminus B) + \mu(A \cap B) + \mu(B \setminus A) + \mu(A \cap B) = \mu(A) + \mu(B)$.
(2) If $A \subset B$, then $B = A \cup (B \setminus A)$ and $\mu(B) = \mu(A) + \mu(B \setminus A) \geq \mu(A)$.
(3) If $\mu(A) < +\infty$, then $\mu(B \setminus A) = \mu(B) - \mu(A)$. Observe that if $\mu(A) = +\infty$, then $\mu(B) = +\infty$. Hence, it would not be possible to determine $\mu(B \setminus A)$.

□

Exercise 2.29. Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. Show that for any sequence of measurable sets $A_1, A_2, \dots \in \mathcal{F}$ we have

$$\mu\left(\bigcup_{n\geq 1} A_n\right) \leq \sum_{n\geq 1} \mu(A_n).$$

A proposition is said to be valid μ-*almost everywhere* (μ-a.e.), if it holds on a set of full measure μ.

Exercise 2.30. Consider two sets each one having full measure. Show that their intersection also has full measure.

Proposition 2.31. *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and $A_1, A_2, \dots \in \mathcal{F}$.*

(1) *If $A_n \uparrow A$, then*
$$\mu(A) = \lim_{n\to+\infty} \mu(A_n).$$
(2) *If $A_n \downarrow A$ and $\mu(A_1) < +\infty$, then*
$$\mu(A) = \lim_{n\to+\infty} \mu(A_n).$$

Proof.

(1) If there is i such that $\mu(A_i) = +\infty$, then $\mu(A_n) = +\infty$ for $n \geq i$. So, $\lim_n \mu(A_n) = +\infty$. On the other hand, as $A_i \subset \bigcup_n A_n$, we have $\mu(\bigcup_n A_n) = +\infty$. It remains to consider the case where $\mu(A_n) < +\infty$ for any n. Let $A_0 = \emptyset$ e $B_n = A_n \setminus A_{n-1}$, $n \geq 1$, a sequence of pairwise disjoint sets. Then, $\bigcup_n A_n = \bigcup_n B_n$ and $\mu(B_n) = \mu(A_n) - \mu(A_{n-1})$. Finally,
$$\begin{aligned}\mu\left(\bigcup_n A_n\right) &= \mu\left(\bigcup_n B_n\right) \\ &= \lim_{n\to+\infty} \sum_{i=1}^{n} (\mu(A_n) - \mu(A_{n-1})) \\ &= \lim_{n\to+\infty} \mu(A_n).\end{aligned} \tag{2.2}$$

(2) Since $\mu(A_1) < +\infty$ any subset of A_1 also has finite measure. Notice that

$$\bigcap_n A_n = \left(\bigcup_n A_n^c\right)^c = A_1 \setminus \bigcup_n C_n,$$

where $C_n = A_n^c \cap A_1$. We also have $C_k \subset C_{k+1}$. Hence, by the previous case,

$$\begin{aligned} \mu\left(\bigcap_n A_n\right) &= \mu(A_1) - \mu\left(\bigcup_n C_n\right) \\ &= \lim_{n\to+\infty} (\mu(A_1) - \mu(C_n)) \\ &= \lim_{n\to+\infty} \mu(A_n). \end{aligned} \tag{2.3}$$

□

EXAMPLE 2.32. Consider the counting measure μ. Let

$$A_n = \{n, n+1, \dots\}.$$

Therefore, $A = \bigcap_{n=1}^{+\infty} A_n = \emptyset$ and $A_{n+1} \subset A_n$. However, $\mu(A_n) = +\infty$ does not converge to $\mu(A) = 0$. Notice that the previous proposition does not apply because $\mu(A_1) = +\infty$.

The next theorem gives us a way to construct probability measures on an algebra.

THEOREM 2.33. *Let $\mathcal{A}$ be an algebra of a set Ω. Then, $P\colon \mathcal{A} \to \mathbb{R}$ is a probability measure on $\mathcal{A}$ iff*

(1) $P(\Omega) = 1$,
(2) $P(A \cup B) = P(A) + P(B)$ *for every disjoint pair* $A, B \in \mathcal{A}$,
(3) $\lim_{n\to+\infty} P(A_n) = 0$ *for all* $A_1, A_2 \dots \in \mathcal{A}$ *such that* $A_n \downarrow \emptyset$.

EXERCISE 2.34. *Prove it.

EXERCISE 2.35. Let $(\Omega, \mathcal{F}, P)$ probability space, $A_1, A_2, \dots \in \mathcal{F}$ and B is the set of points in Ω that belong to an infinite number of A_n's:

$$B = \bigcap_{n=1}^{+\infty} \bigcup_{k=n}^{+\infty} A_k.$$

Show that:

(1) (First Borel-Cantelli lemma) If

$$\sum_{n=1}^{+\infty} P(A_n) < +\infty,$$

then $P(B) = 0$.

(2) *(Second Borel-Cantelli lemma) If

$$\sum_{n=1}^{+\infty} P(A_n) = +\infty$$

and

$$P\left(\bigcap_{i=1}^{n} A_i\right) = \prod_{i=1}^{n} P(A_i),$$

for every $n \in \mathbb{N}$ (i.e. the events are mutually independent), then $P(B) = 1$.

4.2. *Carathéodory extension theorem. In the definition of σ-additivity is not very convenient to check whether we are choosing only sets $A_1, A_2, \dots$ in the algebra $\mathcal{A}$ such that their union is still in $\mathcal{A}$. That would be guaranteed by considering a σ-algebra instead of an algebra.

Theorem 2.36 below assures the extension of the measure to a σ-algebra containing $\mathcal{A}$. So, we only need to construct a measure on an algebra in order to have it well determined on a larger σ-algebra. Before stating the theorem we need several definitions.

Let μ be a measure on an algebra $\mathcal{A}$ of Ω. We say that a sequence of disjoint sets $A_1, A_2, \dots \in \mathcal{A}$ is a cover of $A \in \mathcal{P}$ if

$$A \subset \bigcup_j A_j.$$

Consider the function $\mu^* \colon \mathcal{P} \to \bar{\mathbb{R}}$ given by

$$\mu^*(A) = \inf_{A_1, A_2, \dots \text{ cover } A} \sum_j \mu(A_j), \quad A \in \mathcal{P},$$

where the infimum is taken over all covers $A_1, A_2, \dots \in \mathcal{A}$ of A. Notice that $\mu^*(\emptyset) = 0$ as the empty set covers itself. Also, $\mu^*(A) \geq 0$. To show that μ^* is a measure it is enough to determine its σ-additivity.

Consider now the collection of subsets of Ω defined as

$$\mathcal{M} = \{A \in \mathcal{P} \colon \mu^*(B) = \mu^*(B \cap A) + \mu^*(B \cap A^c), B \in \mathcal{P}\}.$$

THEOREM 2.36 (Carathéodory extension). *μ^* is a measure on the σ-algebra $\mathcal{M}$ such that $\mathcal{A} \subset \sigma(\mathcal{A}) \subset \mathcal{M}$ and*

$$\mu^*(A) = \mu(A), \quad A \in \mathcal{A}$$

(i.e. μ^* extends μ to $\mathcal{M}$). Moreover, if μ is finite, μ^* is the unique extension to $\sigma(\mathcal{A})$ and it is also finite.

The remaining part of this section is devoted to the proof of the above theorem.

LEMMA 2.37. *μ^* is σ-subadditive on $\mathcal{P}$.*

Proof. Take $A_1, A_2, \dots \in \mathcal{P}$ pairwise disjoint and $\varepsilon > 0$. For each A_n, $n \in \mathbb{N}$, consider the cover $A_{n,1}, A_{n,2}, \dots \in \mathcal{A}$ such that

$$\sum_j \mu(A_{n,j}) < \mu^*(A_n) + \frac{\varepsilon}{2^n}.$$

Then, because μ is a measure,

$$\begin{aligned}\mu^*\left(\bigcup_n A_n\right) &\leq \sum_j \mu\left(\bigcup_n A_{n,j}\right)\\ &\leq \sum_{n,j} \mu\left(A_{n,j}\right)\\ &< \sum_n \mu^*\left(A_n\right) + \varepsilon.\end{aligned}$$

Since $\varepsilon > 0$ is arbitrary, μ^* is σ-subadditive. $\square$

From the σ-subadditivity we know that for any $A, B \in \mathcal{P}$ we have

$$\mu^*(B) \leq \mu^*(B \cap A) + \mu^*(B \cap A^c).$$

So, an element A of $\mathcal{M}$ has to verify the other inequality:

$$\mu^*(B) \geq \mu^*(B \cap A) + \mu^*(B \cap A^c)$$

for every $B \in \mathcal{P}$.

Lemma 2.38. *μ^* is finitely additive on $\mathcal{M}$.*

Proof. Let $A_1, A_2 \in \mathcal{M}$ disjoint. Then,

$$\mu^*(A_1 \cup A_2) = \mu^*((A_1 \cup A_2) \cap A_1) + \mu^*((A_1 \cup A_2) \cap A_1^c) = \mu^*(A_1) + \mu^*(A_2).$$

By induction we obtain the finite additivity on $\mathcal{M}$. $\square$

Notice that μ^* is monotonous on $\mathcal{P}$, i.e. $\mu^*(C) \leq \mu^*(D)$ whenever $C \subset D$. This is because a cover of D is also a cover of C.

Lemma 2.39. *$\mathcal{M}$ is a σ-algebra.*

Proof. Let $B \in \mathcal{P}$. From $\mu^*(B \cap \emptyset) + \mu^*(B \cap \Omega) = \mu^*(\emptyset) + \mu^*(B) = \mu^*(B)$ we obtain that $\emptyset \in \mathcal{M}$. If $A \in \mathcal{M}$ it is clear that A^c is also in $\mathcal{M}$.

Now, let $A_1, A_2 \in \mathcal{M}$. Their union is also in $\mathcal{M}$ because

$$\begin{aligned}\mu^*(B \cap (A_1 \cup A_2)) &= \mu^*((B \cap A_1) \cup (B \cap A_2 \cap A_1^c))\\ &\leq \mu^*(B \cap A_1) + \mu^*(B \cap A_2 \cap A_1^c)\end{aligned}$$

and

$$\mu^*(B \cap (A_1 \cup A_2)^c) = \mu^*(B \cap A_1^c \cap A_2^c),$$

whose sum gives

$$\begin{aligned}\mu^*(B\cap(A_1\cup A_2))+\mu^*(B\cap(A_1\cup A_2)^c)&\leq \mu^*(B\cap A_1)+\mu^*(B\cap A_1^c)\\ &=\mu^*(B),\end{aligned}$$

where we have used the fact that A_1, A_2 are in $\mathcal{M}$.

By induction any finite union of sets in $\mathcal{M}$ is also in $\mathcal{M}$. It remains to deal with the countable union case. Suppose that $A_1, A_2, \dots \in \mathcal{M}$ and write

$$F_n = A_n \setminus \left(\bigcup_{k=1}^{n-1} A_k\right), \quad n\in\mathbb{N},$$

which are pairwise disjoint and satisfy

$$\bigcup_n F_n = \bigcup_n A_n.$$

Moreover, since each F_n is a finite union of sets in $\mathcal{M}$, then $F\in\mathcal{M}$.

Finally,

$$\begin{aligned}\mu^*(B)&=\mu^*\left(B\cap\bigcup_{n=1}^N F_n\right)+\mu^*\left(B\cap\left(\bigcup_{n=1}^N F_n\right)^c\right)\\ &\geq \sum_{n=1}^N \mu^*\left(B\cap F_n\right)+\mu^*\left(B\cap\left(\bigcup_{n=1}^{+\infty} F_n\right)^c\right)\end{aligned}$$

where we have used the finite additivity of μ^* on $\mathcal{M}$ along with the facts that

$$\left(\bigcup_{n=1}^{+\infty} F_n\right)^c \subset \left(\bigcup_{n=1}^{N} F_n\right)^c$$

and μ^* is monotonous. Taking the limit $N\to+\infty$ we obtain

$$\begin{aligned}\mu^*(B)&\geq \sum_{n=1}^{+\infty} \mu^*\left(B\cap F_n\right)+\mu^*\left(B\cap\left(\bigcup_{n=1}^{+\infty} F_n\right)^c\right)\\ &\geq \mu^*\left(B\cap\bigcup_{n=1}^{+\infty} F_n\right)+\mu^*\left(B\cap\left(\bigcup_{n=1}^{+\infty} F_n\right)^c\right),\end{aligned}$$

proving that $\mathcal{M}$ is a σ-algebra. □

Lemma 2.40. *$\mathcal{A}\subset\sigma(\mathcal{A})\subset\mathcal{M}$.*

Proof. Let $A\in\mathcal{A}$. Given any $\varepsilon>0$ and $B\in\mathcal{P}$ consider $A_1, A_2, \dots \in\mathcal{A}$ covering B such that

$$\mu^*(B)\leq\sum_n \mu(A_n) < \mu^*(B)+\varepsilon.$$

On the other hand, the sets $A_1 \cap A, A_2 \cap A, \dots \in \mathcal{A}$ cover $B \cap A$ and $A_1 \cap A^c, A_2 \cap A^c, \dots \in \mathcal{A}$ cover $B \cap A^c$. From $\sum_n \mu(A_n) = \sum_n \mu(A_n \cap A) + \mu(A_n \cap A^c) \geq \mu^*(B \cap A) + \mu^*(B \cap A^c)$ we obtain

$$\mu^*(B \cap A) + \mu^*(B \cap A^c) < \mu^*(B) + \varepsilon.$$

Since ε is arbitrary and from the σ-subadditivity of μ^* we have

$$\mu^*(B) = \mu^*(B \cap A) + \mu^*(B \cap A^c).$$

So, $A \in \mathcal{M}$ and $\mathcal{A} \subset \mathcal{M}$.

Finally, as $\mathcal{M}$ is a σ-algebra that contains $\mathcal{A}$ it also contains $\sigma(\mathcal{A})$. $\square$

LEMMA 2.41. *μ^* is σ-additive on $\mathcal{M}$.*

PROOF. Let $A_1, A_2, \dots \in \mathcal{M}$ pairwise disjoint. By the finite additivity and the monotonicity of μ^* on $\mathcal{M}$ we have

$$\sum_{n=1}^{N} \mu^*(A_n) = \mu^* \left(\bigcup_{n=1}^{N} A_n \right) \leq \mu^* \left(\bigcup_{n=1}^{+\infty} A_n \right).$$

Taking the limit $N \to +\infty$ we obtain

$$\sum_{n=1}^{+\infty} \mu^*(A_n) \leq \mu^* \left(\bigcup_{n=1}^{+\infty} A_n \right).$$

As μ^* is σ-subadditive on $\mathcal{M}$, it is also σ-additive. $\square$

It simple to check that μ^* agrees with μ for any $A \in \mathcal{A}$. In fact, we can choose the cover equal to A itself, so that

$$\mu^*(A) = \mu(A).$$

Suppose now that μ is finite. Since the cover of a set A consists of disjoint sets $A_1, A_2, \dots \in \mathcal{A}$, then $\sum_j \mu(A_j) = \mu(\bigcup_j A_j) \leq \mu(\Omega)$. Thus μ^* is also finite.

The uniqueness of the extension comes from the following lemma.

LEMMA 2.42. *Let μ be finite. If μ_1^* and μ_2^* are extensions of μ to $\mathcal{M}$, then $\mu_1^* = \mu_2^*$ on $\sigma(\mathcal{A})$.*

PROOF. The collection to where one has a unique extension is

$$\mathcal{F} = \{A \in \mathcal{M} \colon \mu_1^*(A) = \mu_2^*(A)\}.$$

Taking an increasing sequence $A_1 \subset A_2 \subset \dots$ in $\mathcal{F}$ we have

$$\begin{aligned}\mu_1^*\left(\bigcup_{n=1}^{+\infty} A_n\right) &= \mu_1^*\left(\bigcup_n A_n \setminus A_{n-1}\right)\\ &= \sum_n \mu_1^*(A_n \setminus A_{n-1})\\ &= \sum_n \mu_2^*(A_n \setminus A_{n-1})\\ &= \mu_2^*\left(\bigcup_n A_n\right),\end{aligned}$$

where $A_0 = \emptyset$. Thus, $A_n \uparrow \bigcup_n A_n \in \mathcal{F}$. Similarly, for a decreasing sequence we also obtain $A_n \downarrow \bigcap_n A_n \in \mathcal{F}$. Thus, $\mathcal{F}$ is a monotone class. According to Theorem 2.19, since $\mathcal{F}$ contains the algebra $\mathcal{A}$ it also contains $\sigma(\mathcal{A})$. $\square$

5. Examples

5.1. Dirac measure. Let $a \in \Omega$ and $P\colon \mathcal{P}(\Omega) \to \mathbb{R}$ given by

$$P(A) = \begin{cases} 1, & a \in A \\ 0, & \text{other cases.} \end{cases}$$

If $A_1, A_2, \dots$ are pairwise disjoint sets, then only one of the following alternatives can hold:

(1) There exists a unique j such that $a \in A_j$. So, $P(\bigcup_n A_n) = 1 = P(A_j) = \sum_n P(A_n)$.
(2) For all n we have that $a \notin A_n$. Therefore, $P(\bigcup_n A_n) = 0 = \sum_n P(A_n)$.

This implies that P is σ-additive. Since $P(\Omega) = 1$, P is a probability measure called *Dirac measure at a* .

Exercise 2.43. Is

$$P = \sum_{n=1}^{+\infty} \frac{1}{2^n}\delta_{1/n}.$$

a probability measure on $\mathcal{P}(\mathbb{R})$?

5.2. Lebesgue measure. Consider the Borel algebra $\mathcal{A} = \mathcal{A}(\mathbb{R})$ on $\mathbb{R}$ and a function $m\colon \mathcal{A} \to \bar{\mathbb{R}}$. For a sequence of disjoint intervals $]a_n, b_n]$ with $a_n \leq b_n$, $n \in \mathbb{N}$, such that their union is in $\mathcal{A}$, we define

$$m\left(\bigcup_{n=1}^{+\infty}]a_n, b_n]\right) = \sum_{n=1}^{+\infty}(b_n - a_n)$$

corresponding to the sum of the lengths of the intervals. This is the σ-additivity. Moreover, $m(\emptyset) = m(]a,a]) = 0$ and $m(A) \geq 0$ for any $A \in \mathcal{A}$. Therefore, m is a measure on the algebra $\mathcal{A}$. By the Carathéodory extension theorem (Theorem 2.36) there is a unique extension of m to the Borel σ-algebra $\mathcal{B}$ also denoted by $m \colon \mathcal{B} \to \mathbb{R}$. This is called the *Lebesgue measure* on $\mathcal{B}$.

REMARK 2.44. There is a larger σ-algebra to where we can extend m. It is called the Lebesgue σ-algebra $\mathcal{M}$ and includes all sets $\Lambda \subset \mathbb{R}$ such that there are Borelian sets $A, B \in \mathcal{B}$ satisfying $A \subset \Lambda \subset B$ and $m(B \setminus A) = 0$. Clearly, $\mathcal{B} \subset \mathcal{M}$. We set $m(\Lambda) = m(A) = m(B)$. The extended function $m \colon \mathcal{M} \to \mathbb{R}$ is called the Lebesgue measure on $\mathcal{M}$.

EXERCISE 2.45. Compute the Lebesgue measure of the following Borel sets: $\{a\}$, $[a,b]$, $]-\infty, a[$, $[b,+\infty[$.

5.3. Product measure. Let $(\Omega_1, \mathcal{F}_1, \mu_1)$ and $(\Omega_2, \mathcal{F}_2, \mu_2)$ be measure spaces. Consider the product space $\Omega = \Omega_1 \times \Omega_2$ with the product σ-algebra $\mathcal{F} = \sigma(\mathcal{A})$, where $\mathcal{A}$ is the product algebra introduced in (2.1).

We start by defining the *product measure* $\mu = \mu_1 \times \mu_2$ as the measure on Ω that satisfies

$$\mu(A_1 \times A_2) = \mu_1(A_1)\, \mu_2(A_2)$$

for any measurable rectangle $A_1 \times A_2 \in \mathcal{F}_1 \times \mathcal{F}_2$. For other sets in $\mathcal{A}$, i.e. finite union of measurable rectangles, we define μ as to make it σ-additive.

EXERCISE 2.46.

(1) Prove that if $A \in \mathcal{A}$ can be written as the finite disjoint union of measurable rectangles in two different ways, i.e. we can find measurable rectangles $A_i \times B_j$, $i = 1, \dots, N$, $j = 1, \dots, M$, and also $A'_i \times B'_j$, $i = 1, \dots, N'$, $j = 1, \dots, M'$ such that

$$A = \bigcup_i \bigcup_j A_i \times B_j = \bigcup_i \bigcup_j A'_i \times B'_j,$$

then

$$\sum_{i,j} \mu(A_i \times B_j) = \sum_{i,j} \mu(A'_i \times B'_j).$$

So, $\mu(A)$ is well-defined.

(2) Show that μ can be uniquely extended to every measurable set in the product σ-algebra $\mathcal{F}$.

5.4. Non-unique extensions. The Carathéodory extension theorem guarantees an extension to $\sigma(\mathcal{A})$ of a measure initially defined on an algebra $\mathcal{A}$. If the measure is not finite ($\mu(\Omega) = +\infty$), then there might be more than one extension, say μ_1^* and μ_2^*. They both agree with μ for sets in $\mathcal{A}$ but are different in some sets in $\sigma(\mathcal{A}) \setminus \mathcal{A}$. Here are two examples.

Consider the algebra $\mathcal{A} = \mathcal{A}(\mathbb{R})$ and the measure $\mu\colon \mathcal{A} \to \bar{\mathbb{R}}$ given by

$$\mu(A) = \begin{cases} 0, & A = \emptyset \\ +\infty, & \text{o.c.} \end{cases}$$

Two extensions of μ to $\mathcal{B} = \sigma(\mathcal{A})$ are

$$\mu_1(A) = \begin{cases} 0, & A = \emptyset \\ +\infty, & \text{o.c.} \end{cases} \quad \text{and} \quad \mu_2(A) = \#A,$$

for any $A \in \mathcal{B}$. Notice that μ_1 is the one given by the construction in the proof of the Carathéodory extension theorem.

Another example is the following. Let $\Omega = [0,1] \times [0,1]$ and the product algebra $\mathcal{A} = \mathcal{B}([0,1]) \times \mathcal{P}([0,1])$. Define the product measure $\mu = m \times \nu$ on $\mathcal{A}$ where ν is the counting measure. Thus, two extensions of μ are

$$\mu_1(A) = \sum_{y\colon (x,y)\in A} m(A_y) \quad \text{and} \quad \mu_2(A) = \int_0^1 n_A(x)\, dx,$$

for any $A \in \sigma(\mathcal{A})$, where

$$A_x = \{y \in [0,1] \colon (x,y) \in A\}, \quad A_y = \{x \in [0,1] \colon (x,y) \in A\}$$

and $n_A(x) = \#A_x$. In particular, $D = \{(x,x) \in \Omega \colon x \in [0,1]\}$ is in $\sigma(\mathcal{A})$ but not in $\mathcal{A}$, and

$$\mu_1(D) = 0 \quad \text{and} \quad \mu_2(D) = 1.$$

The extension given by construction in the proof of the Carathéodory extension theorem is

$$\mu_3(A) = \begin{cases} m \times \nu(A), & \cup_x A_x \text{ is countable and } m(\cup_y A_y) = 0 \\ +\infty, & \text{o.c.} \end{cases}$$

So, $\mu_3(D) = +\infty$.

CHAPTER 3

Measurable functions

1. Definition

Let $(\Omega_1, \mathcal{F}_1)$ and $(\Omega_2, \mathcal{F}_2)$ be measurable spaces and consider a function between those spaces $f\colon \Omega_1 \to \Omega_2$. We say that f is $(\mathcal{F}_1, \mathcal{F}_2)$-measurable iff

$$f^{-1}(B) \in \mathcal{F}_1, \quad B \in \mathcal{F}_2.$$

That is, the pre-image of a measurable set is also measurable (with respect to the respective σ-algebras). This definition will be useful in order to determine the measure of a set in $\mathcal{F}_2$ by looking at the measure of its pre-image in $\mathcal{F}_1$. Whenever there is no ambiguity, namely the σ-algebras are known and fixed, we will simply say that the function is measurable.

REMARK 3.1. Notice that the pre-image can be seen as a function between the collection of subsets, i.e. $f^{-1}\colon \mathcal{P}(\Omega_2) \to \mathcal{P}(\Omega_1)$. So, f is measurable iff the image under f^{-1} of $\mathcal{F}_2$ is contained in $\mathcal{F}_1$, i.e.

$$f^{-1}(\mathcal{F}_2) \subset \mathcal{F}_1.$$

EXERCISE 3.2. Show the following propositions:

(1) If f is $(\mathcal{F}_1, \mathcal{F}_2)$-measurable, it is also $(\mathcal{F}, \mathcal{F}_2)$-measurable for any σ-algebra $\mathcal{F} \supset \mathcal{F}_1$.
(2) If f is $(\mathcal{F}_1, \mathcal{F}_2)$-measurable, it is also $(\mathcal{F}_1, \mathcal{F})$-measurable for any σ-algebra $\mathcal{F} \subset \mathcal{F}_2$.

We do not need to check the conditon of measurability for every measurable set in $\mathcal{F}_2$. In fact, it is only required for a collection that generates $\mathcal{F}_2$.

PROPOSITION 3.3. *Let $\mathcal{I} \subset \mathcal{P}(\Omega_2)$. Then, f is $(\mathcal{F}_1, \sigma(\mathcal{I}))$-measurable iff $f^{-1}(\mathcal{I}) \subset \mathcal{F}_1$.*

PROOF.

(1) $(\Rightarrow)$ Since any $I \in \mathcal{I}$ also belongs to $\sigma(\mathcal{I})$, if f is measurable then $f^{-1}(I)$ is in $\mathcal{F}_1$.
(2) $(\Leftarrow)$ Let

$$\mathcal{F} = \{B \in \sigma(\mathcal{I})\colon f^{-1}(B) \in \mathcal{F}_1\}.$$

Notice that $\mathcal{F}$ is a σ-algebra because

- $f^{-1}(\emptyset) = \emptyset \in \mathcal{F}_1$, so $\emptyset \in \mathcal{F}$.
- If $B \in \mathcal{F}$, then $f^{-1}(B^c) = f^{-1}(B)^c \in \mathcal{F}_1$. Hence, B^c is also in $\mathcal{F}$.
- Let $B_1, B_2, \dots \in \mathcal{F}$. Then,
$$f^{-1}\left(\bigcup_{n=1}^{+\infty} B_n\right) = \bigcup_{n=1}^{+\infty} f^{-1}(B_n) \in \mathcal{F}_1.$$
So, $\bigcup_{n=1}^{+\infty} B_n$ is also in $\mathcal{F}$.

Since $\mathcal{I} \subset \mathcal{F}$ we have $\sigma(\mathcal{I}) \subset \mathcal{F} \subset \sigma(\mathcal{I})$. That is, $\mathcal{F} = \sigma(\mathcal{I})$.

□

We will be particularly interested in the case of scalar functions, i.e. with values in $\mathbb{R}$. Fix the Borel σ-algebra $\mathcal{B}$ on $\mathbb{R}$. Recall that $\mathcal{B}$ can be generated by the collection $\mathcal{I} = \{]-\infty, x] \colon x \in \mathbb{R}\}$. So, from Proposition 3.3 we say that $f \colon \Omega \to \mathbb{R}$ is *$\mathcal{F}$-measurable* iff

$$f^{-1}(]-\infty, x]) \in \mathcal{F}, \quad x \in \mathbb{R}.$$

In probability theory, these functions are called *random variables*.

Remark 3.4. The following notation is widely used (especially in probability theory) to represent the pre-image of a set in $\mathcal{I}$:

$$\{f \leq x\} = \{\omega \in \Omega \colon f(\omega) \leq x\} = f^{-1}(]-\infty, x]).$$

Example 3.5. Consider the constant function $f(\omega) = a$, $\omega \in \Omega$, where $a \in \mathbb{R}$. Then,

$$f^{-1}(]-\infty, x]) = \begin{cases} \Omega, & x \geq a \\ \emptyset, & x < a. \end{cases}$$

So, $f^{-1}(]-\infty, x])$ belongs to any σ-algebra on Ω. That is, a constant function is always measurable regardless of the σ-algebra considered.

Example 3.6. Let $A \subset \Omega$. The *indicator function* of A is defined by

$$\mathcal{X}_A(\omega) = \begin{cases} 1, & \omega \in A \\ 0, & \omega \in A^c. \end{cases}$$

Therefore,

$$\mathcal{X}_A^{-1}(]-\infty, x]) = \begin{cases} \Omega, & x \geq 1 \\ A^c, & 0 \leq x < 1 \\ \emptyset, & x < 0. \end{cases}$$

So, $\mathcal{X}_A$ is $\mathcal{F}$-measurable iff $A \in \mathcal{F}$.

Exercise 3.7. Show that:

(1) $\mathcal{X}_{f^{-1}(A)} = \mathcal{X}_A \circ f$ for any $A \subset \Omega_2$ and $f \colon \Omega_1 \to \Omega_2$.

(2) $\mathcal{X}_A\mathcal{X}_B = \mathcal{X}_{A\cap B}$ for $A, B \subset \Omega_1$.

EXAMPLE 3.8. Recall that $f\colon \mathbb{R}^n \to \mathbb{R}$ is a continuous function iff the pre-image of any open set is open. Since open sets are Borel sets, it follows that any continuous function is $\mathcal{B}(\mathbb{R}^n)$-measurable.

REMARK 3.9. In order to simplify the language and whenever there is no ambiguity, we often write measurable without the explicit reference to the σ-algebra.

PROPOSITION 3.10. *Let $f, g\colon \Omega \to \mathbb{R}$ be measurable functions. Then, their sum $f+g$ and product fg are also measurable.*

PROOF. Let $F\colon \mathbb{R}^2 \to \mathbb{R}$ be a continuous function and $h\colon \Omega \to \mathbb{R}$ given by

$$h(x) = F(f(x), g(x)).$$

Since F is continuous, we have that $F^{-1}(]-\infty, a[)$ is open. Thus, for any $a \in \mathbb{R}$ we can write $F^{-1}(]-\infty, a[) = \bigcup_n I_n \times J_k$, where I_n and J_n are open intervals. So,

$$h^{-1}(]-\infty, a[) = \bigcup_n f^{-1}(I_n) \cap g^{-1}(J_n) \in \mathcal{F}.$$

That is, h is measurable. We complete the proof by applying this to $F(u,v) = u+v$ and $F(u,v) = uv$. □

PROPOSITION 3.11. *Let $f\colon \Omega_1 \to \Omega_2$ and $g\colon \Omega_2 \to \Omega_3$ be measurable functions. Then, $g \circ f$ is also measurable.*

PROOF. Prove it. □

2. Simple functions

Consider N pairwise disjoint sets $A_1, \dots, A_N \subset \Omega$ whose union is Ω. A function $\varphi\colon \Omega \to \mathbb{R}$ is a *simple function* on $A_1, \dots, A_N$ if there are different numbers $c_1, \dots, c_N \in \mathbb{R}$ such that

$$\varphi = \sum_{j=1}^{N} c_j \mathcal{X}_{A_j}.$$

That is, a simple function is constant on a finite number of sets that cover Ω: $\varphi(A_j) = c_j$. Hence φ has only a finite number of possible values.

PROPOSITION 3.12. *A function is simple on $A_1, \dots, A_N$ iff it is $\sigma(\{A_1, \dots, A_N\})$-measurable.*

PROOF.

(⇒) For $x \in \mathbb{R}$, take the set $J(x) = \{j : c_j \leq x\} \subset \{1, \dots, N\}$. Hence,

$$\varphi^{-1}(]-\infty, x]) = \bigcup_{j \in J(x)} \varphi^{-1}(c_j) = \bigcup_{j \in J(x)} A_j.$$

So, φ is $\sigma(\{A_1, \dots, A_N\})$-measurable.

(⇐) Suppose that f is not simple. So, for some j it is not constant on A_j (and so A_j has more than one element). Then, there are $\omega_1, \omega_2 \in A_j$ and $x \in \mathbb{R}$ such that

$$f(\omega_1) < x < f(\omega_2).$$

Hence, ω_1 is in $f^{-1}(]-\infty, x])$ but ω_2 is not. This means that this set can not be any of $A_1, \dots, A_N$, their complements or their unions. Therefore, f is not $\sigma(\{A_1, \dots, A_N\})$-measurable.

□

REMARK 3.13. From the above proposition it follows that a function is constant (i.e. a simple function on a unique set) iff it is measurable with respect to the trivial σ-algebra (thus to any other).

EXERCISE 3.14. Consider a simple function φ. Write $|\varphi|$ and determine if it is also simple.

REMARK 3.15. Consider two simple functions

$$\varphi = \sum_{j=1}^{N} c_j \mathcal{X}_{A_j} \quad \text{and} \quad \varphi' = \sum_{j'=1}^{N'} c'_{j'} \mathcal{X}_{A'_{j'}}.$$

Their sum is also a simple function given by

$$\varphi + \varphi' = \sum_{j=1}^{N} \sum_{j'=1}^{N'} (c_j + c'_{j'}) \mathcal{X}_{A_j \cap A'_{j'}},$$

and their product is the simple function

$$\varphi\varphi' = \sum_{j=1}^{N} \sum_{j'=1}^{N'} c_j c'_{j'} \mathcal{X}_{A_j \cap A'_{j'}}.$$

3. Convergence of sequences of measurable functions

Consider countably many measurable functions $f_n : \Omega \to \mathbb{R}$ ordered by $n \in \mathbb{N}$. This defines a sequence of measurable functions $f_1, f_2, \dots$. We denote such a sequence by its general term f_n. There are several notions of its convergence:

- f_n *converges pointwisely*[1] to f (i.e. $f_n \to f$) iff
$$\lim_{n\to+\infty} f_n(x) = f(x) \quad \text{for every } x \in \Omega.$$
We also say that f is the limit of f_n.
- f_n *converges uniformly* to f (i.e. $f_n \xrightarrow{u} f$) iff
$$\lim_{n\to+\infty} \sup_{x\in\Omega} |f_n(x) - f(x)| = 0.$$
- f_n *converges almost everywhere* to f (i.e. $f_n \xrightarrow{a.e.} f$) iff there is $A \in \mathcal{F}$ such that $\mu(A) = 0$ and
$$\lim_{n\to+\infty} f_n(x) = f(x) \qquad \text{for every } x \in A^c.$$
- f_n *converges in measure* to f (i.e. $f_n \xrightarrow{\mu} f$) iff for every $\varepsilon > 0$
$$\lim_{n\to+\infty} \mu\left(\{x \in \Omega \colon |f_n(x) - f(x)| \geq \varepsilon\}\right) = 0.$$

REMARK 3.16. In the case of a probability measure we refer to convergence almost everywhere (a.e.) as almost surely (a.s.), and convergence in measure as convergence in probability.

EXERCISE 3.17. Let $([0,1], \mathcal{B}([0,1]), m)$ the Lebesgue measure space and $f_n(x) = x^n$, $x \in [0,1]$, $n \in \mathbb{N}$. Determine the convergence of f_n.

EXERCISE 3.18. Determine the convergence of $\mathcal{X}_{A_n}$ when

(1) $A_n \uparrow A$
(2) $A_n \downarrow A$
(3) the sets $A_1, A_2, \dots$ are pairwise disjoint.

A function $f \colon \Omega \to \mathbb{R}$ is called *bounded* if there is $M > 0$ such that for every $\omega \in \Omega$ we have $|f(\omega)| \leq M$.

We use the notation
$$f_n \nearrow f$$
to mean that $f_n \to f$ and $f_n \leq f$.

PROPOSITION 3.19.

(1) *For every measurable function f there is a sequence of simple functions such that $\varphi_n \nearrow f$.*
(2) *For every bounded measurable function f there is a sequence of simple functions such that $\varphi_n \nearrow f$ and the convergence is uniform.*

PROOF.

[1] or simply, f_n converges to f

(1) Consider the simple functions

$$\varphi_n = \sum_{j=0}^{n2^{n+1}} \left(-n + \frac{j}{2^n}\right) \mathcal{X}_{A_{n,j}} + n\mathcal{X}_{f^{-1}([n,+\infty[)} - n\mathcal{X}_{f^{-1}(]-\infty,-n[)}$$

where

$$A_{n,j} = f^{-1}\left(\left[-n + \frac{j}{2^n}, -n + \frac{j+1}{2^n}\right[\right).$$

Notice that for any $\omega \in A_{n,j}$ we have

$$-n + \frac{j}{2^n} \le f(\omega) < -n + \frac{j+1}{2^n}$$

and

$$\varphi_n(\omega) = -n + \frac{j}{2^n}.$$

So,

$$f(\omega) - \frac{1}{2^n} < \varphi_n(\omega) \le f(\omega).$$

Therefore, $\varphi_n \to f$ for every $\omega \in \Omega$ since for n sufficiently large ω belongs to some $A_{n,j}$.

(2) Assume that $|f(\omega)| \le M$ for every $\omega \in \Omega$. Given $n \in \mathbb{N}$, let

$$c_j = -M + \frac{2(j-1)M}{n}, \quad j = 1, \dots, n.$$

Define the intervals $I_j = [c_j, c_j + 2M/n[$ for $j = 1, \dots, n-1$ and $I_n = [c_n, M]$. Clearly, these n intervals are pairwise disjoint and their union is $[-M, M]$. Take also $A_j = f^{-1}(I_j)$ which are pairwise disjoint measurable sets and cover Ω, and the sequence of simple functions

$$\varphi_n = \sum_{j=1}^{n} c_j \mathcal{X}_{A_j}.$$

On each A_j the function is valued in I_j, and it is always $2M/n$ close to c_j (corresponding to the length of I_j. Then,

$$\sup_{\omega \in \Omega} |\varphi_n(\omega) - f(\omega)| \le \frac{2M}{n}.$$

As $n \to +\infty$ we obtain $\varphi_n \xrightarrow{u} f$.

□

Recall the definitions of liminf and limsup for a sequence u_n in $\mathbb{R}$:

$$\liminf_{n\to+\infty} u_n = \sup_{n\in\mathbb{N}} \inf_{k\ge n} u_k, \tag{3.1}$$

$$\limsup_{n\to+\infty} u_n = \inf_{n\in\mathbb{N}} \sup_{k\ge n} u_k. \tag{3.2}$$

These numbers always exist in $\mathbb{R}\cup\{-\infty,+\infty\}$. Moreover, the sequence u_n converges if they are finite and equal to the limit of u_n ($\lim u_n$).

PROPOSITION 3.20.

(1) *The infimum, supremum, minimum and maximum of a sequence of measurable functions is also measurable.*
(2) *The liminf, limsup and limit (if it exists) of a sequence of measurable functions is also measurable.*

EXERCISE 3.21. Prove it.

4. Induced measure

Let $(\Omega_1, \mathcal{F}_1, \mu_1)$ be a measure space, $(\Omega_2, \mathcal{F}_2)$ a measurable space and $f\colon \Omega_1 \to \Omega_2$ a measurable function. Notice that $\mu_1 \circ f^{-1}$ defines a function $\mathcal{F}_2 \to \mathbb{R}$ since $f^{-1}\colon \mathcal{F}_2 \to \mathcal{F}_1$ and $\mu_1\colon \mathcal{F}_1 \to \mathbb{R}$.

PROPOSITION 3.22. *The function*

$$\mu_2 = \mu_1 \circ f^{-1}$$

is a measure on $\mathcal{F}_2$ called the induced measure. Moreover, if μ_1 is a probability measure, then μ_2 is also a probability measure.

EXERCISE 3.23. Prove it.

REMARK 3.24.

(1) The induced measure $\mu_1 \circ f^{-1}$ is sometimes called push-forward measure and denoted by $f_*\mu_1$.
(2) In probability theory the induced probability measure is also known as *distribution* of f. If $\Omega_2 = \mathbb{R}$ and $\mathcal{F}_2 = \mathcal{B}(\mathbb{R})$ it is know as well as probability distribution. We will always refer to it as distribution.

EXERCISE 3.25. Consider the measure space $(\Omega, \mathcal{P}, \delta_a)$ where δ_a is the Dirac measure at $a \in \Omega$. If $f\colon \Omega \to \mathbb{R}$ is measurable, what is its induced measure (distribution)?

EXERCISE 3.26. Compute $m \circ f^{-1}$ where $f(x) = 2x$, $x \in \mathbb{R}$, and m is the Lebesgue measure on $\mathbb{R}$.

5. Generation of σ-algebras by measurable functions

Consider a function $f\colon \Omega \to \mathbb{R}$. The smallest σ-algebra of Ω for which f is measurable is

$$\sigma(f) = \sigma(\{f^{-1}(B) \in \mathcal{F}\colon B \in \mathcal{B}\}).$$

It is called the σ-algebra generated by f. Notice that f will be also measurable for any other σ-algebra containing $\sigma(f)$.

When we have a finite set of funtions $f_1, \dots, f_n$, the smallest σ-algebra for which all these functions are measurable is

$$\sigma(f_1, \dots, f_n) = \sigma(\{f_i^{-1}(B) \in \mathcal{F} \colon B \in \mathcal{B}, i = 1, \dots, n\}).$$

We also refer to it as the σ-algebra generated by $f_1, \dots, f_n$.

Example 3.27. Let $A \subset \Omega$ and take the indicator function $\mathcal{X}_A$. Then,

$$\sigma(\mathcal{X}_A) = \sigma(\{\emptyset, \Omega, A^c\}) = \{\emptyset, \Omega, A, A^c\} = \sigma(\{A\}).$$

Similarly, for $A_1, \dots, A_n \subset \Omega$,

$$\sigma(\mathcal{X}_{A_1}, \dots, \mathcal{X}_{A_n}) = \sigma(\{A_1, \dots, A_n\}).$$

Exercise 3.28. Decide if the following propositions are true:

(1) $\sigma(f) = \sigma(\{f^{-1}(]-\infty, x] \colon x \in \mathbb{R}\})$.
(2) $\sigma(f + g) = \sigma(f, g)$.

6. Extended real-valued functions

We many applications it is convenient to consider the case of functions with values in $\bar{\mathbb{R}} = \mathbb{R} \cup \{-\infty, +\infty\}$. For such cases it is enough to consider the Borel σ-algebra of $\bar{\mathbb{R}}$. That is defined as

$$\mathcal{B}(\bar{\mathbb{R}}) = \sigma(\{[-\infty, a] \colon a \in \mathbb{R}\}).$$

Let $(\Omega, \mathcal{F})$ be a measurable space. We say that $f \colon \Omega \to \bar{\mathbb{R}}$ is $\mathcal{F}$-measurable (a random variable) iff it is $(\mathcal{F}, \mathcal{B}(\bar{\mathbb{R}}))$-measurable. This is equivalent to check that $f^{-1}([-\infty, a]) \in \mathcal{F}$ for every $a \in \mathbb{R}$.

CHAPTER 4

Lebesgue integral

Here we define the Lebesgue integral of a measurable function f on a measurable set A with respect to a measure μ. This is a huge generalization of the Riemann integral in $\mathbb{R}$ introduced in first year Calculus. There, the functions have anti-derivatives, sets are intervals and there is no mention of the measure, althought it is the Lebesgue measure that is being used (the length of the intervals).

Roughly speaking, the Lebesgue integral is a "sum" of the values of f at all points in A times a weight given by the measure μ. For probability measures it can be thought as the weighted average of f on A.

In the following, in order to simplify the language, we will drop the name Lebesgue when referring to the integral.

1. Definition

We will define the integral first for simple functions, then for non-negative measurable functions, and finally for measurable functions.

1.1. Integral of non-negative simple functions. Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and $\varphi\colon \Omega \to \mathbb{R}$ a non-negative simple function ($\varphi \geq 0$) of the form

$$\varphi = \sum_{j=1}^{N} c_j \mathcal{X}_{A_j},$$

where $c_j \geq 0$. The integral of a simple function φ with respect to the measure μ is

$$\int \varphi \, d\mu = \sum_{j=1}^{N} c_j \mu(A_j).$$

It is a number in $[0, +\infty[\cup \{+\infty\}$.

Remark 4.1.

(1) If $c_j = 0$ and $\mu(A_j) = +\infty$ we set $c_j \mu(A_j) = 0$. So, $\int 0 \, d\mu = 0$ for any measure μ.

(2) Consider the simple function $\varphi(x) = c_1\mathcal{X}_A + c_2\mathcal{X}_{A^c}$ where $\mu(A) = \mu(A^c) = +\infty$. If we had allowed $c_1 > 0$ and $c_2 < 0$, then there would be an indetermination $c_1\mu(A) + c_2\mu(A^c) = +\infty - \infty$. This is why in the definition of the above integral we restrict to non-negative simple functions.

(3) We frequently use the following notation so that the variable of integration is explicitly written:

$$\int \varphi\, d\mu = \int \varphi(x)\, d\mu(x).$$

Proposition 4.2. *Let $\varphi_1, \varphi_2 \geq 0$ be simple functions and $a_1, a_2 \geq 0$. Then,*

(1)
$$\int (a_1\varphi_1 + a_2\varphi_2)\, d\mu = a_1 \int \varphi_1\, d\mu + a_2 \int \varphi_2\, d\mu.$$

(2) *If $\varphi_1 \leq \varphi_2$, then*
$$\int \varphi_1\, d\mu \leq \int \varphi_2\, d\mu.$$

Proof.

(1) Let $\varphi, \widetilde{\varphi}$ be simple functions in the form
$$\varphi = \sum_{j=1}^{N} c_j \mathcal{X}_{A_j}, \qquad \widetilde{\varphi} = \sum_{j=1}^{\widetilde{N}} \widetilde{c}_j \mathcal{X}_{\widetilde{A}_j}$$
where $A_j = \varphi^{-1}(c_j)$, $\widetilde{A}_j = \widetilde{\varphi}^{-1}(\widetilde{c}_j)$. Then,
$$\begin{aligned}\int (\varphi + \widetilde{\varphi})\, d\mu &= \sum_{i,j} (c_i + \widetilde{c}_i)\mu(A_i \cap \widetilde{A}_j) \\ &= \sum_i c_i \sum_j \mu(A_i \cap \widetilde{A}_j) + \sum_j \widetilde{c}_j \sum_i \mu(A_i \cap \widetilde{A}_j).\end{aligned}$$
Notice that $\sum_j \mu(A_i \cap \widetilde{A}_j) = \mu(A_i)$ because the sets $\widetilde{A}_j$ are pairwise disjoint and its union is Ω. The same applies to $\sum_i \mu(A_i \cap \widetilde{A}_j) = \mu(\widetilde{A}_j)$. Hence,
$$\int (\varphi + \widetilde{\varphi})\, d\mu = \int \varphi\, d\mu + \int \widetilde{\varphi}\, d\mu.$$

(2) Prove it.

□

1.2. Integral of non-negative measurable functions. Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and $f\colon \Omega \to \mathbb{R}$ a non-negative measurable function ($f \geq 0$). Consider the set of all possible values of the integrals of non-negative simple functions that are not above f, i.e.

$$I(f) = \left\{ \int \varphi \, d\mu \colon 0 \leq \varphi \leq f, \varphi \text{ is simple} \right\}.$$

PROPOSITION 4.3. *There is $a \in \mathbb{R}_0^+ \cup \{+\infty\}$ such that $I(f) = [0, a[$ or $I(f) = [0, a]$.*

PROOF. Since $\int \varphi \, d\mu \geq 0$, then $I(f) \subset \mathbb{R}_0^+ \cup \{+\infty\}$. Moreover, $0 \in I(f)$ because $\int 0 \, d\mu = 0$ for the simple function $\varphi = 0$ and $0 \leq f$.

Suppose now that $x \in I(f)$ with $x > 0$. This means that there is a simple function $0 \leq \varphi \leq f$ such that $\int \varphi \, d\mu = x$. Considering $y \in [0, x]$, let $\tilde{\varphi} = \frac{y}{x}\varphi$. This is also a simple function satisfying $0 \leq \tilde{\varphi} \leq \varphi \leq f$. Furthermore,

$$\int \tilde{\varphi} \, d\mu = \frac{y}{x} \int \varphi \, d\mu = y \in I(f).$$

Therefore, $[0, x] \subset I(f)$.

The only sets which have the property $[0, x] \subset I(f)$ for every $x \in I(f)$ are the intervals $[0, a[$ and $[0, a]$ for some $a \geq 0$ or $a = +\infty$. $\square$

The integral of $f \geq 0$ with respect to the measure μ is defined to be

$$\int f \, d\mu = \sup I(f).$$

So, the integral always exists and it is either a finite number in $[0, +\infty[$ or $+\infty$.

1.3. Integral of measurable functions. Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and $f\colon \Omega \to \mathbb{R}$ a measurable function. There is a simple decomposition of f into its positive and negative parts:

$$\begin{aligned} f^+(x) &= \max\{f(x), 0\} \geq 0 \\ f^-(x) &= \max\{-f(x), 0\} \geq 0. \end{aligned}$$

Hence,

$$f(x) = f^+(x) - f^-(x)$$

and also

$$|f(x)| = \max\{f^+(x), f^-(x)\} = f^+(x) + f^-(x).$$

A measurable function f is *integrable* with respect to μ iff $\int |f| \, d\mu < +\infty$. Its integral is defined as

$$\int f \, d\mu = \int f^+ \, d\mu - \int f^- \, d\mu.$$

The integral of f on $A \in \mathcal{F}$ with respect to μ is

$$\int_A f\,d\mu = \int f\,\mathcal{X}_A\,d\mu.$$

EXERCISE 4.4. Consider a simple function φ (not necessarily non-negative). Show that:

(1) $\varphi.\mathcal{X}_A = \sum_j c_j \mathcal{X}_{A_j \cap A}$.
(2)
$$\int_A \varphi\,d\mu = \sum_{j=1}^{N} c_j \mu(A_j \cap A).$$

In probability theory the integral of an integrable random variable X on a probability space $(\Omega, \mathcal{F}, P)$, is denoted by

$$E(X) = \int X\,dP$$

and called the *expected value*[1] of X.

REMARK 4.5. As for simple functions we will also be using the notation:

$$\int_A f\,d\mu = \int_A f(x)\,d\mu(x).$$

2. Properties

PROPOSITION 4.6. *Let f and g be integrable functions and $A, B \in \mathcal{F}$.*

(1) *If $f \leq g$ μ-a.e. then $\int f\,d\mu \leq \int g\,d\mu$.*
(2) *If $A \subset B$, then $\int_A |f|\,d\mu \leq \int_B |f|\,d\mu$.*
(3) *If $\mu(A) = 0$ then $\int_A f\,d\mu = 0$.*
(4) *If $\mu(A \cap B) = 0$ then $\int_{A\cup B} f\,d\mu = \int_A f\,d\mu + \int_B f\,d\mu$.*
(5) *If $f = 0$ μ-a.e. then $\int f\,d\mu = 0$.*
(6) *If $f \geq 0$ and $\lambda > 0$ then*
$$\mu\left(\{x \in \Omega \colon f(x) \geq \lambda\}\right) \leq \frac{1}{\lambda} \int f\,d\mu \quad \textit{(Markov inequality).}$$
(7) *If $f \geq 0$ and $\int f\,d\mu = 0$, then $f = 0$ μ-a.e.*
(8) *$(\inf f)\,\mu(\Omega) \leq \int f\,d\mu \leq (\sup f)\,\mu(\Omega)$.*

PROOF.

[1]It is also known as expectation, mathematical expectation, mean value, mean, average or first moment. It is sometimes denoted by $E[X]$, $\mathbb{E}(X)$ or $\langle X \rangle$.

(1) Any simple function satisfying $\varphi \leq f^+$ a.e. also satisfies $\varphi \leq g^+$ a.e. since $f^+ \leq g^+$ a.e. So, $I(f^+) \leq I(g^+)$ and $\int f^+ \, d\mu \leq \int g^+ \, d\mu$. Similarly, $g^- \leq f^-$ a.e. and $\int g^- \, d\mu \leq \int f^- \, d\mu$. Finally, $\int f^+ \, d\mu - \int f^- \, d\mu \leq \int g^+ \, d\mu - \int g^- \, d\mu$.
(2) Notice that $\int_A |f| \, d\mu = \int |f| \mathcal{X}_A \, d\mu$ and similarly for the integral in B. Since $|f|\mathcal{X}_A \leq |f|\mathcal{X}_B$, by the previous property, $\int |f|\mathcal{X}_A \, d\mu \leq \int |f|\mathcal{X}_B \, d\mu$.
(3) For any simple function $\int \varphi \, d\mu = \sum_i a_i \mu(A_i \cap A) = 0$. Thus, $\int_A |f| \, d\mu = 0$.
(4) Suppose that $f \geq 0$. For any $C \in \mathcal{F}$ we have

$$\begin{aligned}\mu((A \cup B) \cap C) &= \mu((A \cap C) \cup (B \cap C)) \\ &= \mu(A \cap C) + \mu(B \cap C)\end{aligned}$$

as $A \cap B \cap C \subset A \cap B$ has zero measure. Given a simple function $0 \leq \varphi \leq f$ we have

$$\begin{aligned}\int_{A \cup B} \varphi \, d\mu &= \sum_{j=1}^{N} c_j \mu(A_j \cap (A \cup B)) \\ &= \sum_{j=1}^{N} c_j (\mu(A_j \cap A) + \mu(A_i \cap B)) \\ &= \int_A \varphi \, d\mu + \int_B \varphi \, d\mu.\end{aligned}$$

Using the relation $\sup(g_1 + g_2) \leq \sup g_1 + \sup g_2$ for any functions g_1, g_1, we obtain

$$\int_{A \cup B} f \, d\mu \leq \int_A f \, d\mu + \int_B f \, d\mu.$$

Now, consider simple functions $0 \leq \varphi_1, \varphi_2 \leq f$, Since A and B are disjoint and $0 \leq \varphi_1 \mathcal{X}_A + \varphi_2 \mathcal{X}_B \leq f$ we get

$$\begin{aligned}\int_A \varphi_1 \, d\mu + \int_B \varphi_2 \, d\mu &= \int_{A \cup B} (\varphi_1 \mathcal{X}_A + \varphi_2 \mathcal{X}_B) \, d\mu \\ &\leq \int_{A \cup B} f \, d\mu.\end{aligned}$$

Considering the supremum over the simple functions, we get

$$\int_A f \, d\mu + \int_B f \, d\mu \leq \int_{A \cup B} f \, d\mu.$$

We have thus proved that $\int_{A\cup B} f^\pm \, d\mu = \int_A f^\pm \, d\mu + \int_B f^\pm \, d\mu$. This implies that

$$\begin{aligned}\int_{A\cup B} f\, d\mu &= \int_{A\cup B} f^+ \, d\mu - \int_{A\cup B} f^- \, d\mu \\ &= \int_A f^+ \, d\mu + \int_B f^+ \, d\mu - \int_A f^- \, d\mu - \int_B f^- \, d\mu \\ &= \int_A f\, d\mu + \int_B f\, d\mu.\end{aligned}$$

(5) We have $0 \leq f \leq 0$ a.e. Then, by the the first property, $\int 0\, d\mu \leq \int f\, d\mu \leq \int 0\, d\mu$.

(6) Let $A = \{x \in \Omega \colon f(x) \geq \lambda\}$. Then,

$$\int f\, d\mu \geq \int_A f\, d\mu \geq \int_A \lambda\, d\mu = \lambda\mu(A).$$

(7) We want to show that $\mu(\{x \in \Omega \colon f(x) > 0\}) = \mu \circ f^{-1}(]0, +\infty[) = 0$. The Markov inequality implies that for any $n \in \mathbb{N}$,

$$\mu \circ f^{-1}\left(\left[\frac{1}{n}, +\infty\right[\right) \leq n \int f\, d\mu = 0.$$

Since

$$f^{-1}(]0, +\infty[) = \bigcup_{n=1}^{+\infty} f^{-1}\left(\left[\frac{1}{n}, +\infty\right[\right),$$

we have

$$\mu \circ f^{-1}(]0, +\infty[) \leq \sum_{n=1}^{+\infty} \mu \circ f^{-1}\left(\left[\frac{1}{n}, +\infty\right[\right) = 0.$$

(8) It is enough to notice that $\inf f \leq f \leq \sup f$.

□

3. Examples

We present now two fundamental examples of integrals, constructed with the Dirac and the Lebesgue measures.

3.1. Integral for the Dirac measure. Consider the measure space $(\Omega, \mathcal{P}, \delta_a)$ where δ_a is the Dirac measure at $a \in \Omega$. We start by determining the integral of a simple function $\varphi \geq 0$ written in the usual form $\varphi = \sum_{j=1}^N c_j \mathcal{X}_{A_j}$. So, there is a unique $1 \leq k \leq N$ such that $a \in A_k$ (since the sets A_j are pairwise disjoint and their union is Ω) and c_k is the value in A_k. In particular, $\varphi(a) = c_k$. This implies that

$$\int \varphi\, d\delta_a = \sum_j c_j \delta_a(A_j) = c_k = \varphi(a).$$

Any function $f\colon \Omega \to \mathbb{R}$ is measurable for the σ-algebra considered. Take any $f^+ \geq 0$. Its integral is computed from the fact that

$$I(f) = \left\{\varphi(a)\colon 0 \leq \varphi \leq f^+, \varphi \text{ simple}\right\} = [0, f^+(a)].$$

Therefore, for any function $f = f^+ - f^-$ we have

$$\int f\, d\delta_a = \int f^+\, d\delta_a - \int f^-\, d\delta_a = f^+(a) - f^-(a) = f(a).$$

3.2. Integral for the Lebesgue measure. Let $(\mathbb{R}, \mathcal{B}, m)$ be the measure space associated to the Lebesgue measure m and a measurable function $f\colon I \to \mathbb{R}$ where $I \subset \mathbb{R}$ is an interval. We use the notation

$$\int_a^b f(t)\, dt = \begin{cases} \int_{[a,b]} f\, dm, & a \leq b \\ -\int_{[b,a]} f\, dm, & b < a, \end{cases}$$

where $a, b \in I$. Notice that we write $dm(t) = dt$ when the measure is the Lebesgue one.

Consider some $a \in I$ and the function $F\colon I \to \mathbb{R}$ given by

$$F(x) = \int_a^x f(t)\, dt.$$

Theorem 4.7. *If f is continuous at $x \in \operatorname{Int} I$, then F is differentiable at x and*

$$F'(x) = f(x).$$

Proof. By the definition, the derivative of F at x is, if it exists, given by

$$F'(x) = \lim_{y \to x} \frac{F(y) - F(x)}{y - x} = \lim_{y \to x} \frac{\int_x^y f(t)\, dt}{y - x}.$$

Now, if $x < y$,

$$\inf_{[x,y]} f \leq \frac{\int_x^y f(t)\, dt}{y - x} \leq \sup_{[x,y]} f.$$

As $y \to x^+$ we get $\inf_{[x,y]} f \to f(x)$ and $\sup_{[x,y]} f \to f(x)$ because f is continuous at x. Similarly for the case $y < x$. So, $F'(x) = f(x)$. $\square$

Remark 4.8. We call F an anti-derivative of f. It is not unique, there are other functions whose derivative is equal to f.

Exercise 4.9. Show that if F_1 and F_2 are anti-derivatives of f, then $F_1 - F_2$ is a constant function.

Theorem 4.10. *If f is continuous in I and has an anti-derivative F, then for any $a, b \in I$ we have*

$$\int_a^b f(t)\, dt = F(b) - F(a).$$

Proof. Theorem 4.7 and the fact that the anti-derivative is determined up to a constant, imply that $\int_a^b f(t)\,dt = F(b) + c$ where c is a constant. To determine c it is enough to compute $0 = \int_a^a f(t)\,dt = F(a) + c$, thus $c = -F(a)$. $\square$

Example 4.11. Let $f\colon \mathbb{R} \to \mathbb{R}$ given by $f(x) = e^{-|x|}$. It is a continuous function on $\mathbb{R}$ and

$$\int_0^x e^{-|t|}\,dt = \begin{cases} 1 - e^{-x}, & x \geq 0 \\ -(1 - e^{x}), & x < 0. \end{cases}$$

4. Convergence theorems

The computation of the integral of a function $f \geq 0$ is not direct in most cases. It requires considering all simple functions below f and determine the supremum of the set of all their integrals. As a measurable function is the limit of a sequence of simple functions φ_n, it would be very convenient to have the integral of f as just the limit of the integrals of φ_n. This indeed is a particular case of the convergence theorems (monotone and dominated) which study the relation between the limit and the integral of sequences of functions.

4.1. Monotone convergence. We start by a preliminar result that will be used later. Let $(\Omega, \mathcal{F}, \mu)$ be a measure space.

Lemma 4.12 (Fatou). *Let f_n be a sequence of measurable functions such that $f_n \geq 0$. Then,*

$$\int \liminf_{n\to+\infty} f_n\,d\mu \leq \liminf_{n\to+\infty} \int f_n\,d\mu.$$

Proof. Consider any simple function $0 \leq \varphi \leq \liminf f_n$, $0 < c < 1$ and the increasing sequence of measurable functions

$$g_n = \inf_{k\geq n} f_k.$$

Thus, for a sufficiently large n we have

$$c\varphi < g_n \leq \sup g_n = \liminf f_n.$$

Let

$$A_n = \{x \in \Omega\colon g_n(x) \geq c\varphi(x)\}.$$

So, $A_n \subset A_{n+1}$ and $\bigcup_n A_n = \Omega$. In addition,

$$\int_{A_n} c\varphi\,d\mu \leq \int_{A_n} g_n\,d\mu \leq \int_{A_n} f_k\,d\mu \leq \int f_k\,d\mu$$

for any $k \geq n$. Finally,

$$\int_{A_n} c\varphi\,d\mu \leq \inf_{k\geq n} \int f_k\,d\mu \leq \liminf \int f_n\,d\mu.$$

Therefore, since the previous inequality is valid for any $0 < c < 1$ and any n large,

$$\int \varphi\, d\mu \leq \liminf \int f_n\, d\mu.$$

It remains to observe that the definition of the integral requires that

$$\int \liminf f_n\, d\mu = \sup \left\{ \int \varphi\, d\mu \colon 0 \leq \varphi \leq \liminf f_n \right\}.$$

The claim follows immediately. $\square$

EXAMPLE 4.13. Consider $f_n = \mathcal{X}_{[n,n+1]}$. So, $\int_{\mathbb{R}} f_n\, dm = 1$ for any $n \in \mathbb{N}$. Thus, $\liminf_{n\to+\infty} \int_{\mathbb{R}} f_n\, dm = 1$. On the other hand, $\liminf_{n\to+\infty} f_n(x) = 0$ for all $x \in \mathbb{R}$. Therefore, $\int_{\mathbb{R}} \liminf_{n\to+\infty} f_n\, dm = 0$, which agrees with the theorem.

The next result is the first one for limits and not just liminf.

THEOREM 4.14 (Monotone convergence). *Let $f_n \geq 0$ be a sequence of measurable functions. If $f_n \nearrow f$ a.e., then*

$$\int \lim_{n\to+\infty} f_n\, d\mu = \lim_{n\to+\infty} \int f_n\, d\mu.$$

PROOF. Notice that $\int f_n \leq \int \lim_{n\to+\infty} f_n$. Hence,

$$\limsup_{n\to+\infty} \int f_n \leq \int \lim_{n\to+\infty} f_n = \int \liminf_{n\to+\infty} f_n \leq \liminf_{n\to+\infty} \int f_n$$

where we have used Fatou's lemma. Since $\liminf$ is always less or equal to $\limsup$, the above inequality implies that they have to be the same and equal to lim. $\square$

REMARK 4.15. This result applied to a sequence of random variables $X_n \geq 0$ on a probability space is the following: if $X_n \nearrow X$ a.s., then $E(\lim X_n) = \lim E(X_n)$.

4.2. More properties.

PROPOSITION 4.16. *Let f, g integrable functions on $(\Omega, \mathcal{F}, P)$ and $\alpha, \beta \in \mathbb{R}$.*

(1) $\int (\alpha f + \beta g)\, d\mu = \alpha \int f\, d\mu + \beta \int g\, d\mu$.
(2) *If $\int_A f\, d\mu \leq \int_A g\, d\mu$ for all $A \in \mathcal{F}$, then $f \leq g$ μ-a.e.*
(3) *If $\int_A f\, d\mu = \int_A g\, d\mu$ for all $A \in \mathcal{F}$, then $f = g$ μ-a.e.*
(4) $|\int f\, d\mu| \leq \int |f|\, d\mu$.

PROOF.

(1) Consider sequences of non-negative simple functions $\varphi_n \nearrow f^+$ and $\widetilde{\varphi}_n \nearrow g^+$. By Proposition 4.2 and the monotone convergence theorem applied twice,

$$\begin{aligned}\int (\alpha f^+ + \beta g^+)\, d\mu &= \int \lim(\alpha\varphi_n + \beta\psi_n)\, d\mu \\ &= \lim \int (\alpha\varphi_n + \beta\psi_n)\, d\mu \\ &= \lim \alpha \int \varphi_n\, d\mu + \lim \beta \int \psi_n\, d\mu \\ &= \alpha \int \lim \varphi_n\, d\mu + \beta \int \lim \psi_n\, d\mu \\ &= \alpha \int f^+\, d\mu + \beta \int g^+\, d\mu.\end{aligned}$$

The same is done for $\alpha f^- + \beta g^-$ and the result follows immediately.

(2) By writing $\int_A (g - f)\, d\mu \geq 0$ for all $A \in \mathcal{F}$, we want to show that $h = g - f \geq 0$ a.e. or equivalently that $h^- = 0$ a.e. Let

$$A = \{x \in \Omega \colon h(x) < 0\} = \{x \in \Omega \colon h^-(x) > 0\}.$$

Then, on A we have $h = -h^-$ and

$$0 \leq \int_A h\, d\mu = \int_A -h^-\, d\mu \leq 0.$$

That is, $\int_A h^-\, d\mu = 0$ and $h^- \geq 0$. So, by (7) of Proposition 4.6 we obtain that $h^- = 0$ a.e.

(3) Notice that $\int_A f\, d\mu = \int_A g\, d\mu$ implies that

$$\int_A f\, d\mu \leq \int_A g\, d\mu \leq \int_A f\, d\mu.$$

By (2) we get $f \leq g$ on a set of full measure and $g \leq f$ on another set of full measure. Since the intersection of both sets has still full measure, we have $f = g$ a.e.

(4) From the definition of the integral

$$\begin{aligned}\left| \int f\, d\mu \right| &= \left| \int f^+\, d\mu - \int f^-\, d\mu \right| \\ &\leq \left| \int f^+\, d\mu \right| + \left| \int f^-\, d\mu \right| \\ &= \int f^+\, d\mu + \int f^-\, d\mu \\ &= \int |f|\, d\mu.\end{aligned}$$

□

Proposition 4.17. *Let $(\Omega_1, \mathcal{F}_1, \mu_1)$ be a measure space, $(\Omega_2, \mathcal{F}_2)$ a measurable space and $f\colon \Omega_1 \to \Omega_2$ measurable. If $\mu_2 = \mu_1 \circ f^{-1}$ is the induced measure, then*

$$\int_{\Omega_2} g\, d\mu_2 = \int_{\Omega_1} g \circ f\, d\mu_1$$

for any $g\colon \Omega_2 \to \mathbb{R}$ measurable.

Proof. Consider a simple function φ in the form

$$\varphi = \sum_{j=1}^{+\infty} c_j \mathcal{X}_{A_j}.$$

Then,

$$\begin{aligned}
\int_{\Omega_2} \varphi\, d\mu_2 &= \sum_{j=1}^{+\infty} c_j \int_{\Omega_2} \mathcal{X}_{A_j}\, d\mu_2 \\
&= \sum_{j=1}^{+\infty} c_j \mu_1 \circ f^{-1}(A_j) \\
&= \sum_{j=1}^{+\infty} c_j \int_{f^{-1}(A_j)} d\mu_1 \\
&= \sum_{j=1}^{+\infty} c_j \int_{\Omega_1} \mathcal{X}_{f^{-1}(A_j)}\, d\mu_1 \\
&= \int_{\Omega_1} \varphi \circ f\, d\mu_1.
\end{aligned}$$

So, the result is proved for simple functions.

Take a sequence of non-negative simple functions $\varphi_n \nearrow g^+$ (we can use a similar approach for g^-) noting that $\varphi_n \circ f \nearrow g^+ \circ f$. We can therefore use the monotone convergence theorem for the sequences of simple functions φ_n and $\varphi_n \circ f$. Thus,

$$\begin{aligned}
\int_{\Omega_2} g^+\, d\mu_2 &= \int_{\Omega_2} \lim \varphi_n\, d\mu_2 \\
&= \lim \int_{\Omega_2} \varphi_n\, d\mu_2 \\
&= \lim \int_{\Omega_1} \varphi_n \circ f\, d\mu_1 \\
&= \int_{\Omega_1} \lim \varphi_n \circ f\, d\mu_1 \\
&= \int_{\Omega_1} g^+ \circ f\, d\mu_1.
\end{aligned}$$

□

EXAMPLE 4.18. Consider a probability space $(\Omega, \mathcal{F}, P)$ and $X\colon \Omega \to \mathbb{R}$ a random variable. By setting the induced measure $\alpha = P \circ X^{-1}$ and $g\colon \mathbb{R} \to \mathbb{R}$ we have

$$E(g(X)) = \int g \circ X \, dP = \int g(x) \, d\alpha(x).$$

In particular, $E(X) = \int x \, d\alpha(x)$.

PROPOSITION 4.19. *Consider the measure*

$$\mu = \sum_{n=1}^{+\infty} a_n \mu_n,$$

where μ_n is a measure and $a_n \geq 0$, $n \in \mathbb{N}$. If $f\colon \Omega \to \mathbb{R}$ satisfies

$$\sum_{n=1}^{+\infty} a_n \int |f| \, d\mu_n < +\infty,$$

then f is also μ-integrable and

$$\int f \, d\mu = \sum_{n=1}^{+\infty} a_n \int f \, d\mu_n.$$

PROOF. Recall Exercise 2.27 showing that μ is a measure. Suppose that $f \geq 0$. Take a sequence of simple functions

$$\varphi_k = \sum_j c_j \mathcal{X}_{A_j}$$

such that $\varphi_k \nearrow f$ as $k \to +\infty$. Then,

$$\int \varphi_k \, d\mu = \sum_j c_j \mu(A_j) = \sum_{n=1}^{+\infty} a_n \sum_j c_j \mu_n(A_j) = \sum_{n=1}^{+\infty} a_n \int \varphi_k \, d\mu_n,$$

which is finite for every k because $\varphi_k \leq |f|$. Now, using the monotone convergence theorem, f is μ-integrable and

$$\int f \, d\mu = \lim_{k \to +\infty} \int \varphi_k \, d\mu = \lim_{k \to +\infty} \lim_{m \to +\infty} b_{k,m},$$

where

$$b_{k,m} = \sum_{n=1}^{m} a_n \int \varphi_k \, d\mu_n.$$

Notice that $b_{k,m} \geq 0$, it is increasing both on k and on m and bounded from above. So, $A = \sup_k \sup_m b_{k,m} = \lim_k \lim_m b_{k,m}$. Define also $B = \sup_{k,m} b_{k,m}$. We want to show that $A = B$.

For all k and m we have $B \geq b_{k,m}$, so $B \geq A$. Given any $\varepsilon > 0$ we can find k_0 and m_0 such that $B - \varepsilon \leq b_{k_0,m_0} \leq B$. This implies that

$$A = \sup_k \sup_m b_{k,m} \geq \sup_k b_{k,m_0} \geq b_{k_0,m_0} \geq B - \varepsilon.$$

Taking $\varepsilon \to 0$ we get $A = B$. The same arguments above can be used to show that $A = B = \lim_m \lim_k b_{k,m}$.

We can thus exchange the limits order and, again by the monotone convergence theorem,

$$\int f \, d\mu = \sum_{n=1}^{+\infty} a_n \int \lim_k \varphi_k \, d\mu_n = \sum_{n=1}^{+\infty} a_n \int f \, d\mu_n.$$

Consider now f not necessarily ≥ 0. Using the decomposition $f = f^+ - f^-$ with $f^+, f^- \geq 0$, we have $|f| = f^+ + f^-$. Thus, f is μ-integrable and

$$\int f \, d\mu = \int f^+ \, d\mu - \int f^- \, d\mu = \sum_{n=1}^{+\infty} a_n \int (f^+ - f^-) \, d\mu_n.$$

$\square$

4.3. Dominated convergence.

THEOREM 4.20 (Dominated convergence). *Let f_n be a sequence of measurable functions and g an integrable function. If f_n converges a.e. and for any $n \in \mathbb{N}$ we have*

$$|f_n| \leq g \quad a.e.,$$

then,

$$\int \lim_{n\to+\infty} f_n \, d\mu = \lim_{n\to+\infty} \int f_n \, d\mu.$$

PROOF. Suppose that $0 \leq f_n \leq g$. By Fatou's lemma,

$$\int \lim_{n\to+\infty} f_n \, d\mu \leq \liminf_{n\to+\infty} \int f_n \, d\mu.$$

It remains to show that $\limsup_{n\to+\infty} \int f_n \, d\mu \leq \int \lim_{n\to+\infty} f_n \, d\mu$.

Again using Fatou's lemma,

$$\begin{aligned}
\int g \, d\mu - \int \lim_{n\to+\infty} f_n \, d\mu &= \int \lim_{k\to+\infty} (g - f_n) \, d\mu \\
&\leq \liminf_{n\to+\infty} \int (g - f_n) \, d\mu \\
&= \int g \, d\mu - \limsup_{n\to+\infty} \int f_n \, d\mu.
\end{aligned}$$

This implies that

$$\limsup_{n\to+\infty} \int f_n \, d\mu \leq \int \lim_{n\to+\infty} f_n \, d\mu.$$

For $|f_n| \leq g$, we have $\max\{f_n^+, f_n^-\} \leq g$ and $\lim_{n\to+\infty} \int f_n^\pm \, d\mu = \int \lim_{n\to+\infty} f_n^\pm \, d\mu$. □

EXAMPLE 4.21.

(1) Consider $\Omega =]0,1[$ and

$$f_n(x) = \frac{n \sin x}{1 + n^2\sqrt{x}}.$$

So,

$$|f_n(x)| \leq \frac{n}{1 + n^2\sqrt{x}} \leq \frac{1}{\sqrt{x}}.$$

As $g(x) = 1/\sqrt{x}$ is integrable,

$$\lim_{n\to+\infty} \int f_n \, dm = \int \lim_{n\to+\infty} f_n \, dm = 0.$$

(2)

$$\lim_{n\to+\infty} \int_{\mathbb{R}^2} e^{-(x^2+y^2)^n} \, dx \, dy = \int_{\mathbb{R}^2} \lim_{n\to+\infty} e^{-(x^2+y^2)^n} \, dx \, dy = \int_D dm = \pi,$$

where we have used the fact that $|e^{-(x^2+y^2)^n}| \leq e^{-(x^2+y^2)}$ is integrable and

$$\lim e^{-(x^2+y^2)^n} = \begin{cases} \frac{1}{e}, & (x,y) \in \partial D \\ 0, & (x,y) \in \mathbb{R}^2 \setminus \overline{D} \\ 1, & (x,y) \in D \end{cases}$$

with $D = \{(x,y) \in \mathbb{R}^2 \colon x^2 + y^2 < 1\}$.

EXERCISE 4.22. Determine the following limits:

(1) $\lim_{n\to+\infty} \int_0^{+\infty} \frac{r^n}{1+r^{n+2}} dr$
(2) $\lim_{n\to+\infty} \int_0^{\pi} \frac{\sqrt[n]{x}}{1+x^2} \, dx$
(3) $\lim_{n\to+\infty} \int_{-\infty}^{+\infty} e^{-|x|} \cos^n(x) \, dx$
(4) $\lim_{n\to+\infty} \int_{\mathbb{R}^2} \frac{1+\cos^n(x-y)}{(x^2+y^2+1)^2} \, dx \, dy$

5. Fubini theorem

Let $(\Omega_1, \mathcal{F}_1, P_1)$ and $(\Omega_2, \mathcal{F}_2, P_2)$ be probability spaces. Consider the product probability space $(\Omega, \mathcal{F}, P)$. Given $x_1 \in \Omega_1$ and $x_2 \in \Omega_2$ take $A \in \mathcal{F}$ and its sections

$$A_{x_1} = \{x_2 \in \Omega_2 \colon (x_1, x_2) \in A\},$$
$$A_{x_2} = \{x_1 \in \Omega_1 \colon (x_1, x_2) \in A\}.$$

EXERCISE 4.23. Show that

(1) for any $A \subset \Omega$,

$$(A^c)_{x_1} = (A_{x_1})^c \quad \text{and} \quad (A^c)_{x_2} = (A_{x_2})^c. \tag{4.1}$$

(2) for any $A_1, A_2, \dots \subset \Omega$,

$$\left(\bigcup_{n\in\mathbb{N}} A_n\right)_{x_1} = \bigcup_{n\in\mathbb{N}} (A_n)_{x_1} \quad \text{and} \quad \left(\bigcup_{n\in\mathbb{N}} A_n\right)_{x_2} = \bigcup_{n\in\mathbb{N}} (A_n)_{x_2}. \tag{4.2}$$

PROPOSITION 4.24. *For every $x_1 \in \Omega_1$ and $x_2 \in \Omega_2$, we have $A_{x_2} \in \mathcal{F}_1$ and $A_{x_1} \in \mathcal{F}_2$.*

PROOF. Consider the collection

$$\mathcal{G} = \{A \in \mathcal{F} \colon A_{x_2} \in \mathcal{F}_1, x_2 \in \Omega_2\}.$$

We want to show that $\mathcal{G} = \mathcal{F}$.

Notice that any measurable rectangle $B = B_1 \times B_2$ with $B_1 \in \mathcal{F}_1$ and $B_2 \in \mathcal{F}_2$ is in $\mathcal{G}$. In fact, $B_{x_2} = B_1$ if $x_2 \in B_2$, otherwise it is empty.

If $\mathcal{I}$ is the collection of all measurable rectangles, then $\mathcal{I} \subset \mathcal{G} \subset \mathcal{F}$. This implies that $\mathcal{F} = \sigma(\mathcal{I}) \subset \sigma(\mathcal{G}) \subset \mathcal{F}$ and $\sigma(\mathcal{G}) = \mathcal{F}$. It is now enough to show that $\mathcal{G}$ is a σ-algebra. This follows easily by using (4.1) and (4.2). □

PROPOSITION 4.25. *Let $A \in \mathcal{F}$.*

(1) *The function $x_1 \mapsto P_2(A_{x_1})$ on Ω_1 is measurable.*
(2) *The function $x_2 \mapsto P_1(A_{x_2})$ on Ω_2 is measurable.*
(3)

$$P(A) = \int P_2(A_{x_1})\, dP_1(x_1) = \int P_1(A_{x_2})\, dP_2(x_2).$$

PROOF. Given $A \in \mathcal{F}$ write $f_A(x_1) = P_2(A_{x_1})$ and $g_A(x_2) = P_1(A_{x_2})$. Denote the collection of all measurable rectangles by $\mathcal{I}$ and consider

$$\mathcal{G} = \left\{A \in \mathcal{F} \colon f_A \text{ and } g_A \text{ are measurable}, \int f_A\, dP_1 = \int g_A\, dP_2\right\}.$$

We want to show that $\mathcal{G} = \mathcal{F}$.

We start by looking at measurable rectangles whose collection we denote by $\mathcal{I}$. For each $B = B_1 \times B_2 \in \mathcal{I}$ we have that $f_B = P_2(B_2)\mathcal{X}_{B_1}$ and $g_B = P_1(B_1)\mathcal{X}_{B_2}$ are simple functions, thus measurable for $\mathcal{F}_1$ and $\mathcal{F}_2$, respectively. In addition,

$$P(A) = \int f_A\, dP_1 = \int g_A\, dP_2 = P_1(B_1)P_2(B_2).$$

So, $\mathcal{I} \subset \mathcal{G}$. The same can be checked for the finite union of measurable rectangles, corresponding to an algebra $\mathcal{A}$, so that $\mathcal{A} \subset \mathcal{G}$.

We now show that $\mathcal{G}$ is a monotone class. Take an increasing sequence $A_n \uparrow A$ in $\mathcal{G}$. Hence, their sections are increasing as well as f_{A_n} and g_{A_n}. Moreover, $f_A = \lim f_{A_n}$ and $g_A = \lim g_{A_n}$ are measurable. Finally, since $\int f_{A_n}\, dP_1 = \int g_{A_n}\, dP_2$ holds for every n, by the monotone convergence theorem $\int f_A\, dP1 = \int g\, dP_2$. That means that $A \in \mathcal{G}$. The same argument can be carried over to decreasing sequences $A_n \downarrow A$. Therefore, $\mathcal{G}$ is a monotone class.

By Theorem 2.19 we know that $\sigma(\mathcal{A}) \subset \mathcal{G}$. Since $\mathcal{F} = \sigma(\mathcal{A})$ and $\mathcal{G} \subset \mathcal{F}$ we obtain that $\mathcal{G} = \mathcal{F}$. Also, $P(A) = \int f_A\, dP_1$ for any $A \in \mathcal{F}$ by extending this property for measurable rectangles. □

Remark 4.26. There exist examples of non-measurable sets ($A \subset \Omega$ but $A \notin \mathcal{F}$) with measurable sections and measurable functions $P_2(A_{x_1})$ and $P_1(A_{x_2})$ whose integrals differ.

Consider now a measurable function $f \colon \Omega \to \mathbb{R}$. Given $x_1 \in \Omega_1$ we define

$$f_{x_1} \colon \Omega_2 \to \mathbb{R}, \quad f_{x_1}(x_2) = f(x_1, x_2).$$

Similarly, for $x_2 \in \Omega_2$ let

$$f_{x_2} \colon \Omega_1 \to \mathbb{R}, \quad f_{x_2}(x_1) = f(x_1, x_2).$$

Exercise 4.27. Show that if f is measurable, then f_{x_1} and f_{x_2} are measurable for each x_1 and x_2, respectively.

Define

$$I_1 \colon \Omega_1 \to \mathbb{R}, \quad I_1(x_1) = \int f_{x_1}\, dP_2$$

and

$$I_2 \colon \Omega_2 \to \mathbb{R}, \quad I_2(x_2) = \int f_{x_2}\, dP_1.$$

Exercise 4.28. Show that if f is measurable, then I_1 and I_2 are measurable.

Theorem 4.29 (Fubini). *Let $f \colon \Omega \to \mathbb{R}$*

(1) *If f is an integrable function, then f_{x_1} and f_{x_2} are integrable for a.e. x_1 and x_2, respectively. Moreover, I_1 and I_2 are integrable functions and*
$$\int f\, dP = \int I_1\, dP_1 = \int I_2\, dP_2.$$
(2) *If $f \geq 0$ and I_1 is an integrable function, then f is integrable.*

Exercise 4.30. Prove it.

EXERCISE 4.31. Consider the Lebesgue probability space $([0,1],\mathcal{B},m)$. Write an example of a measurable function f such that I_1 and I_2 are integrable but $\int I_1\,dP_1 \neq \int I_2\,dP_2$.

CHAPTER 5

Distributions

From now on we focus on probability theory and use its notations and nomenclatures. That is, we interpret Ω as the set of outcomes of an experiment, $\mathcal{F}$ as the collection of events (sets of outcomes), P as a probability measure, and measurable functions as random variables (numerical result of an observation).

In this chapter we are going to explore a correspondence between distributions and two types of functions: distribution functions and characteristic functions. It is simpler to study functions than measures. Determining a measure requires knowing its value for every measurable set, a much harder task than to understand a function.

1. Definition

Let $(\Omega, \mathcal{F}, P)$ be a probability space and $X\colon \Omega \to \mathbb{R}$ a random variable. The *distribution of* X (or the *law* of X) is the induced probability measure $\alpha\colon \mathcal{B}(\mathbb{R}) \to \mathbb{R}$,

$$\alpha = P \circ X^{-1}.$$

In general we say that any probability measure P on $\mathbb{R}$ is a distribution by considering the identity random variable $X(x) = x$ so that $\alpha = P$.

It is common in probability theory to use several notations that are appropriate in the context. We list below some of them:

(1) $P(X \in A) = P(X^{-1}(A)) = \alpha(A)$
(2) $P(X \in A, X \in B) = \alpha(A \cap B)$
(3) $P(X \in A \text{ or } X \in B) = \alpha(A \cup B)$
(4) $P(X \notin A) = \alpha(A^c)$
(5) $P(X \in A, X \notin B) = \alpha(A \setminus B)$
(6) $P(X = a) = \alpha(\{a\})$
(7) $P(X \leq a) = \alpha(]-\infty, a])$
(8) $P(a < X \leq b) = \alpha(]a, b])$

EXERCISE 5.1. Suppose that $f\colon \mathbb{R} \to \mathbb{R}$ is a $\mathcal{B}(\mathbb{R})$-measurable function and α is the distribution of a random variable X. Find the distribution of $f \circ X$.

When the random variable is multidimensional, i.e. $X\colon \Omega \to \mathbb{R}^d$ and $X = (X_1, \dots, X_d)$, we call the induced measure $\alpha\colon \mathcal{B}(\mathbb{R}^d) \to \mathbb{R}$ given by $\alpha = P \circ X^{-1}$ the *joint distribution of* $X_1, \dots, X_d$. Here we denote the product σ-algebra by $\mathcal{B}(\mathbb{R}^d)$.

In most applications it is the distribution α that really matters. For example, suppose that Ω is the set of all possible states of the atmosphere. If X is the function that gives the temperature (°C) in Lisbon for a given state of the atmosphere and $I = [20, 21]$,

$$\alpha(I) = P(X^{-1}(I)) = P(X \in I) = P(20 \leq X \leq 21)$$

is the probability of the temperature being between 20°C and 21°C. That is, we first compute the set $X^{-1}(I)$ of all states that correspond to a temperature in Lisbon inside the interval I, and then find its probability measure.

It is important to be aware that for the vast majority of systems in the real world, we do not know Ω and P. So, one needs to guess α. Finding the right distribution is usually a very difficult task, if not impossible. Nevertheless, a frequently convenient way to acquire some knowledge of α is by treating statistically the data from experimental observations. In particular, it is possible to determine good approximations of each moment of order n of α (if it exists):

$$m_n = E(X^n) = \int x^n \, d\alpha(x), \quad n \in \mathbb{N}, \quad m_0 = 1,$$

Knowing the moments is a first step towards a choice of the distribution could be, but in general it does not determine it uniquely. Notice that $E(X^n)$ exists if $E(|X^n|) < +\infty$ (i.e. X^n is integrable).

Remark 5.2.

(1) The moment of order 1 is the expectation of X,

$$m_1 = E(X).$$

(2) The variance of X is defined as

$$\operatorname{Var}(X) = E(X - E(X))^2 = E(X^2) - E(X)^2 = m_2 - m_1^2.$$

(3) Given two integrable random variables X and Y, their covariance is

$$\operatorname{Cov}(X, Y) = E((X - E(X))(Y - E(Y))) = E(XY) - E(X)E(Y).$$

If $\operatorname{Cov}(X, Y) = 0$, then we say that X and Y are uncorrelated.

Exercise 5.3. Show that $\operatorname{Var}(X) = 0$ iff $P(X = E(X)) = 1$.

Exercise 5.4. Show that for each distribution α there is n_0 such that m_n exists for every $n \leq n_0$ and it does not exist otherwise.

EXERCISE 5.5. Consider $X_1, \dots, X_n$ integrable random variables. Show that if $\mathrm{Cov}(X_i, X_j) = 0$, $i \neq j$, then

$$\mathrm{Cov}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{Var}(X_i).$$

EXERCISE 5.6. Let X be a random variable and $\lambda \geq 0$. Prove the Tchebychev inequalities:

(1)
$$P(|X| \geq \lambda) \leq \frac{1}{\lambda^k} E(|X|^k).$$

(2) For $k \in \mathbb{N}$,
$$P(|X - E(X)| \geq \lambda) \leq \frac{\mathrm{Var}(X)}{\lambda^2}.$$

2. Simple examples

Here are some examples of random variables for which one can find explicitly their distributions.

EXAMPLE 5.7. Consider X to be constant, i.e. $X(x) = c$ for any $x \in \Omega$ and some $c \in \mathbb{R}$. Then, given $B \in \mathcal{B}$ we obtain

$$X^{-1}(B) = \begin{cases} \emptyset, & c \notin B \\ \Omega, & c \in B. \end{cases}$$

Hence,

$$\alpha(B) = P(X^{-1}(B)) = \begin{cases} P(\emptyset) = 0, & c \notin B \\ P(\Omega) = 1, & c \in B. \end{cases}$$

That is, $\alpha = \delta_c$ is the Dirac distribution at c. Finally, $E(g(X)) = \int g(x)\, d\alpha(x) = g(c)$, so $m_n = c^n$ and in particular

$$E(X) = c \quad \text{and} \quad \mathrm{Var}(X) = 0.$$

EXAMPLE 5.8. Given $A \in \mathcal{F}$ and constants $c_1, c_2 \in \mathbb{R}$, let $X = c_1\mathcal{X}_A + c_2\mathcal{X}_{A^c}$. Then, for $B \in \mathcal{B}$ we get

$$X^{-1}(B) = \begin{cases} A, & c_1 \in B, c_2 \notin B \\ A^c, & c_1 \notin B, c_2 \in B \\ \Omega, & c_1, c_2 \in B \\ \emptyset, & \text{o.c.} \end{cases}$$

So, the distribution of $c\mathcal{X}_A$ is

$$\alpha(B) = \begin{cases} p, & c_1 \in B, c_2 \notin B \\ 1-p, & c_1 \notin B, c_2 \in B \\ 1, & c_1, c_2 \in B \\ 0, & \text{o.c.} \end{cases}$$

where $p = P(A)$. That is,

$$\alpha = p\delta_{c_1} + (1-p)\delta_{c_2}$$

is the so-called *Bernoulli distribution*. Hence, $E(g(X)) = \int g(x)\,d\alpha(x) = pg(c_1) + (1-p)g(c_2)$ and $m_n = pc_1^n + (1-p)c_2^n$. In particular,

$$E(X) = pc_1 + (1-p)c_2, \qquad \mathrm{Var}(X) = p(1-p)(c_1+c_2)^2.$$

EXERCISE 5.9. Find the distribution of a simple function in the form

$$X = \sum_{j=1}^{N} c_j \mathcal{X}_{A_j}$$

and compute its moments.

3. Distribution functions

A function $F\colon \mathbb{R} \to \mathbb{R}$ is a *distribution function* iff

(1) it is increasing, i.e. for any $x_1 < x_2$ we have $F(x_1) \le F(x_2)$,
(2) it is continuous from the right at every point, i.e. $F(x^+) = F(x)$,
(3) $F(-\infty) = 0$, $F(+\infty) = 1$.

The next theorem states that there is a one-to-one correspondence between distributions and distribution functions.

THEOREM 5.10 (Lebesgue).

(1) *If α is a distribution, then*

$$F(x) = \alpha(]-\infty, x]), \qquad x \in \mathbb{R},$$

is a distribution function.
(2) *If F is a distribution function, then there is a unique distribution α such that*

$$\alpha(]-\infty, x]) = F(x), \qquad x \in \mathbb{R}.$$

REMARK 5.11. The function F as above is called the *distribution function* of α. Whenever α is the distribution of a random variable X, we also say that F is the distribution function of X and

$$F(x) = P(X \le x).$$

PROOF.

(1) For any $x_1 \leq x_2$ we have $]-\infty, x_1] \subset]-\infty, x_2]$. Thus, $F(x_1) \leq F(x_2)$ and F is increasing. Now, given any sequence $x_n \to a^+$,

$$\begin{aligned}\lim_{n\to+\infty} F(x_n) &= \lim_{n\to+\infty} \alpha(]-\infty, x_n]) \\ &= \alpha\left(\bigcap_{n=1}^{+\infty}]-\infty, x_n]\right) \\ &= \alpha(]-\infty, a]) = F(a).\end{aligned}$$

That is, F is continuous from the right for any $a \in \mathbb{R}$. Finally, using Proposition 2.31,

$$\begin{aligned}F(-\infty) &= \lim_{n\to+\infty} F(-n) \\ &= \lim_{n\to+\infty} \alpha(]-\infty, -n]) \\ &= \alpha\left(\bigcap_{n=1}^{+\infty}]-\infty, -n]\right) \\ &= \alpha(\emptyset) = 0.\end{aligned}$$

and

$$\begin{aligned}F(+\infty) &= \lim_{n\to+\infty} F(n) \\ &= \lim_{n\to+\infty} \alpha(]-\infty, n]) \\ &= \alpha\left(\bigcup_{n=1}^{+\infty}]-\infty, n]\right) \\ &= \alpha(\mathbb{R}) = 1.\end{aligned}$$

(2) Consider the algebra $\mathcal{A}(\mathbb{R})$ that contains every finite union of intervals of the form $]a, b]$ (see section 1.2). Take a sequence of disjoint intervals $]a_n, b_n]$, $-\infty \leq a_n \leq b_n \leq +\infty$, whose union is in $\mathcal{A}(\mathbb{R})$ and define

$$\alpha\left(\bigcup_{n=1}^{+\infty}]a_n, b_n]\right) = \sum_{n=1}^{+\infty} (F(b_n) - F(a_n)).$$

Thus, α is σ-additive, $\alpha(\emptyset) = \alpha(]a, a]) = 0$, $\alpha(A) \geq 0$ for any $A \in \mathcal{A}(\mathbb{R})$ because F is increasing, and $\alpha(\mathbb{R}) = F(+\infty) - F(-\infty) = 1$. Thus, α is a probability measure on $\mathcal{A}(\mathbb{R})$. In particular, $\alpha(]-\infty, x]) = F(x)$, $x \in \mathbb{R}$.

Finally, Carathéodory extension theorem guarantees that α can be uniquely extended to a distribution in $\sigma(\mathcal{A}(\mathbb{R})) = \mathcal{B}(\mathbb{R})$.

$\square$

EXERCISE 5.12. Show that for $-\infty \leq a \leq b \leq +\infty$ we have

(1) $\alpha(\{a\}) = F(a) - F(a^-)$
(2) $\alpha(]a,b[) = F(b^-) - F(a)$
(3) $\alpha([a,b[) = F(b^-) - F(a^-)$
(4) $\alpha([a,b]) = F(b) - F(a^-)$
(5) $\alpha(]-\infty,b[) = F(b^-)$
(6) $\alpha([a,+\infty[) = 1 - F(a^-)$
(7) $\alpha(]a,+\infty[) = 1 - F(a)$

Exercise 5.13. Compute the distribution function of the following distributions:

(1) The Dirac distribution δ_a at $a \in \mathbb{R}$.
(2) The Bernoulli distribution $p\delta_a + (1-p)\delta_b$ with $0 \leq p \leq 1$ and $a, b \in \mathbb{R}$.
(3) The uniform distribution on a bounded interval $I \subset \mathbb{R}$

$$m_I(A) = \frac{m(A \cap I)}{m(I)}, \qquad A \in \mathcal{B}(\mathbb{R}),$$

where m is the Lebesgue measure.
(4) $\alpha = c_1\delta_a + c_2 m_I$ on $\mathcal{B}(\mathbb{R})$ where $c_1, c_2 \geq 0$ and $c_1 + c_2 = 1$.
(5)

$$\alpha = \sum_{n=1}^{+\infty} \frac{1}{2^n}\delta_{-1/n}.$$

4. Classification of distributions

Consider a distribution α and its correspondent distribution function $F\colon \mathbb{R} \to \mathbb{R}$. The set of points where F is discontinuous is denoted by

$$D = \left\{x \in \mathbb{R}\colon F(x^-) < F(x)\right\}.$$

Proposition 5.14.

(1) *$\alpha(\{a\}) > 0$ iff $a \in D$.*
(2) *D is countable.*
(3) *$D = \emptyset$ iff $\alpha(D) = 0$.*

Exercise 5.15. Prove this.

4.1. Discrete. A distribution function $F\colon \mathbb{R} \to \mathbb{R}$ is called *discrete* if it is piecewise constant, i.e. for any $n \in \mathbb{N}$ we can find $a_n \in \mathbb{R}$ and $p_n \geq 0$ such that $\sum_n p_n = 1$ and

$$F(x) = \sum_{n=1}^{+\infty} p_n \mathcal{X}_{]-\infty,a_n]}(x) = \sum_{a_n \geq x} p_n.$$

A distribution is called *discrete* iff its distribution function is discrete. So, a discrete distribution α is given by

$$\alpha(A) = \sum_{n=1}^{+\infty} p_n \delta_{a_n}(A) = \sum_{a_n \in A} p_n, \quad A \in \mathcal{B}(\mathbb{R}).$$

EXERCISE 5.16. Show that $\alpha(D) = 1$ iff α is a discrete distribution.

EXAMPLE 5.17. Consider a random variable X such that $P(X \in \mathbb{N}) = 1$. So,

$$P(X \geq n) = \sum_{i=n}^{+\infty} P(X = i).$$

Its expected value is then

$$\begin{aligned} E(X) &= \int X\, dP = \sum_{i=1}^{+\infty} iP(X = i) \\ &= \sum_{i=1}^{+\infty} \sum_{n=1}^{i} P(X = i) \\ &= \sum_{n=1}^{+\infty} \sum_{i=n}^{+\infty} P(X = i) \\ &= \sum_{n=1}^{+\infty} P(X \geq n). \end{aligned}$$

4.2. Absolutely continuous. We say that a distribution function is *absolutely continuous* if there is an integrable function $f \geq 0$ with respect to the Lebesgue measure m such that

$$F(x) = \int_{-\infty}^{x} f\, dm, \quad x \in \mathbb{R}.$$

In particular, F is continuous ($D = \emptyset$). Recall that by the fundamental theorem of calculus, if F is differentiable at x, then $F'(x) = f(x)$. Moreover, if f is continuous, then F is differentiable.

A distribution is called *absolutely continuous* iff its distribution function is absolutely continuous. An absolutely continuous distribution α is given by

$$\alpha(A) = \int_A f\, dm, \quad A \in \mathcal{B}(\mathbb{R}).$$

The function f is known as the density of α.

EXAMPLE 5.18. Take $\Omega = \mathbb{R}$, $\mathcal{F} = \mathcal{B}(\mathbb{R})$ and the Lebesgue measure m on $[0, 1]$. For a fixed $r > 0$ consider the random variable

$$X(\omega) = \omega^r \mathcal{X}_{[0,+\infty[}(\omega).$$

Thus,

$$\{X \leq x\} = \begin{cases} \emptyset, & x < 0 \\ ]-\infty, x^{1/r}], & x \geq 0 \end{cases}$$

and the distribution function of X is

$$F(x) = m(X \leq x) = \begin{cases} 0, & x < 0 \\ x^{1/r}, & 0 \leq x < 1 \\ 1, & x \geq 1. \end{cases}$$

The function F is absolutely continuous since

$$F(x) = \int_{-\infty}^{x} f(t)\, dt,$$

where $f(t) = F'(t)$ is the density function given by

$$f(t) = \frac{1}{r} t^{1/r-1}\, \mathcal{X}_{[0,1]}(t).$$

4.3. Singular continuous. We say that a distribution function is *singular continuous* if F is continuous ($D = \emptyset$) but not absolutely continuous.

A distribution is called *singular continuous* iff its distribution function is singular continuous.

4.4. Mixed. A distribution is called *mixed* iff it is not discrete, absolutely continuous or singular continuous.

5. Convergence in distribution

Consider a sequence of random variables X_n and the sequence of their distributions $\alpha_n = P \circ X_n^{-1}$. Moreover, we take the sequence of the corresponding distribution functions F_n and of the characteristic functions ϕ_n.

We say that X_n *converges in distribution* to a random variable X iff

$$\lim_{n\to+\infty} F_n(x) = F(x), \quad x \in D^c,$$

where F is the distribution function of X and D the set of its discontinuity points. We use the notation

$$X_n \xrightarrow{d} X.$$

Moreover, we say that α_n *converges weakly* to a distribution α iff

$$\lim_{n\to+\infty} \int f\, d\alpha_n = \int f\, d\alpha$$

for every $f\colon \mathbb{R} \to \mathbb{R}$ continuous and bounded. We use the notation

$$\alpha_n \xrightarrow{w} \alpha.$$

It turns out that it is enough to check the convergence of the above integral for a one-parameter family of complex-valued functions $g_t(x) = e^{itx}$ where $t \in \mathbb{R}$. Recall that $e^{itx} = \cos(tx) + i\sin(tx)$, so that

$$\int e^{itx}\, d\alpha(x) = \int \cos(tx)\, d\alpha(x) + i \int \sin(tx)\, d\alpha(x).$$

THEOREM 5.19. *Let α_n be a sequence of distributions. If*

$$\lim_{n\to+\infty} \int e^{itx}\, d\alpha_n$$

exists for every $t \in \mathbb{R}$ and it is continuous at 0, then there is a distribution α such that $\alpha_n \xrightarrow{w} \alpha$.

PROOF. Let F_n be the distribution function of each α_n. Let r_j be a sequence ordering the rational numbers. As $F_n(r_1) \in [0,1]$ there is a subsequence $k_1(n)$ (that is, $k_1\colon \mathbb{N} \to \mathbb{N}$ is strictly increasing) for which $F_{k_1(n)}(r_1)$ converges when $n \to +\infty$, say to b_{r_1}. Again, $F_{k_1(n)}(r_2) \in [0,1]$ implies that there is a subsequence $k_2(n)$ of $k_1(n)$ (meaning that $k_2\colon \mathbb{N} \to k_1(\mathbb{N})$ is strictly increasing) giving $F_{k_2(n)}(r_2) \to b_{r_2}$. Inductively, we can find subsequences $k_j(n)$ of $k_{j-1}(n)$ such that $F_{k_j(n)}(r_j) \to b_{r_j}$. Notice that the sequence $m(n) = k_n(n)$ is a subsequence of $k_j(n)$ when $n \geq j$. Therefore, for any j we have that $F_{m(n)}(r_j) \to b_{r_j}$. Since each distribution function F_n is increasing, for rationals $r < r'$ we have for any sufficiently large n that $F_{m(n)}(r) \leq F_{m(n)}(r')$.

Define $G_n = F_{m(n)}$. So, $G_n(r) \to b_r$ for any $r \in \mathbb{Q}$. In addition, for rationals $r < r'$ it holds $b_r \leq b_{r'}$. We now choose the function $G\colon \mathbb{R} \to [0,1]$ by

$$G(x) = \inf_{r>x} b_r.$$

We will now show that G is also a distribution function.

For $x_1 < x_2$ it is simple to check that

$$G(x_1) = \inf_{r>x_1} b_r \leq \inf_{r>x_2} b_r = G(x_2),$$

so that G is increasing. Take a sequence $x_n \to x^+$. Hence $G(x) =$

□

The following theorem shows that convergence in distribution for sequences of random variables is the same as weak convergence for their distributions. Moreover, this is equivalent to showing convergence of the integrals for a specific complex function $x \mapsto e^{itx}$ for each $t \in \mathbb{R}$. This last fact will be explored in the next section, and this integral will be called the characteristic function of the distribution.

THEOREM 5.20 (Lévy-Cramer continuity). *For each $n \in \mathbb{N}$ consider a distribution α_n with distribution function F_n. Let α be a distribution α with distribution function F. The following propositions are equivalent:*

(1) $F_n \to F$ *on* D^c,
(2) $\alpha_n \xrightarrow{w} \alpha$,
(3) *for each* $t \in \mathbb{R}$,

$$\lim_{n\to+\infty} \int e^{itx}\, d\alpha_n(x) = \int e^{itx}\, d\alpha(x).$$

PROOF.

(1)⇒(2) Assume that $F_n \to F$ on the set D^c of continuity points of F. Let $\varepsilon > 0$ and $a, b \in D^c$ such that $a < b$, $F(a) \leq \varepsilon$ and $F(b) \geq 1 - \varepsilon$. Then, there is $n_0 \in \mathbb{N}$ satisfying

$$F_n(a) \leq 2\varepsilon \quad \text{and} \quad F_n(b) \geq 1 - 2\varepsilon$$

for all $n \geq n_0$.

Let $\delta > 0$ and f continuous such that $|f(x)| \leq M$ for some $M > 0$. Take the following partition

$$]a, b] = \bigcup_{j=1}^{N} I_j, \quad I_j =]a_j, a_{j+1}],$$

where $a = a_1 < \cdots < a_{N+1} = b$ with $a_i \in D^c$ such that

$$\max_{I_j} f - \min_{I_j} f < \delta.$$

Consider now the simple function

$$h(x) = \sum_{j=1}^{N} f(a_j) \mathcal{X}_{I_j}.$$

Hence,

$$|f(x) - h(x)| \leq \delta, \quad x \in]a, b].$$

In addition,

$$\begin{aligned}
\left| \int (f-h)\, d\alpha_n \right| &= \left| \int_{]a,b]} (f-h)\, d\alpha_n + \int_{]a,b]^c} f\, d\alpha_n \right| \\
&\leq \delta\, \alpha_n(]a,b]) + (\max|f|)(F_n(a) + 1 - F_n(b)) \\
&\leq \delta + 4M\varepsilon.
\end{aligned}$$

Similarly,

$$\left| \int (f-h)\, d\alpha \right| \leq \delta + 2M\varepsilon.$$

In addition,

$$\alpha_n(I_j) - \alpha(I_j) = F_n(a_{j+1}) - F(a_{j+1}) - (F_n(a_j) - F(a_j))$$

converges to zero as $n \to +\infty$ and the same for

$$\left|\int h\,d\alpha_n - \int h\,d\alpha\right| = \left|\sum_{j=1}^{N} f(a_j)\,(\alpha_n(I_j) - \alpha(I_j))\right|$$

Therefore, using

$$\begin{aligned}\left|\int f\,d\alpha_n - \int f\,d\alpha\right| &= \left|\int (f-h)\,d\alpha_n - \int (f-h)\,d\alpha\right.\\ &\quad \left. + \int h\,d\alpha_n - \int h\,d\alpha\right|\end{aligned}$$

we obtain

$$\limsup_{n\to+\infty}\left|\int f\,d\alpha_n - \int f\,d\alpha\right| \leq 2\delta + 6M\varepsilon.$$

Being ε and δ arbitrary, we get $\alpha_n \xrightarrow{w} \alpha$.

(2)⇒(1) Let y be a continuity point of F. So, $\alpha(\{y\}) = 0$. Consider $A =]-\infty, y[$ and the sequence of functions

$$f_k(x) = \begin{cases} 1, x \leq y - \frac{1}{2^k} \\ -2^k(x-y), & y - \frac{1}{2^k} < x \leq y \\ 0, & x > y, \end{cases}$$

where $k \in \mathbb{N}$. Notice that $f_k \nearrow \mathcal{X}_A$. Thus, using the dominated convergence theorem

$$F(y) = \alpha(A) = \int \mathcal{X}_A\,d\alpha = \int \lim_k f_k\,d\alpha = \lim_k \int f_k\,d\alpha.$$

Since f_k is continuous and bounded, and $f_k \leq \mathcal{X}_A$,

$$\begin{aligned}\lim_k \int f_k\,d\alpha &= \lim_k \lim_n \int f_k\,d\alpha_n \\ &\leq \lim_k \liminf_n \int \mathcal{X}_A\,d\alpha_n = \liminf_n F_n(y)\end{aligned}$$

where it was also used the fact that $F_n(y^-) \leq F_n(y)$.

Now, take $A =]-\infty, y]$ and

$$f_k(x) = \begin{cases} 1, x \leq y \\ -2^k(x-y)+1, & y < x \leq y + \frac{1}{2^k} \\ 0, & x > y + \frac{1}{2^k}. \end{cases}$$

Similarly to above, as $f_k \searrow \mathcal{X}_A$,

$$F(y) = \lim_k \lim_n \int f_k\,d\alpha_n \geq \lim_k \limsup_n \int \mathcal{X}_A\,d\alpha_n = \limsup_n F_n(y).$$

Combining the two inequalities,

$$\limsup_n F_n(y) \leq F(y) \leq \liminf_n F_n(y)$$

we conclude that $F(y) = \lim_n F_n(y)$.

(2)⇒(3) Define $g_t(x) = e^{itx} = \cos(tx) + i\sin(tx)$ for each $t \in \mathbb{R}$. Since $\cos(tx)$ and $\sin(tx)$ are continuous and bounded as functions of x, by (2) we have $\lim_n \int g_t(x)\, d\alpha_n(x) = \int g_t(x)\, d\alpha(x)$.

(3)⇒(2) This follows from Theorem 5.19 by noticing that $t \mapsto \int e^{itx}\, d\alpha(x)$ is continuous at 0.

$\square$

EXERCISE 5.21. Show that if X_n converges in distribution to a constant, then it also converges in probability.

6. Characteristic functions

A function $\phi\colon \mathbb{R} \to \mathbb{C}$ is a *characteristic function* iff

(1) ϕ is continuous at 0,
(2) ϕ is positive definite, i.e.
$$\sum_{i=1}^{n}\sum_{j=1}^{n} \phi(t_i - t_j) z_i \overline{z}_j \in \mathbb{R}_0^+$$
for all $z_1, \dots, z_n \in \mathbb{C}$, $t_1, \dots, t_n \in \mathbb{R}$ and $n \in \mathbb{N}$,
(3) $\phi(0) = 1$.

The next theorem states that there is a one-to-one correspondence between distributions and characteristic functions.

THEOREM 5.22 (Bochner).

(1) *If α is a distribution, then*
$$\phi(t) = \int e^{itx}\, d\alpha(x), \qquad t \in \mathbb{R},$$
is a characteristic function.
(2) *If ϕ is a characteristic function, then there is a unique distribution α such that*
$$\int e^{itx}\, d\alpha(x) = \phi(t), \qquad t \in \mathbb{R}.$$

The above theorem is proved in section 6.3.

REMARK 5.23. The function ϕ as above is called the *characteristic function*[1] of α. Whenever α is the distribution of a random variable X, we also say that ϕ is the characteristic function of X and
$$\phi(t) = E(e^{itX}).$$

---

[1]It is also known as the *Fourier transform* of α. Notice that if α is absolutely continuous then ϕ is the Fourier transform of the density function.

Example 5.24. The characteristic function of the Dirac distribution δ_a at a point $a \in \mathbb{R}$ is

$$\phi(t) = \int e^{itx}\, d\delta_a(x) = e^{ita}.$$

Exercise 5.25. Let X and $Y = aX + b$ be random variables with $a, b \in \mathbb{R}$ and $a \neq 0$. Show that if ϕ_X is the characteristic function of the distribution of X, then

$$\phi_Y(t) = e^{itb}\phi_X(at), \qquad t \in \mathbb{R},$$

is the characteristic function of the distribution of Y.

Exercise 5.26. Let X be a random variable and ϕ_X its characteristic function. Show that the characteristic function of $-X$ is

$$\phi_{-X}(t) = \phi_X(-t).$$

Exercise 5.27. Let ϕ be the characteristic function of the distribution α. Prove that ϕ is real-valued (i.e. $\phi(t) \in \mathbb{R}$, $t \in \mathbb{R}$) iff α is symmetric around the origin (i.e. $\alpha(A) = \alpha(-A)$, $A \in \mathcal{B}$).

6.1. Regularity of the characteristic function. We start by presenting some facts about positive definite functions.

Exercise 5.28. Show the following statements:

(1) If ϕ is positive definite, then for any $a \in \mathbb{R}$ the function $\psi(t) = e^{ita}\phi(t)$ is also positive definite.
(2) If $\phi, \dots, \phi_n$ are positive definite functions and $a_1, \dots, a_n > 0$, then $\sum_{i=1}^{n} a_i\phi_i$ is also positive definite.

Lemma 5.29. *Suppose that $\phi\colon \mathbb{R} \to \mathbb{C}$ is a positive definite function. Then,*

(1) $0 \leq |\phi(t)| \leq \phi(0)$ *and* $\phi(-t) = \overline{\phi(t)}$ *for every* $t \in \mathbb{R}$.
(2) *for any* $s, t \in \mathbb{R}$,
$$|\phi(t) - \phi(s)|^2 \leq 4\phi(0)\, |\phi(0) - \phi(t-s)|.$$
(3) ϕ *is continuous at* 0 *iff it is uniformly continuous on* $\mathbb{R}$.

Proof.

(1) Take $n = 2$, $t_1 = 0$ and $t_2 = t$. Hence,
$$\phi(0)z_1\overline{z}_1 + \phi(-t)z_1\overline{z}_2 + \phi(t)z_2\overline{z}_1 + \phi(0)z_2\overline{z}_2 \in \mathbb{R}_0^+$$
for any choice of $z_1, z_2 \in \mathbb{C}$. In particular, using $z_1 = 1$ and $z_2 = 0$ we obtain $\phi(0) \in \mathbb{R}_0^+$. On the other hand, $z_1 = z_2 = 1$ implies that the imaginary part of $\phi(-t) + \phi(t)$ is zero. For $z_1 = 1$ and $z_2 = i$, we get that the real part of $\phi(-t) - \phi(t)$ is zero. Finally, $z_1 = \overline{z}_2 = \sqrt{-\phi(t)}$ yields that $|\phi(t)| \leq \phi(0)$.

(2) Fixing $n \in \mathbb{N}$ and $t_1, \dots, t_n \in \mathbb{R}$ we have that the matrix $[\phi(t_i - t_j)]_{i,j}$ is positive definite and Hermitian. In particular, by choosing $n = 3$, $t_1 = t$, $t_2 = s$ and $t_3 = 0$, we obtain that

$$\begin{bmatrix} \phi(0) & \phi(t-s) & \phi(t) \\ \overline{\phi(t-s)} & \phi(0) & \phi(s) \\ \overline{\phi(t)} & \overline{\phi(s)} & \phi(0) \end{bmatrix}$$

has a non-negative determinant given by

$$\phi(0)^3 + 2\operatorname{Re}(\phi(t-s)\phi(s)\overline{\phi(t)}) - \phi(0)(|\phi(t)|^2 + |\phi(s)|^2 + |\phi(t-s)|^2) \geq 0.$$

Hence, assuming that $\phi(0) > 0$ (otherwise the result is immediate),

$$\begin{aligned} |\phi(t) - \phi(s)| &= |\phi(t)|^2 + |\phi(s)|^2 - 2\operatorname{Re}\phi(s)\overline{\phi(t)} \\ &\leq \phi(0)^2 + 2\operatorname{Re}(\phi(t-s) - \phi(0))\frac{\phi(s)\overline{\phi(t)}}{\phi(0)} - |\phi(t-s)|^2 \\ &\leq (\phi(0) - |\phi(t-s)|)(\phi(0) + |\phi(t-s)| + 2\phi(0)) \\ &\leq 4\phi(0)|\phi(0) - \phi(t-s)|. \end{aligned}$$

(3) This follows from the previous estimate.

□

The previous lemma implies that any characteristic function ϕ is continuous everywhere and its absolute value is between 0 and 1. In the following we find a condition for the differentiability of ϕ.

Proposition 5.30. *If there is $k \in \mathbb{N}$ such that*

$$\int |x|^k \, d\alpha(x) < +\infty,$$

then ϕ is C^k and $\phi^{(k)}(0) = i^k m_k$.

Proof. Let $k = 1$. Then,

$$\begin{aligned} \phi'(t) &= \lim_{s \to 0} \frac{\phi(t+s) - \phi(t)}{s} \\ &= \lim_{s \to 0} \int e^{itx} \frac{e^{isx} - 1}{s} \, d\alpha(x) \\ &= \lim_{s \to 0} \int e^{itx} \sum_{n=1}^{+\infty} \frac{(ix)^n}{n!} s^{n-1} \, d\alpha(x) \\ &= \int e^{itx} \sum_{n=1}^{+\infty} \frac{(ix)^n}{n!} \lim_{s \to 0} s^{n-1} \, d\alpha(x) \\ &= \int e^{itx} ix \, d\alpha(x). \end{aligned}$$

This integral exists (it is finite) because

$$\left|\int e^{itx} ix \, d\alpha(x)\right| \leq \int |x| \, d\alpha(x) < +\infty$$

by hypothesis. Therefore, ϕ' exists and it is a continuous function of t. In addition, $\phi'(0) = i \int x \, d\alpha(x)$. The claim is proved for $k = 1$.

We can now proceed by induction for the remaining cases $k \geq 2$. This is left as an exercise for the reader. □

6.2. Examples. In the following there is a list of widely used distributions. For some of the examples below there is a special notation to indicate the distribution of a random variable X. For instance, if it is the Uniform distribution on $[a, b]$, we write $X \sim \mathrm{U([a, b])}$. Other cases are included in the next examples.

EXERCISE 5.31. Find the characteristic functions of the following discrete distributions $\alpha(A) = P(X \in A)$, $A \in \mathcal{B}(\mathbb{R})$:

(1) Dirac (or degenerate or atomic) distribution

$$\alpha(A) = \begin{cases} 1, & a \in A \\ 0, & \text{o.c.} \end{cases}$$

where $a \in \mathbb{R}$.

(2) Binomial distribution ($X \sim \mathrm{Bin}$) with $n \in \mathbb{N}$:

$$\alpha(\{k\}) = C^n_k \, p^k (1-p)^{n-k}, \quad 0 \leq k \leq n.$$

(3) Poisson distribution ($X \sim \mathrm{Poisson}(\lambda)$) with $\lambda > 0$:

$$\alpha(\{k\}) = \frac{\lambda^k}{k! e^\lambda}, \quad k \in \mathbb{N} \cup \{0\}.$$

This describes the distribution of 'rare' events with rate λ.

(4) Geometric distribution ($X \sim \mathrm{Geom(p)}$) with $0 \leq p \leq 1$:

$$\alpha(\{k\}) = (1-p)^k p, \quad k \in \mathbb{N} \cup \{0\}.$$

This describes the distribution of the number of unsuccessful attempts preceding a success with probability p.

(5) Negative binomial distribution ($X \sim \mathrm{NBin}$):

$$\alpha(\{k\}) = C^{n+k-1}_k (1-p)^k p^n, \quad k \in \mathbb{N} \cup \{0\}.$$

This describes the distribution of the number of accumulated failures before n successes. *Hint*: Recall the Taylor series of $\frac{1}{1-x} = \sum_{i=0}^{+\infty} x^i$ for $|x| < 1$. Differentiate this n times and use the result.

EXERCISE 5.32. Find the characteristic functions of the following absolutely continuous distributions $\alpha(A) = P(X \in A) = \int_A f(x) \, dx$, $A \in \mathcal{B}(\mathbb{R})$ where f is the density function:

(1) Uniform distribution on $[a,b]$ ($X \sim \mathrm{U([a,b])}$):

$$f(x) = \begin{cases} \frac{1}{b-a}, & x \in [a,b] \\ 0, & \text{o.c.} \end{cases}$$

(2) Exponential distribution ($X \sim \mathrm{Exp}(\lambda)$) with $\lambda > 0$:

$$f(x) = \lambda e^{-\lambda x}, \quad x \geq 0$$

(3) The two-sided exponential distribution ($X \sim \mathrm{Exp}(\lambda)$), with $\lambda > 0$:

$$f(x) = \frac{\lambda}{2} e^{-\lambda |x|}, \quad x \in \mathbb{R}$$

(4) The Cauchy distribution ($X \sim \mathrm{Cauchy}$):

$$f(x) = \frac{1}{\pi} \frac{1}{1+x^2}, \quad x \in \mathbb{R}$$

Hint: Use the residue theorem of complex analysis.

(5) The normal (Gaussian) distribution ($X \sim \mathrm{N}(\mu, \sigma)$) with mean μ and variance $\sigma^2 > 0$:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbb{R}$$

Exercise 5.33. Let $X_n \sim \mathrm{U}([0,1])$ and $Y_n = -\frac{1}{\lambda} \log(1 - X_n)$ with $\lambda > 0$. Show that $Y_n \sim \mathrm{Exp}(\lambda)$.

6.3. Proof of Bochner theorem.

Proposition 5.34. *Consider a distribution α and the function*

$$\phi(t) = \int e^{itx} \, d\alpha(x), \quad t \in \mathbb{R}.$$

Then, ϕ is a characteristic function, i.e.

(1) $\phi(0) = 1$,
(2) *ϕ is uniformly continuous,*
(3) *ϕ is positive definite.*

Proof.

(1) $\phi(0) = \int d\alpha = 1$.
(2) For any $s, t \in \mathbb{R}$ we have

$$\begin{aligned} |\phi(t) - \phi(s)| &= \left| \int (e^{itx} - e^{isx}) \, d\alpha(x) \right| \\ &\leq \int |e^{isx}| \, |e^{i(t-s)x} - 1| \, d\alpha(x) \\ &= \int |e^{i(t-s)x} - 1| \, d\alpha(x). \end{aligned}$$

Taking $s \to t$ we can use the dominated convergence theorem to show that

$$\lim_{s\to t} \int |e^{i(t-s)x} - 1| \, d\alpha(x) = \int \lim_{s\to t} |e^{i(t-s)x} - 1| \, d\alpha(x) = 0,$$

being enough to notice that $|e^{i(t-s)x} - 1|$ is bounded. So,

$$\lim_{s\to t} |\phi(t) - \phi(s)| = 0,$$

meaning that ϕ is uniformly continuous.

(3) For all $z_1, \dots, z_n \in \mathbb{C}$, $t_1, \dots, t_n \in \mathbb{R}$ and $n \in \mathbb{N}$,

$$\begin{aligned}\sum_{i,j=1}^{n} \phi(t_i - t_j) z_i \overline{z}_j &= \sum_{i,j=1}^{n} z_i \overline{z}_j \int e^{i(t_i - t_j)x} \, d\alpha(x) \\ &= \int \sum_{i=1}^{n} z_i e^{it_i x} \sum_{j=1}^{n} \overline{z_j e^{it_j x}} \, d\alpha(x) \\ &= \int \left| \sum_{i=1}^{n} z_i e^{it_i x} \right|^2 d\alpha(x) \geq 0.\end{aligned}$$

□

Proposition 5.35. *If $\phi\colon \mathbb{R} \to \mathbb{C}$ is a characteristic function, then there is a unique distribution α such that*

$$\int e^{itx} \, d\alpha(x) = \phi(t).$$

Proof. Visit the library. □

CHAPTER 6

Limit theorems

Let $(\Omega, \mathcal{F}, P)$ be a probability space and $(\mathbb{R}, \mathcal{B})$ the Borel measurable space. A *discrete-time stochastic process* is a sequence of random variables $X_n \colon \Omega \to \mathbb{R}$, $n \in \mathbb{N}$. For simplicity, we denote the stochastic process simply by its n-th term X_n.

A *continuous-time stochastic process* is a one-parameter family of random variables $X_t \colon \Omega \to \mathbb{R}$, $t \geq 0$. At every instant of time t we have a random variable X_t.

In this chapter we are only interested in discrete-time stochastic processes that are sequences of independent random variables all with the same distribution. We call those sequences iid (independent and identically distributed). In the next chapters we will study examples of stochastic processes that are dependent and do not have the same distribution.

In the following, whenever there is a property valid for a set of full probability measure, we will use the initials a.s. (almost surely) instead of a.e. (almost everywhere).

1. Independent random variables

Two events A_1 and A_2 are independent if they do not influence each other in terms of probability. This notion is fundamental in probability theory and it is stated in general in the following way. Let $(\Omega, \mathcal{F}, P)$ be a probability space. We say that $A_1, A_2 \in \mathcal{F}$ are *independent events* iff

$$P(A_1 \cap A_2) = P(A_1)\, P(A_2).$$

EXERCISE 6.1. Show that:

(1) If A_1 and A_2 are independent, then A_1^c and A_2 are also independent.
(2) Any full probability event is independent of any other event. The same for any zero probability event.
(3) Two disjoint events are independent iff at least one of them has zero probability.
(4) Consider two events $A_1 \subset A_2$. They are independent iff A_1 has zero probability or A_2 has full probability.

EXAMPLE 6.2. Consider the Lebesgue measure m on $\Omega = [0, 1]$ and the event $I_1 = [0, \frac{1}{2}]$. Any other interval $I_2 = [a, b]$ with $0 \leq a < b \leq 1$ that is independent of I_1 has to satisfy the relation $P(I_2 \cap [0, \frac{1}{2}]) = \frac{1}{2}(b-a)$. Notice that $a \leq \frac{1}{2}$ (otherwise $I_1 \cap I_2 = \emptyset$) and $b \geq \frac{1}{2}$ (otherwise $I_2 \subset I_1$). So, $b = 1 - a$. That is, any interval $[a, 1 - a]$ with $0 \leq a \leq \frac{1}{2}$ is independent of $[0, \frac{1}{2}]$.

EXERCISE 6.3. Suppose that A and C are independent events as well as B and C with $A \cap B = \emptyset$. Show that $A \cup B$ and C are also independent.

EXERCISE 6.4. Give examples of probability measures P_1 and P_2, and of events A_1 and A_2 such that $P_1(A_1 \cap A_2) = P_1(A_1)\, P_1(A_2)$ but $P_2(A_1 \cap A_2) \neq P_2(A_1)\, P_2(A_2)$. Recall that the definition of independence depends on the probability measure.

Two random variables X, Y are *independent random variables* iff

$$P(X \in B_1, Y \in B_2) = P(X \in B_1)\, P(Y \in B_2), \quad B_1, B_2 \in \mathcal{B}.$$

REMARK 6.5. The independence between X and Y is equivalent to any of the following propositions. For any $B_1, B_2 \in \mathcal{B}$,

(1) $X^{-1}(B_1)$ and $Y^{-1}(B_2)$ are independent events.
(2) $P((X, Y) \in B_1 \times B_2) = P(X \in B_1)\, P(Y \in B_2)$.
(3) $\alpha_Z(B_1 \times B_2) = \alpha_X(B_1)\, \alpha_Y(B_2)$, where $\alpha_Z = P \circ Z^{-1}$ is the joint distribution of $Z = (X, Y)$, $\alpha_X = P \circ X^{-1}$ and $\alpha_Y = P \circ Y^{-1}$ are the distributions of X and Y, respectively. We can therefore show that the joint distribution is the product measure

$$\alpha_Z = \alpha_X \times \alpha_Y.$$

EXAMPLE 6.6. Consider simple functions

$$X = \sum_{i=1}^{N} c_i \mathcal{X}_{A_i}, \quad Y = \sum_{j=1}^{N'} c'_j \mathcal{X}_{A'_j}.$$

Then, for any $B_1, B_2 \in \mathcal{B}$,

$$X^{-1}(B_1) = \bigcup_{i\colon c_i \in B_1} A_i, \quad Y^{-1}(B_2) = \bigcup_{j\colon c'_j \in B_2} A'_j.$$

These are independent events iff A_i and A'_j are independent for every i, j.

PROPOSITION 6.7. *Let X and Y be independent random variables. Then, there are sequences φ_n and φ'_n of simple functions such that*

(1) *$\varphi_n \nearrow X$ and $\varphi'_n \nearrow Y$,*
(2) *φ_n and φ'_n are independent for every $n \in \mathbb{N}$.*

PROOF. We follow the idea in the proof of Proposition 3.19. The construction there guarantees that we get $\varphi_n \nearrow X$ and $\varphi'_n \nearrow Y$ by considering the simple functions

$$\varphi_n = \sum_{j=0}^{n2^{n+1}} \left(-n + \frac{j}{2^n}\right) \mathcal{X}_{A_{n,j}} + n\mathcal{X}_{X^{-1}([n,+\infty[)} - n\mathcal{X}_{X^{-1}(]-\infty,-n[)}$$

where

$$A_{n,j} = X^{-1}\left(\left[-n + \frac{j}{2^n}, -n + \frac{j+1}{2^n}\right[\right),$$

and

$$\varphi'_n = \sum_{j=0}^{n2^{n+1}} \left(-n + \frac{j}{2^n}\right) \mathcal{X}_{A_{n,j}} + n\mathcal{X}_{Y^{-1}([n,+\infty[)} - n\mathcal{X}_{Y^{-1}(]-\infty,-n[)}$$

where

$$A'_{n,j} = Y^{-1}\left(\left[-n + \frac{j}{2^n}, -n + \frac{j+1}{2^n}\right[\right).$$

It remains to check that φ_n and φ'_n are independent for any given n. This follows from the fact that X and Y are independent, since any pre-image of a Borel set by X and Y are independent. $\square$

PROPOSITION 6.8. *If X and Y are independent, then*

$$E(XY) = E(X)\,E(Y)$$

and

$$\operatorname{Var}(X+Y) = \operatorname{Var}(X) + \operatorname{Var}(Y).$$

PROOF. We start by considering two independent simple functions $\varphi = \sum_j c_j \mathcal{X}_{A_j}$ and $\varphi' = \sum_{j'} c'_{j'} \mathcal{X}_{A'_{j'}}$. The independence implies that

$$P(A_j \cap A'_{j'}) = P(A_j)\,P(A'_{j'}).$$

So,

$$E(\varphi\varphi') = \sum_{j,j'} c_j c'_{j'} P(A_j \cap A'_{j'}) = E(\varphi)\,E(\varphi').$$

The claim follows from the application of the monotone convergence theorem to sequences of simple functions $\varphi_n \nearrow X$ and $\varphi'_n \nearrow Y$ which are independent.

Finally, it is simple to check that $\operatorname{Var}(X+Y) = \operatorname{Var}(X) + \operatorname{Var}(Y) + 2E(XY) - 2E(X)E(Y)$. So, by the previous relation we complete the proof. $\square$

PROPOSITION 6.9. *If X and Y are independent random variables and f and g are $\mathcal{B}$-measurable functions on $\mathbb{R}$, then*

(1) *$f(X)$ and $g(Y)$ are independent.*

(2) $E(f(X)g(Y)) = E(f(X))\,E(g(Y))$ *if* $E(|f(X)|), E(|g(Y)|) < +\infty$.

EXERCISE 6.10. Prove it.

EXAMPLE 6.11. Let $f(x) = x^2$ and $g(y) = e^y$. If X and Y are independent random variables, then X^2 and e^Y are also independent.

The random variables in a sequence $X_1, X_2, \dots$ are independent iff for any $n \in \mathbb{N}$ and $B_1, \dots, B_n \in \mathcal{B}$ we have

$$P(X_1 \in B_1, \dots, X_n \in B_n) = P(X_1 \in B_1) \cdots P(X_n \in B_n).$$

That is, the joint distribution of $(X_1, \dots, X_n)$ is equal to the product of the individual distributions for any $n \in \mathbb{N}$.

EXERCISE 6.12. Suppose that the random variables X and Y have only values in $\{0, 1\}$. Show that if $E(XY) = E(X)E(Y)$, then X, Y are independent.

Recall the definition of variance of a random variable X,

$$\operatorname{Var}(X) = E(X^2) - E(X)^2,$$

and of covariance between X and Y,

$$\operatorname{Cov}(X, Y) = E(XY) - E(X)E(Y).$$

Notice that if X, Y are independent, then they are uncorrelated since $\operatorname{Cov}(X, Y) = 0$.

EXERCISE 6.13. Construct an example of two uncorrelated random variables that are not independent.

EXERCISE 6.14. Show that if $\operatorname{Var}(X) \neq \operatorname{Var}(Y)$, then $X + Y$ and $X - Y$ are not independent.

Two σ-algebras $\mathcal{F}_1, \mathcal{F}_2$ are independent iff every $A_1 \in \mathcal{F}_1$ and $A_2 \in \mathcal{F}_2$ are independent.

EXERCISE 6.15. Show that two random variables X, Y are independent iff $\sigma(X)$ and $\sigma(Y)$ are independent.

2. Sums of random variables

Let $(\Omega, \mathcal{F}, P)$ be a probability space and X_1, X_2 random variables with distributions α_1, α_2, respectively. Consider the measurable function $f\colon \mathbb{R}^2 \to \mathbb{R}$ given by $f(x_1, x_2) = x_1 + x_2$, which is measurable with respect to the product Borel σ-algebra $\mathcal{G}$.

The *convolution of* α_1 *and* α_2 is defined to be the induced product measure on $\mathcal{B}$

$$\alpha_1 * \alpha_2 = (\alpha_1 \times \alpha_2) \circ f^{-1}.$$

In addition, $\alpha_1 * \alpha_2(\Omega) = 1$. So, the convolution is a distribution which turns out to be of the random variable $X_1 + X_2$.

PROPOSITION 6.16.

(1) *For every $A \in \mathcal{B}$,*
$$(\alpha_1 * \alpha_2)(A) = \int \alpha_1(A - x_2)\, d\alpha_2(x_2),$$
where $A - x_2 = \{y - x_2 \in \mathbb{R} \colon y \in A\}$.

(2) *The characteristic function of $\alpha_1 * \alpha_2$ is*
$$\phi_{\alpha_1 * \alpha_2} = \phi_{\alpha_1}\phi_{\alpha_2},$$
where ϕ_{α_i} is the characteristic function of α_i.

(3) $\alpha_1 * \alpha_2 = \alpha_2 * \alpha_1$.

PROOF. Writing $f_{x_2}(x_1) = f(x_1, x_2)$ and using Proposition 4.17 and the Fubini theorem we get:

(1)
$$\begin{aligned}(\alpha_1 * \alpha_2)(A) &= \int \mathcal{X}_A(y)\, d(\alpha_1 * \alpha_2)(y) \\ &= \int \mathcal{X}_A \circ f(x_1, x_2)\, d(\alpha_1 \times \alpha_2)(x_1, x_2) \\ &= \int\int \mathcal{X}_{f_{x_2}^{-1}(A)}(x_1)\, d\alpha_1(x_1) d\alpha_2(x_2) \\ &= \int \alpha_1(f_{x_2}^{-1}(A)) d\alpha_2(x_2) \\ &= \int \alpha_1(A - x_2) d\alpha_2(x_2).\end{aligned}$$

(2)
$$\begin{aligned}\phi_{\alpha_1 * \alpha_2}(t) &= \int e^{ity}\, d(\alpha_1 * \alpha_2)(y) \\ &= \int e^{itf(x_1,x_2)}\, d(\alpha_1 \times \alpha_2)(x_1, x_2) \\ &= \int e^{itx_1}\, d\alpha_1 \int e^{itx_2}\, d\alpha_2.\end{aligned}$$

(3) By the previous result, it follows from the fact that the characteristic functions are equal.

$\square$

PROPOSITION 6.17. *Let $X_1, \dots, X_n$ be independent random variables with distributions $\alpha_1, \dots, \alpha_n$, respectively, and*

$$S_n = \sum_{i=1}^{n} X_i.$$

Then,

(1) $\mu_n = \alpha_1 * \dots * \alpha_n$ *is the distribution of* S_n.
(2) $\phi_{\mu_n} = \phi_{\alpha_1} \dots \phi_{\alpha_n}$ *is the characteristic function of* μ_n.
(3) $E(S_n) = \sum_{i=1}^n E(X_i)$.
(4) $\mathrm{Var}(S_n) = \sum_{i=1}^n \mathrm{Var}(X_i)$.

EXERCISE 6.18. Prove it.

EXAMPLE 6.19. Let $\lambda \in]0,1[$. Consider a sequence of independent random variables $X_n \colon \Omega \to \{-\lambda^n, \lambda^n\}$ with Bernoulli distributions

$$\alpha_n = \frac{1}{2}(\delta_{-\lambda^n} + \delta_{\lambda^n}).$$

The characteristic function of α_n is

$$\phi_{X_n}(t) = \cos(t\lambda^n).$$

Write now

$$S_n = \sum_{i=1}^{n} X_i$$

whose distribution $\mu_n = \alpha_1 * \dots * \alpha_n$ is called Bernoulli convolution. It has characteristic function

$$\phi_{S_n}(t) = \prod_{i=1}^{n} \cos(t\lambda^i).$$

3. Law of large numbers

Let $(\Omega, \mathcal{F}, P)$ be a probability space and $X_1, X_2, \dots$ a sequence of random variables. We say that the sequence is *i.d.d.* if the random variables are independent and identically distributed. That is, all of the random variables are independent and share the same distribution $\alpha = P \circ X_n^{-1}$, $n \in \mathbb{N}$.

THEOREM 6.20 (Weak law of large numbers). *Let $X_1, X_2, \dots$ be an i.i.d. sequence of random variables. If $E(|X_1|) < +\infty$, then*

$$\frac{1}{n} \sum_{i=1}^{n} X_i \xrightarrow{P} E(X_1).$$

PROOF. Let ϕ be the characteristic function of the distribution of X_1 (it is the same for every X_n, $n \in \mathbb{N}$). Since $E(|X_1|) < +\infty$, we have $\phi'(0) = iE(X_1)$. So, the first order Taylor expansion of ϕ around 0 is

$$\phi(t) = 1 + iE(X_1)t + o(t),$$

where $|t| < r$ for some sufficiently small $r > 0$. For any fixed $t \in \mathbb{R}$ and n sufficiently large such that $|t|/n < r$ we have

$$\phi\left(\frac{t}{n}\right) = 1 + iE(X_1)\frac{t}{n} + o\left(\frac{t}{n}\right).$$

Thus, for those values of t and n, the random variable $M_n = \frac{1}{n}S_n$ has characteristic function

$$\phi_n(t) = \phi\left(\frac{t}{n}\right)^n = \left[1 + iE(X_1)\frac{t}{n} + o\left(\frac{t}{n}\right)\right]^n.$$

Finally, using the fact that $(1+a/n+o(1/n))^n \to e^a$ whenever $n \to +\infty$, we get

$$\lim_{n\to+\infty} \phi_n(t) = e^{iE(X_1)t}.$$

This is the characteristic function of the Dirac distribution at $E(X_1)$, corresponding to the constant random variable $E(X_1)$. Therefore,M_n converges in distribution to $E(X_1)$ and also in probability by Proposition 5.21. $\square$

REMARK 6.21. Notice that

$$\frac{1}{n}\sum_{i=1}^{n} X_i$$

is the average of the random variables $X_1, \dots, X_n$. So, the weak law of large numbers states that the average converges in probability to the expected value.

EXERCISE 6.22. Show that the weak law of large numbers does not hold for the Cauchy distribution.

The law of large numbers can be improved by getting a stronger convergence. For the theorem below we will use the following estimate.

EXERCISE 6.23. Show that for any $t, p > 0$ we have

$$P(X \geq t) \leq \frac{1}{t^p}E(|X|^p). \tag{6.1}$$

THEOREM 6.24 (Strong law of large numbers). *Let $X_1, X_2, \dots$ be an i.i.d. sequence of random variables. If $E(|X_1|^4) < +\infty$, then*

$$\frac{1}{n}\sum_{i=1}^{n} X_i \to E(X_1) \quad a.s.$$

PROOF. Suppose $E(X_1) = 0$ for simplicity (the general result is left as an exercise). So, writing $S_n = \sum_{i=1}^n X_i$ we can show by induction that $E(S_n^2) = nE(X_1^2)$ and

$$E(S_n^4) = nE(X_1^4) + 3n(n-1)E(X_1^2)^2.$$

Therefore, $E(S_n^4) \leq nE(|X_1|^4) + 3n^2\sigma^4$.

Now, for any $\delta > 0$, using (6.1),

$$P\left(\left|\frac{S_n}{n}\right| \geq \delta\right) = P(|S_n| \geq n\delta) \leq \frac{nE(|X_1|^4) + 3n^2\sigma^4}{n^4\delta^4}.$$

This implies that

$$\sum_{n\geq 1} P\left(\left|\frac{S_n}{n} \geq \delta\right|\right) < +\infty$$

and the claim follows from the first Borel-Cantelli lemma (Exercise 2.35), i.e. $S_n/n \to 0$ a.s. □

4. Central limit theorem

Let $(\Omega, \mathcal{F}, P)$ be a probability space.

THEOREM 6.25 (Central limit theorem). *Let $X_1, X_2, \dots$ be an i.i.d. sequence of random variables. If $E(X_1) = 0$ and $\sigma^2 = \mathrm{Var}(X_1) < +\infty$, then for every $x \in \mathbb{R}$,*

$$P\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n X_i \leq x\right) \to \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^x e^{-t^2/2\sigma^2}\,dt.$$

REMARK 6.26. The central limit theorem states that $S_n/\sqrt{n}$ converges in distribution to a random variable with the normal distribution.

PROOF. It is enough to show that the characteristic function ϕ_n of

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n X_i$$

converges to the characteristic function of the normal distribution.

Let ϕ be the characteristic function of X_n for any n. Its Taylor expansion of second order at 0 is

$$\phi(t) = 1 - \frac{1}{2}\sigma^2 t^2 + o(t^2),$$

with $|t| < r$ for some $r > 0$. So, for a fixed $t \in \mathbb{R}$ and n satisfying $|t|/\sqrt{n} < r$ (i.e. $n > t^2/r^2$),

$$\phi\left(\frac{t}{\sqrt{n}}\right) = 1 - \frac{1}{2}\sigma^2\frac{t^2}{n} + o\left(\frac{t^2}{n}\right).$$

Then,
$$\phi_n(t) = \phi\left(\frac{t}{\sqrt{n}}\right)^n = \left[1 - \frac{1}{2}\sigma^2\frac{t^2}{n} + o\left(\frac{t^2}{n}\right)\right]^n.$$
Taking the limit as $n \to +\infty$ we obtain
$$\phi_n(t) \to e^{-\sigma^2 t^2/2}.$$
□

Exercise 6.27. Write the statement of the central limit theorem for sequences of i.i.d. random variables X_n with mean μ. *Hint:* Apply the theorem to $X_n - \mu$ which has zero mean.

CHAPTER 7

Conditional expectation

In this chapter we introduce the concept of conditional expectation. It will be used in the construction of stochastic processes which are not sequences of i.i.d. random variables. We start by looking at an important result in the theory of the Lebesgue integral, the Radon-Nikodym theorem.

1. Radon-Nikodym theorem

Let $(\Omega, \mathcal{F}, \mu)$ be a measure space.

THEOREM 7.1. *Let f be an integrable function. Then,*

$$\nu(A) = \int_A f\,d\mu, \qquad A \in \mathcal{F},$$

defines a σ-additive function $\nu\colon \mathcal{F} \to \mathbb{R}$. Moreover, if $f \geq 0$ a.e. then ν is a measure and

$$\int_A g\,d\nu = \int_A g\,f\,d\mu \tag{7.1}$$

for any function g integrable with respect to ν and $A \in \mathcal{F}$.

REMARK 7.2. In the conditions of the above theorem we get $\nu(A) = \int_A d\nu = \int_A f\,d\mu$. It is therefore natural to use the notation

$$d\nu = f\,d\mu.$$

PROOF. Let $A_1, A_2, \dots \in \mathcal{F}$ be pairwise disjoint and $B = \bigcup_{i=1}^{+\infty} A_i$. Hence,

$$\nu\,(B) = \int_B f\,d\mu = \int f\mathcal{X}_B\,d\mu.$$

Define $g_n = f\mathcal{X}_{B_n}$ where $B_n = \bigcup_{i=1}^{n} A_i$. So, $g_n \leq f\mathcal{X}_B = \lim g_n$. By the monotone convergence theorem, $\int \lim g_n\,d\mu = \lim \int g_n\,d\mu$. That is,

$$\int f\mathcal{X}_B\,d\mu = \lim_{n\to+\infty} \int_{B_n} f\,d\mu = \sum_{i=1}^{+\infty} \int_{A_i} f\,d\mu = \sum_{i=1}^{+\infty} \nu(A_i)$$

where we have used $\int_{B_n} f\,d\mu = \sum_{i=1}^{n} \int_{A_i} f\,d\mu$ obtained by induction of the property in Proposition 4.6. Therefore, ν is σ-additive.

By Proposition 4.6, $\nu(\emptyset) = 0$ because $\mu(\emptyset) = 0$. As $f \geq 0$ a.e, we obtain $\nu(A) = \int_A f\,d\mu \geq \int_A 0\,d\mu = 0$, again using Proposition 4.6.

Finally, choose a sequence of simple functions $\varphi_n \nearrow g$ each written in the usual form

$$\varphi_n = \sum_{j=1}^{N} c_j \mathcal{X}_{A_j}.$$

Applying the monotone convergence theorem twice we obtain

$$\begin{aligned}
\int_A g\, d\nu &= \lim_{n\to+\infty} \int_A \varphi_n\, d\nu \\
&= \lim_{n\to+\infty} \sum_j c_j \nu(A_j \cap A) \\
&= \lim_{n\to+\infty} \sum_j c_j \int_A f \mathcal{X}_{A_j}\, d\mu \\
&= \lim_{n\to+\infty} \int_A f\varphi_n\, d\mu = \int_A gf\, d\mu.
\end{aligned}$$

□

EXAMPLE 7.3. Consider $f\colon \mathbb{R} \to \mathbb{R}$ given by

$$f(x) = \begin{cases} e^{-x}, & x \geq 0 \\ 0, & x < 0, \end{cases}$$

and the measure $\nu\colon \mathcal{B}(\mathbb{R}) \to \mathbb{R}$,

$$\nu(A) = \int_A f\, dm.$$

So, $m([n-1,n]) = 1$ for any $n \in \mathbb{N}$, and

$$\nu([n-1,n]) = \int_{n-1}^{n} e^{-t}\, dt = e^{-n}(e-1),$$

which goes to 1 as $n \to +\infty$. Moreover, if $g(x) = 1$, then g is not integrable with respect to m since we would have $\int g\, dm = m(\mathbb{R}) = +\infty$. However,

$$\int g\, d\nu = \int gf\, dm = \int_0^{+\infty} e^{-t}\, dt = 1.$$

Also, $\nu(\mathbb{R}) = 1$ and ν is a probability measure.

Let $(\Omega, \mathcal{F})$ be a measurable space. We say that a function $\lambda\colon \mathcal{F} \to \mathbb{R}$ is a *signed measure* iff λ is σ-additive. Any measure is obviously also a signed measure.

EXERCISE 7.4. Show that if λ is a signed measure and there is $A \in \mathcal{F}$ such that $\lambda(A)$ is finite, then $\lambda(\emptyset) = 0$.

EXAMPLE 7.5. If μ_1 and μ_2 are measures, then $\lambda = \mu_1 - \mu_2$ is a signed measure (but not a measure). In fact, for a sequence $A_1, A_2, \ldots$ of pairwise disjoint measurable sets,

$$\begin{aligned}\lambda\left(\bigcup_{n=1}^{+\infty} A_n\right) &= \mu_1\left(\bigcup_{n=1}^{+\infty} A_n\right) - \mu_2\left(\bigcup_{n=1}^{+\infty} A_n\right) \\ &= \sum_{n=1}^{+\infty} \mu_1(A_n) - \mu_2(A_n) \\ &= \sum_{n=1}^{+\infty} \lambda(A_n)\end{aligned}$$

by the σ-additivity of the measures.

A signed measure λ is *absolutely continuous* with respect to a measure μ iff for any set $A \in \mathcal{F}$ such that $\mu(A) = 0$ we also have $\lambda(A) = 0$. That is, every μ-null set is also λ-null. We use the notation

$$\lambda \ll \mu.$$

EXAMPLE 7.6. Consider the measurable space $(I, \mathcal{B})$ where I is an interval of $\mathbb{R}$ with positive length, the Dirac measure δ_a at some $a \in I$ and the Lebesgue measure m on I. For the set $A = \{a\}$ we have $m(A) = 0$ but $\delta_a(A) = 1$. So, δ_a is not absolutely continuous with respect to m. On the other hand, if $A = I \setminus \{a\}$ we get $\delta_a(A) = 0$ but $m(A) = m(I) > 0$. Hence, m is not absolutely continuous with respect to δ_a.

EXAMPLE 7.7. If μ is a measure and f is integrable, then

$$\lambda(A) = \int_A f \, d\mu$$

is a signed measure by Theorem 7.1. Moreover, if $\mu(A) = 0$ then the integral over A is always equal to zero. So, $\lambda \ll \mu$.

The above example is fundamental because of the next result.

THEOREM 7.8 (Radon-Nikodym). *Let λ be a signed measure and μ a measure both finite (i.e. $|\lambda(A)| < +\infty$ and $\mu(A) < +\infty$ for all $A \in \mathcal{F}$). If $\lambda \ll \mu$, then there is an integrable function f such that*

$$\lambda(A) = \int_A f \, d\mu, \qquad A \in \mathcal{F}.$$

Moreover, f is unique μ-a.e.

PROOF. Visit the library. □

REMARK 7.9.

(1) The function f in the above theorem is called the *Radon-Nikodym derivative* of λ with respect to μ and denoted by

$$\frac{d\lambda}{d\mu} = f.$$

Hence,

$$\lambda(A) = \int_A \frac{d\lambda}{d\mu}\, d\mu, \qquad A \in \mathcal{F}.$$

(2) If λ is also a measure then $\frac{d\lambda}{d\mu} \geq 0$ a.e.

EXAMPLE 7.10.

(1) Consider a countable set $\Omega = \{a_1, a_2, \dots\}$ and $\mathcal{F} = \mathcal{P}(\Omega)$. If μ is a finite measure such that all points in Ω have weight (i.e. $\mu(\{a_n\}) > 0$ for any $n \in \mathbb{N}$) and λ is any finite signed measure, then the only possible subset $A \subset \Omega$ with $\mu(A) = 0$ is the empty set $A = \emptyset$. So, $\lambda(A)$ is also equal to zero and $\lambda \ll \mu$. Since

$$\mu(A) = \sum_{a_n \in A} \mu(\{a_n\})$$

and

$$\lambda(\{a_n\}) = \int_{\{a_n\}} \frac{d\lambda}{d\mu}(x)\, d\mu(x) = \frac{d\lambda}{d\mu}(a_n)\mu(\{a_n\}),$$

we obtain

$$\frac{d\lambda}{d\mu}(a_n) = \frac{\lambda(\{a_n\})}{\mu(\{a_n\})}, \qquad n \in \mathbb{N}.$$

This defines the Radon-Nikodym derivative at μ-almost every point.

(2) Suppose that $\Omega = [0,1] \subset \mathbb{R}$ and $\mathcal{F} = \mathcal{B}([0,1])$. Take the Dirac measure δ_0 at 0, the Lebesgue measure m on $[0,1]$ and $\mu = \frac{1}{2}\delta_0 + \frac{1}{2}m$. If $\mu(A) = 0$ then $\frac{1}{2}\delta_0(A) + \frac{1}{2}m(A) = 0$ which implies that $\delta_0(A) = 0$ and $m(A) = 0$. Therefore, $\delta_0 \ll \mu$ and $m \ll \mu$. Notice that for any integrable function f we have

$$\int_A f\, d\mu = \frac{1}{2}\int_A f\, d\delta_0 + \frac{1}{2}\int_A f\, dm.$$

Therefore, for every $A \in \mathcal{F}$,

$$\delta_0(A) = \frac{1}{2}\int_A \frac{d\delta_0}{d\mu}\, d\delta_0 + \frac{1}{2}\int_A \frac{d\delta_0}{d\mu}\, dm$$

and also

$$m(A) = \frac{1}{2}\int_A \frac{dm}{d\mu}\, d\delta_0 + \frac{1}{2}\int_A \frac{dm}{d\mu}\, dm.$$

Aiming at finding the Radon-Nikodym derivatives, we first choose $A = \{0\}$ so that

$$\frac{d\delta_0}{d\mu}(0) = 2, \qquad \frac{dm}{d\mu}(0) = 0.$$

Moreover, let $A \in \mathcal{F}$ be such that $0 \notin A$. Thus,

$$\int_A \frac{d\delta_0}{d\mu}\, dm = 0, \qquad \int_A \frac{dm}{d\mu}\, dm = 2.$$

By considering the σ-algebra $\mathcal{F}'$ induced by $\mathcal{F}$ on $]0,1]$ we have that the Radon-Nikodym derivatives restricted to the measurable space $(]0,1], \mathcal{F}')$ are measurable and so the above equations imply that

$$\frac{d\delta_0}{d\mu}(x) = \begin{cases} 2, & x = 0 \\ 0, & \text{o.c.} \end{cases} \qquad \frac{dm}{d\mu}(x) = \begin{cases} 0, & x = 0 \\ 2, & \text{o.c.} \end{cases}$$

PROPOSITION 7.11. *Let ν, λ, μ be finite measures. If $\nu \ll \lambda$ and $\lambda \ll \mu$, then*

(1) $\nu \ll \mu$
(2)
$$\frac{d\nu}{d\mu} = \frac{d\nu}{d\lambda}\frac{d\lambda}{d\mu} \quad a.e.$$

PROOF.

(1) If $A \in \mathcal{F}$ is such that $\mu(A) = 0$ then $\lambda(A) = 0$ because $\lambda \ll \mu$. Furthermore, since $\nu \ll \lambda$ we also have $\nu(A) = 0$. This means that $\nu \ll \mu$.
(2) We know that
$$\lambda(A) = \int_A \frac{d\lambda}{d\mu}\, d\mu.$$
So,
$$\nu(A) = \int_A \frac{d\nu}{d\lambda}\, d\lambda = \int_A \frac{d\nu}{d\lambda}\frac{d\lambda}{d\mu}\, d\mu,$$
where we have used (7.1).

□

2. Conditional expectation

Let $(\Omega, \mathcal{F}, P)$ be a probability space and $X : \Omega \to \mathbb{R}$ a random variable (i.e. a $\mathcal{F}$-measurable function). Define the signed measure $\lambda : \mathcal{F} \to \mathbb{R}$ given by

$$\lambda(B) = \int_B X\, dP, \qquad B \in \mathcal{F}.$$

Recall that the Radon-Nikodym derivative is an $\mathcal{F}$-measurable function and it is of course given by

$$\frac{d\lambda}{dP} = X.$$

Consider now a σ-subalgebra $\mathcal{G} \subset \mathcal{F}$ and the restriction of λ to $\mathcal{G}$. That is, $\lambda_{\mathcal{G}} \colon \mathcal{G} \to \mathbb{R}$ such that

$$\lambda_{\mathcal{G}}(A) = \int_A X\, dP, \qquad A \in \mathcal{G}.$$

If the random variable X is not $\mathcal{G}$-measurable, it is not the Radon-Nikodym derivative of $\lambda_{\mathcal{G}}$. We define the *conditional expectation of X given $\mathcal{G}$* as

$$E(X|\mathcal{G}) = \frac{d\lambda_{\mathcal{G}}}{dP} \quad \text{a.s.}$$

which is an $\mathcal{G}$-measurable function. Therefore,

$$\lambda_{\mathcal{G}}(A) = \int_A E(X|\mathcal{G})\, dP = \int_A X\, dP, \quad A \in \mathcal{G}. \tag{7.2}$$

REMARK 7.12. The conditional expectation $E(X|\mathcal{G})$ is a random variable on the probability space $(\Omega, \mathcal{G}, P)$.

PROPOSITION 7.13. *Let X be a random variable and $\mathcal{G} \subset \mathcal{F}$ a σ-algebra.*

(1) *If X is $\mathcal{G}$-measurable, then $E(X|\mathcal{G}) = X$ a.s.* [1]
(2) $E(E(X|\mathcal{G})) = E(X)$.
(3) *If $X \geq 0$, then $E(X|\mathcal{G}) \geq 0$ a.s.*
(4) $E(|E(X|\mathcal{G})|) \leq E(|X|)$.
(5) $E(1|\mathcal{G}) = 1$ *a.s.*
(6) *For every $c_1, c_2 \in \mathbb{R}$ and random variables X_1, X_2,*
$$E(c_1 X_1 + c_2 X_2|\mathcal{G}) = c_1 E(X_1|\mathcal{G}) + c_2 E(X_2|\mathcal{G}).$$
(7) *If $h \colon \Omega \to \mathbb{R}$ is $\mathcal{G}$-measurable and bounded, then*
$$E(hX|\mathcal{G}) = hE(X|\mathcal{G}) \quad \textit{a.s.}$$
(8) *If $\mathcal{G}_1 \subset \mathcal{G}_2 \subset \mathcal{F}$ are σ-algebras, then*
$$E(E(X|\mathcal{G}_2)|\mathcal{G}_1) = E(X|\mathcal{G}_1).$$
(9) *If $\phi \colon \mathbb{R} \to \mathbb{R}$ is convex, then*
$$E(\phi \circ X|\mathcal{G}) \geq \phi \circ E(X|\mathcal{G}) \quad \textit{a.s.}$$

PROOF.

(1) If X is $\mathcal{G}$-measurable, then it is the Radon-Nikodym derivative of $\lambda_{\mathcal{G}}$ with respect to P.
(2) This follows from (7.2) with $A = \Omega$.

[1] In particular $E(X|\mathcal{F}) = X$.

(3) Consider the set $A = \{E(X|\mathcal{G}) < 0\}$ which is in $\mathcal{G}$ since $E(X|\mathcal{G})$ is $\mathcal{G}$-measurable. If $P(A) > 0$, then by (7.2),

$$0 \leq \int_A X\,dP = \int_A E(X|\mathcal{G})\,dP < 0,$$

which is false. So, $P(A) = 0$.

(4) Consider the set $A = \{E(X|\mathcal{G}) \geq 0\} \in \mathcal{G}$. Hence,

$$\begin{aligned} E(|E(X|\mathcal{G})|) &= \int_A E(X|\mathcal{G})\,dP - \int_{A^c} E(X|\mathcal{G})\,dP \\ &= \int_A X\,dP - \int_{A^c} X\,dP \\ &\leq \int_A |X|\,dP + \int_{A^c} |X|\,dP = E(|X|). \end{aligned}$$

(5) Since $X = 1$ is a constant it is $\mathcal{G}$-measurable. Therefore, $E(X|\mathcal{G}) = X$ a.s.

(6) Using the linearity of the integral and (7.2), for every $A \in \mathcal{G}$,

$$\begin{aligned} \int_A E(c_1X_1 + c_2X_2|\mathcal{G})\,dP &= c_1\int_A X_1\,dP + c_2\int_A X_2\,dP \\ &= \int_A (c_1E(X_1|\mathcal{G}) + c_2E(X_2|\mathcal{G}))\,dP. \end{aligned}$$

Since both integrand functions are $\mathcal{G}$-measurable, they agree a.s.

(7) Assume that $h \geq 0$ (the general case follows from the decomposition $h = h^+ - h^-$ with $h^+, h^- \geq 0$). Take a sequence of $\mathcal{G}$-measurable non-negative simple functions $\varphi_n \nearrow h$ of the form

$$\varphi_n = \sum_j c_j \mathcal{X}_{A_j},$$

where each $A_j \in \mathcal{G}$. We will show first that the claim holds for simple functions and later use the monotone convergence theorem to deduce it for h. For any $A \in \mathcal{G}$ we have that $A \cap A_j \in \mathcal{G}$. Hence

$$\begin{aligned} \int_A E(\varphi_n X|\mathcal{G})\,dP &= \int_A \varphi_n X\,dP \\ &= \sum_j c_j \int_{A\cap A_j} X\,dP \\ &= \sum_j c_j \int_{A\cap A_j} E(X|\mathcal{G})\,dP \\ &= \int_A \varphi_n E(X|\mathcal{G})\,dP. \end{aligned}$$

By the monotone convergence theorem applied twice,

$$\begin{aligned}\int_A E(hX|\mathcal{G})\,dP &= \int_A \lim \varphi_n X\,dP \\ &= \lim \int_A E(\varphi_n X|\mathcal{G})\,dP \\ &= \lim \int_A \varphi_n E(X|\mathcal{G})\,dP \\ &= \int_A hE(X|\mathcal{G})\,dP.\end{aligned}$$

(8) Let $A \in \mathcal{G}_1$. Then,

$$\begin{aligned}\int_A E(E(X|\mathcal{G}_2)|\mathcal{G}_1)\,dP &= \int_A E(X|\mathcal{G}_2)\,dP \\ &= \int_A X\,dP = \int_A E(X|\mathcal{G}_1)\,dP\end{aligned}$$

since A is also in $\mathcal{G}_2$.

(9) Do it as an exercise.

□

Remark 7.14. Whenever the σ-algebra is generated by the random variables $Y_1, \dots, Y_n$, we use the notation

$$E(X|Y_1, \dots, Y_n) = E(X|\sigma(Y_1, \dots, Y_n))$$

which reads as the conditional expectation of X given $Y_1, \dots, Y_n$.

Proposition 7.15. *Let $X, Y_1, \dots Y_n$ be independent random variables. Then,*

$$E(X|Y_1, \dots, Y_n) = E(X) \qquad a.s.$$

Exercise 7.16. Prove it. *Hint*: First do it for $\mathcal{X}_B$ for some B independent of $\sigma(Y_1, \dots, Y_n)$. Then, for simple functions that converge to X.

Example 7.17. Fixing some event $B \in \mathcal{F}$ notice that the σ-algebra generated by the random variable $\mathcal{X}_B$ is $\sigma(\mathcal{X}_B) = \{\emptyset, \Omega, B, B^c\}$. As $E(X|\mathcal{X}_B)$ is $\sigma(\mathcal{X}_B)$-measurable it is constant in B and in B^c:

$$E(X|\mathcal{X}_B)(x) = \begin{cases} a_1, & x \in B \\ a_2, & x \in B^c. \end{cases}$$

By (7.2) we obtain the conditions

$$\begin{aligned}a_1 P(B) + a_2 P(B^c) &= E(X) \\ a_1 P(B) &= \int_B X\,dP \\ a_2 P(B^c) &= \int_{B^c} X\,dP.\end{aligned}$$

So, if $0 < P(B) < 1$ we have

$$E(X|\mathcal{X}_B)(x) = \begin{cases} \frac{\int_B X\,dP}{P(B)}, & x \in B \\ \frac{\int_{B^c} X\,dP}{P(B^c)}, & x \in B^c \end{cases}$$
$$= \frac{\int_B X\,dP}{P(B)}\mathcal{X}_B + \frac{\int_{B^c} X\,dP}{P(B^c)}\mathcal{X}_{B^c}.$$

Finally, if $P(B) = 0$ of $P(B) = 1$ we have

$$E(X|\mathcal{X}_B)(x) = E(X) \quad \text{a.e.}$$

REMARK 7.18. In the case that $P(B) > 0$ we define the conditional expectation of X given the event B as the restriction of $E(X|\mathcal{X}_B)$ to B and use the notation

$$E(X|B) = E(X|\mathcal{X}_B)|_B = \frac{\int_B X\,dP}{P(B)}.$$

In particular, for the event $B = \{Y = y\}$ for some random variable Y and $y \in \mathbb{R}$ it is written as $E(X|Y = y)$.

EXERCISE 7.19. Let X be a random variable.

(1) Show that if $0 < P(B) < 1$ and $\alpha, \beta \in \mathbb{R}$, then
$$E(X|\alpha\mathcal{X}_B + \beta\mathcal{X}_{B^c}) = E(X|\mathcal{X}_B).$$
(2) Let $Y = \alpha_1\mathcal{X}_{B_1} + \alpha_2\mathcal{X}_{B_2}$ where $B_1 \cap B_2 = \emptyset$ and $\alpha_1 \neq \alpha_2$. Find $E(X|Y)$.

EXERCISE 7.20. Let $\Omega = \{1, 2, 3, 4, 5, 6\}$, $\mathcal{F} = \mathcal{P}(\Omega)$,

$$P(\{x\}) = \begin{cases} \frac{1}{16}, & x = 1, 2 \\ \frac{1}{4}, & x = 3, 4 \\ \frac{3}{16}, & x = 5, 6, \end{cases}$$

$$X(x) = \begin{cases} 2, & x = 1, 2 \\ 8, & x = 3, 4, 5, 6, \end{cases}$$

and $Y = 4\mathcal{X}_{\{1,2,3\}} + 6\mathcal{X}_{\{4,5,6\}}$. Find $E(X|Y)$.

3. Conditional probability

Let $(\Omega, \mathcal{F}, P)$ be a probability space and $\mathcal{G} \subset \mathcal{F}$ a σ-subalgebra. The conditional probability of an event $B \in \mathcal{F}$ given $\mathcal{G}$ is defined as the $\mathcal{G}$-measurable function

$$P(B|\mathcal{G}) = E(\mathcal{X}_B|\mathcal{G}).$$

REMARK 7.21. From the definition of conditional expectation, we obtain for any $A \in \mathcal{G}$ that

$$\int_A P(B|\mathcal{G})\,dP = \int_A E(\mathcal{X}_B|\mathcal{G})\,dP = \int_A \mathcal{X}_B\,dP = P(A \cap B).$$

THEOREM 7.22.

(1) *If* $P(B) = 0$, *then* $P(B|\mathcal{G}) = 0$ *a.e.*
(2) *If* $P(B) = 1$, *then* $P(B|\mathcal{G}) = 1$ *a.e.*
(3) $0 \leq P(B|\mathcal{G}) \leq 1$ *a.e. for any* $B \in \mathcal{F}$.
(4) *If* $B_1, B_2, \dots \in \mathcal{F}$ *are pairwise disjoint, then*

$$P\left(\bigcup_{n=1}^{+\infty} B_n|\mathcal{G}\right) = \sum_{n=1}^{+\infty} P(B_n|\mathcal{G}) \quad a.e.$$

(5) *If* $B \in \mathcal{A}$, *then* $P(B|\mathcal{G}) = \mathcal{X}_B$ *a.e.*

EXERCISE 7.23. Prove it.

REMARK 7.24.

(1) Similarly to the case of the conditional expectation, we define the conditional probability of B given random variables $Y_1, \dots, Y_n$ by

$$P(B|Y_1, \dots, Y_n) = P(B|\sigma(Y_1, \dots, Y_n)).$$

Moreover, given events $A, B \in \mathcal{F}$ with $P(A) > 0$, we define the conditional expectation of B given A as

$$P(B|A) = E(\mathcal{X}_B|A)|_A = \frac{P(A \cap B)}{P(A)}$$

which is a constant.

(2) Generally, without making any assumption on $P(A)$, the following formula is always true:

$$P(B|A)P(A) = P(A \cap B).$$

(3) Another widely used notation concerns events determined by random variables X and Y. When $B = \{X = x\}$ and $A = \{Y = y\}$ for some $x, y \in \mathbb{R}$, we write

$$P(X = x|Y = y)P(Y = y) = P(X = x, Y = y).$$

EXERCISE 7.25. Show that for $A, B \in \mathcal{F}$:

(1) Assuming that $P(A) > 0$, A and B are independent events iff $P(B|A) = P(B)$.
(2) $P(A|B)\,P(B) = P(B|A)\,P(A)$.
(3) For any sequence $C_1, C_2, \dots \in \mathcal{F}$ such that $P(\bigcup_n C_n) = 1$, we have

$$P(A|B) = \sum_n P(A \cap C_n|B). \tag{7.3}$$

and

$$P(A|B) = \sum_n P(A|C_n \cap B)P(C_n|B). \tag{7.4}$$

CHAPTER 8

Markov chains

1. The Markov property

Let $(\Omega, \mathcal{F}, P)$ be a probability space and $S \subset \mathbb{R}$ a countable set called the state space. For convenience we often choose S to be

$$S = \{1, 2, \ldots, N\}$$

where $N \in \mathbb{N} \cup \{+\infty\}$. We are considering both cases of S finite or infinite.

A stochastic process $X_0, X_1, \ldots$ is a *Markov chain on S* iff for all $n \geq 0$,

(1) $P(X_n \in S) = 1$,
(2) it satisfies the *Markov property*: for every $i_0, \ldots, i_n \in S$,

$$P(X_{n+1} = i_{n+1} | X_n = i_n, \ldots, X_0 = i_0) = P(X_{n+1} = i_{n+1} | X_n = i_n).$$

This means that the next future state (at time $n+1$) only depends on the present one (time n). The system does not have "memory" of the past.

We will see that the distributions of each $X_1, X_2, \ldots$ are determined by the knowledge of the above conditional probabilities (that control the evolution of the system) and the initial distribution of X_0.

Denote by

$$\pi_{i,j}^n = P(X_n = j | X_{n-1} = i)$$

the transition probability of moving from state i to state j at time $n \geq 1$. This defines the *transition probability matrix at time* n given by

$$T_n = \left[\pi_{i,j}^n\right]_{i,j \in S}.$$

Notice that T_n can be an infinite matrix if S is infinite.

Proposition 8.1. *The sum of the coefficients in each row of T_n equals 1.*

PROOF. The sum of the coefficients in the i-th row of T_n is

$$\begin{aligned}\sum_{j\in S}\pi_{i,j}^n &= \sum_{j\in S} P(X_n = j|X_{n-1}=i)\\ &= P\left(\bigcup_{j\in S}\{X_n=j\}|X_{n-1}=i\right)\\ &= P(X_n\in S|X_{n-1}=i)\\ &= 1\end{aligned}$$

because $P(X_n \in S) = 1$. □

A matrix M with dimension $r\times r$ ($r\in\mathbb{N}$ or $r=+\infty$) is called a *stochastic matrix* iff all its coefficients are non-negative and

$$M(1,1,\dots) = (1,1,\dots),$$

i.e. the sum of each row coefficients equals 1. Thus, the product of two stochastic matrices M_1 and M_2 is also a stochastic matrix since the coefficients are again non-negative and $M_1M_2(1,\dots,1) = M_1(1,\dots,1) = (1,\dots,1)$.

The matrices T_n are stochastic by Proposition 8.1. In fact, any sequence of stochastic matrices determines a Markov chain.

PROPOSITION 8.2. *Given a sequence $T_n = [\pi_{i,j}^n]_{i,j\in S}$ of stochastic matrices, any stochastic process X_n satisfying for every $n\geq 1$*

$$P(X_n = j|X_{n-1}=i) = \pi_{i,j}^n$$

is a Markov chain.

EXERCISE 8.3. Prove it.

Let X_n be a Markov chain. The distribution of each X_n, $n\geq 0$, given by $P\circ X_n^{-1}$, can be represented by a vector with dimension equal to $\#S$:

$$\alpha_n = (\alpha_{n,1},\alpha_{n,2},\dots) \quad\text{where}\quad \alpha_{n,j} = P(X_n=j),\quad j\in S.$$

We say that α_n is the distribution of X_n. Notice that

$$\alpha_n\cdot(1,1,\dots) = 1.$$

We can now determine the distribution of the chain at each time n from the initial distribution and the transition probabilities.

PROPOSITION 8.4. *If α_0 is the distribution of X_0, then*

$$\alpha_n = \alpha_0 T_1\dots T_n$$

is the distribution of X_n, $n\geq 1$.

PROOF. If $n = 1$ the formula states that

$$\begin{aligned}\alpha_{1,j} &= \sum_{k\in S} \alpha_{0,k}\pi^1_{k,j}\\ &= \sum_k P(X_0 = k)P(X_1 = j|X_0 = k)\\ &= \sum_k P(X_1 = j, X_0 = k)\\ &= P(X_1 = j)\end{aligned}$$

which is the distribution of X_1. We proceed by induction for $n \geq 2$ assuming that $\alpha_{n-1} = \alpha_0 T_1 \dots T_{n-1}$ is the distribution of X_{n-1}. So,

$$\begin{aligned}\alpha_{n,j} &= \sum_k \alpha_{n-1,k}\pi^n_{k,j}\\ &= \sum_k P(X_{n-1} = k)P(X_n = j|X_{n-1} = k)\\ &= \sum_k P(X_n = j, X_{n-1} = k)\\ &= P(X_n = j)\end{aligned}$$

that is the distribution of X_n. □

A vector $(i_0, i_1, \dots, i_n) \in S \times \cdots \times S$ defines a trajectory of states visited by the stochastic process up to time n. We are interested in computing its probability.

PROPOSITION 8.5. *If α_0 is the distribution of X_0, then for any $i_0, \dots, i_n \in S$ and $n \geq 1$,*

$$P(X_0 = i_0, \dots, X_n = i_n) = \alpha_{0,i_0}\, \pi^1_{i_0,i_1} \dots \pi^n_{i_{n-1},i_n}.$$

PROOF. Starting at $n = 1$ we have

$$\begin{aligned}P(X_0 = i_0, X_1 = i_1) &= P(X_0 = i_0)P(X_1 = i_1|X_0 = i_0)\\ &= \alpha_{0,i_0}\pi^1_{i_0,i_1}.\end{aligned}$$

By induction, for $n \geq 2$ and assuming that $P(X_0 = i_0, \dots, X_{n-1} = i_{n-1}) = \alpha_{0,i_0}\, \pi^1_{i_0,i_1} \dots \pi^{n-1}_{i_{n-2},i_{n-1}}$ we get

$$\begin{aligned}P(X_0 = i_0, \dots, X_n = i_n) =&P(X_0 = i_0, \dots, X_{n-1} = i_{n-1})\\ &P(X_n = i_n|X_0 = i_0, \dots, X_{n-1} = i_{n-1})\\ =&\alpha_{0,i_0}\, \pi^1_{i_0,i_1} \dots \pi^{n-1}_{i_{n-2},i_{n-1}} P(X_n = i_n|X_{n-1} = i_{n-1})\\ =&\alpha_{0,i_0}\, \pi^1_{i_0,i_1} \dots \pi^n_{i_{n-1},i_n},\end{aligned}$$

where we have used the Markov property. □

The probability of a trajectory given an initial state is now simple to obtain. It also follows the n-step transition probability.

Proposition 8.6. *If $P(X_0 = i) > 0$, then*

(1) $P(X_1 = i_1, \dots, X_n = i_n | X_0 = i) = \pi^1_{i,i_1} \dots \pi^n_{i_{n-1},i_n}$.
(2)
$$P(X_n = j | X_0 = i) = \pi^{(n)}_{i,j} \tag{8.1}$$
where $\pi^{(n)}_{i,j}$ is the (i,j)-coefficient of the product matrix $T_1 \dots T_n$.

Proof.

(1) It is enough to observe that
$$P(X_1 = i_1, \dots, X_n = i_n | X_0 = i) = \frac{P(X_0 = i, X_1 = i_1, \dots, X_n = i_n)}{P(X_0 = i)}$$
and use Proposition 8.5.
(2) Using the previous result and (7.3),
$$\begin{aligned} P(X_n = j | X_0 = i) &= \sum_{i_1,\dots,i_{n-1}} P(X_1 = i_1, \dots, X_{n-1} = i_{n-1}, X_n = j | X_0 = i) \\ &= \sum_{i_1,\dots,i_{n-1}} \pi^1_{i,i_1} \pi^2_{i_1,i_2} \dots \pi^{n-1}_{i_{n-2},i_{n-1}} \pi^n_{i_{n-1},j}. \end{aligned}$$
Now, notice that
$$\begin{aligned} \sum_{i_1} \pi^1_{i,i_1} \pi^2_{i_1,i_2} &= \sum_{i_1} P(X_2 = i_2 | X_1 = i_1, X_0 = i)\, P(X_1 = i_1 | X_0 = i) \\ &= P(X_2 = i_2 | X_0 = i), \end{aligned}$$
where we have used the fact that it is a Markov chain and (7.4). Moreover, using the same arguments
$$\begin{aligned} \sum_{i_2} P(X_2 = i_2 | X_0 = i) \pi^3_{i_2,i_3} &= \sum_{i_1} P(X_3 = i_3 | X_2 = i_2, X_0 = i)\, P(X_2 = i_2 | X_0 = i) \\ &= P(X_3 = i_3 | X_0 = i). \end{aligned}$$
Therefore, repeating the same ideia up to the sum in i_{n-1}, we finally prove the claim.

$\square$

Exercise 8.7. Given any increasing sequence of positive integers u_n, show that the sequence of (stochastic) product matrices
$$T_1 \dots T_{u_1}, T_{u_1+1} \dots T_{u_2}, T_{u_2+1} \dots T_{u_3}, \dots$$
corresponds to the transition matrices of the Markov chain
$$X_0, X_{u_1}, X_{u_2}, X_{u_3} \dots.$$

2. Homogeneous Markov chains

We are going to focus now on a special class of Markov chains, when the transition probabilities do not depend on the time n, i.e.

$$T_1 = T_2 = \cdots = T = [\pi_{i,j}]_{i,j \in S}.$$

These are called *homogeneous Markov chains.*

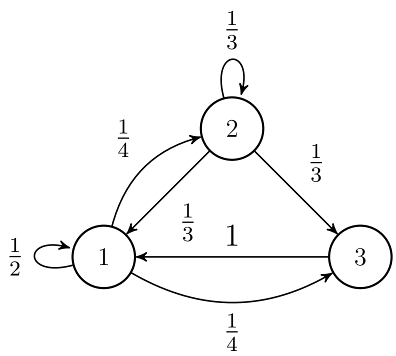
By Proposition 8.4 we have that the distribution of X_n is

$$\alpha_n = \alpha_0 T^n.$$

EXAMPLE 8.8. Let $S = \{1, 2, 3\}$ and

$$T = \begin{bmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ 1 & 0 & 0 \end{bmatrix}.$$

We can represent this stochastic process in graphical mode.



Moreover, starting with some distribution α_0 of X_0 we can get the distribution $\alpha_1 = \alpha_0 T$ of X_1 as

$$\begin{aligned} P(X_1 = 1) &= \sum_{j=1}^{3} \pi_{i,j}^{(1)} P(X_0 = j) \\ &= \frac{1}{2} P(X_0 = 1) + \frac{1}{3} P(X_0 = 2) + P(X_0 = 3) \\ P(X_1 = 2) &= \frac{1}{4} P(X_0 = 1) + \frac{1}{3} P(X_0 = 2) \\ P(X_1 = 3) &= \frac{1}{4} P(X_0 = 1) + \frac{1}{3} P(X_0 = 2). \end{aligned}$$

Similar relations can be obtained for the distribution α_n of X_n for any $n \geq 1$. In addition, given $X_0 = 1$ the probability of a trajectory $(1, 2, 3, 1)$ is

$$P(X_1 = 2, X_2 = 3, X_3 = 1 | X_0 = 1) = \frac{1}{12}.$$

Whenever $\alpha_{0,i} = P(X_0 = i) > 0$ by (8.1) we get

$$P(X_n = j | X_0 = i) = \pi_{i,j}^{(n)}$$

where $\pi_{i,j}^{(n)}$ is the (i,j)-coefficient of T^n. In fact, all the information about the evolution of the stochastic process is derived from the power matrix T^n called the n-step transition matrix. It is also a stochastic matrix. In particular, the sequence of random variables

$$X_0, X_n, X_{2n}, X_{3n}, \ldots$$

is also a Markov chain with transition matrix T^n and called the *n-step Markov chain.*

Notice that we can include the case $n = 0$, since

$$\pi_{i,j}^{(0)} = P(X_0 = j | X_0 = i) = \begin{cases} 1, & i = j \\ 0, & i \neq j. \end{cases}$$

This corresponds to the transition matrix $T^0 = I$ (the identity matrix).

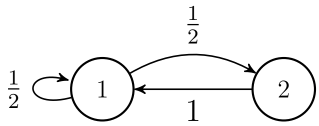
EXAMPLE 8.9. Let

$$T = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ 1 & 0 \end{bmatrix}.$$

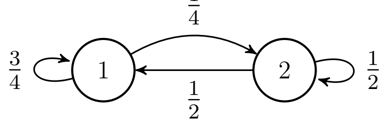
Then,

$$T^2 = \begin{bmatrix} \frac{3}{4} & \frac{1}{4} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix} \quad \text{and} \quad T^3 = \begin{bmatrix} \frac{5}{8} & \frac{3}{8} \\ \frac{3}{4} & \frac{1}{4} \end{bmatrix}.$$

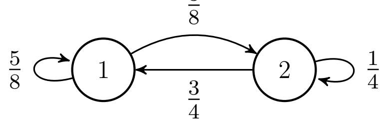
The Markov chain corresponding to T can be represented graphically as



Moreover, the two-step Markov chain given by T^2 looks like



Finally, the three-step Markov chain is



EXERCISE 8.10 (Bernoulli process). Let $S = \mathbb{N}$, $0 < p < 1$ and

$$P(X_{n+1} = i + 1 | X_n = i) = p$$
$$P(X_{n+1} = i | X_n = i) = 1 - p,$$

for every $n \geq 0$, $i \in S$. The random variable X_n could count the number of heads in n tosses of a coin if we set $P(X_0 = 0) = 1$. This is a homogeneous Markov chain with (infinite) transition matrix

$$T = \begin{bmatrix} 1-p & p & & & 0 \\ & 1-p & p & & \\ 0 & & \ddots & \ddots & \end{bmatrix}$$

i.e. for $i, j \in \mathbb{N}$

$$\pi_{i,j} = \begin{cases} 1-p, & i = j \\ p, & i + 1 = j \\ 0, & \text{o.c.} \end{cases}$$

Show that

$$P(X_n = j | X_0 = i) = C^n_{j-i} p^{j-i} (1-p)^{n-j+i}, \quad 0 \leq j - i \leq n.$$

3. Recurrence time

Consider a homogeneous Markov chain X_n on S. The *time of the first visit* to $i \in S$ (regardless of the initial state) is the random variable $t_i \colon \Omega \to \mathbb{N} \cup \{+\infty\}$ given by

$$t_i = \begin{cases} \min\{n \geq 1 \colon X_n = i\}, & \text{if there is } n \text{ such that } X_n = i, \\ +\infty, & \text{if for all } n \text{ we have } X_n \neq i. \end{cases}$$

EXERCISE 8.11. Show that t_i is a random variable.

EXERCISE 8.12. Show that

$$t_i = \sum_{n \geq 1} \mathcal{X}_{\{t_i \geq n\}}.$$

The distribution of t_i is then given by

$$P(t_i = n) = P(X_1 \neq i, \ldots, X_{n-1} \neq i, X_n = i), \quad n \in \mathbb{N},$$

and

$$P(t_i = +\infty) = P\left(X_1 \neq i, X_2 \neq i, \ldots\right).$$

The *mean recurrence time* τ_i of the state i is the expected value of t_i given $X_0 = i$,

$$\tau_i = E(t_i | X_0 = i).$$

Using the convention

$$+\infty.\, a = \begin{cases} 0, & a = 0 \\ +\infty, & a > 0, \end{cases}$$

we can write

$$\tau_i = \sum_{n=1}^{+\infty} nP(t_i = n|X_0 = i) + \infty.P(t_i = +\infty|X_0 = i).$$

Thus, $\tau_i \in [1, +\infty[\cup\{+\infty\}$. Notice also that if $P(X_0 = i) = 0$, then $\tau_i = E(t_i)$.

4. Classification of states

A state i is called *recurrent* iff

$$P(t_i = +\infty|X_0 = i) = 0.$$

This is also equivalent to

$$P(X_1 \neq i, X_2 \neq i, \dots |X_0 = i) = 0.$$

It means that the process returns to the initial state i with full probability. A state i which is not recurrent is said to be *transient*. So,

$$S = R \cup T,$$

where R is the set of recurrent states and T its complementary in S.

REMARK 8.13. Notice that

$$\begin{aligned}\{X_n = i \text{ for some } n \geq 1\} &= \bigcup_{n=1}^{+\infty} \{X_n = i\} \\ &= \left(\bigcap_{n=1}^{+\infty} \{X_n \neq i\}\right)^c.\end{aligned}$$

Hence, $i \in R$ iff

$$P(X_n = i \text{ for some } n \geq 1\ |X_0 = i) = 1.$$

PROPOSITION 8.14. *Let $i \in S$. Then,*

(1) *$i \in R$ iff*

$$\sum_{n=1}^{+\infty} \pi_{i,i}^{(n)} = +\infty.$$

(2) *If $i \in T$, then for any $j \in S$*

$$\sum_{n=1}^{+\infty} \pi_{j,i}^{(n)} < +\infty.$$

REMARK 8.15. Recall that if $\sum_n u_n < +\infty$, then $u_n \to 0$. On the other hand, there are sequences u_n that converge to 0 but the corresponding series does not converge. For example, $\sum_n 1/n = +\infty$.

PROOF.

(1)

($\Rightarrow$) If i is recurrent, then there is $m \geq 1$ such that

$$P(X_m = i|X_0 = i) > 0.$$

From (7.4) and the Markov property, for any $q > m$ we have

$$P(X_q = j|X_0 = i) = \sum_{k \in S} P(X_q = j|X_m = k)P(X_m = k|X_0 = i).$$

Next, we prove by induction that for $q = sm$ with $s \in \mathbb{N}$ and $j = i$ we have

$$P(X_{sm} = i|X_0 = i) \geq P(X_m = i|X_0 = i)^s.$$

This is clear for $s = 1$. Assuming that the relation above is true for s,

$$\begin{aligned} P(X_{(s+1)m} = i|X_0 = i) \geq & P(X_{(s+1)m} = i|X_m = i)P(X_m = i|X_0 = i) \\ = & P(X_{sm} = i|X_0 = i)P(X_m = i|X_0 = i) \\ \geq & P(X_m = i|X_0 = i)^{s+1}. \end{aligned}$$

Thus,

$$\begin{aligned} \sum_{n=1}^{+\infty} \pi_{i,i}^{(n)} &= \sum_{n=1}^{+\infty} P(X_n = i|X_0 = i) \\ &\geq \sum_{s=1}^{+\infty} P(X_m = i|X_0 = i)^2 = +\infty. \end{aligned}$$

($\Leftarrow$) Suppose now that $\sum_n \pi_{i,i}^{(n)} = +\infty$. Using (7.4), since

$$\sum_{k=1}^{+\infty} P(t_i = k) = 1,$$

we have

$$\begin{aligned} \pi_{j,i}^{(n)} &= P(X_n = i|X_0 = j) \\ &= \sum_{k=1}^{+\infty} P(X_n = i|t_i = k, X_0 = j)P(t_i = k|X_0 = j). \end{aligned}$$

The Markov property implies that

$$P(X_n = i|t_i = k, X_0 = j) = P(X_n = i|X_k = i) = \pi_{i,i}^{(n-k)}$$

for $0 \leq k \leq n$ and it vanishes for other values of k. Recall that $\{t_i = k\} = \{X_1 \neq 1, \dots, X_{k-1} \neq i, X_k = i\}$ and $\pi_{i,i}^{(0)} = 1$. Therefore,

$$\pi_{j,i}^{(n)} = \sum_{k=1}^{n} \pi_{i,i}^{(n-k)} P(t_i = k|X_0 = j). \tag{8.2}$$

For $N \geq 1$ and $j = i$, the following holds by resummation

$$\begin{aligned}
\sum_{n=1}^{N} \pi_{i,i}^{(n)} &= \sum_{n=1}^{N} \sum_{k=1}^{n} \pi_{i,i}^{(n-k)} P(t_i = k | X_0 = i) \\
&= \sum_{k=1}^{N} \sum_{n=k}^{N} \pi_{i,i}^{(n-k)} P(t_i = k | X_0 = i) \\
&= \sum_{k=1}^{N} P(t_i = k | X_0 = i) \sum_{n=0}^{N-k} \pi_{i,i}^{(n)} \\
&\leq \sum_{k=1}^{N} P(t_i = k | X_0 = i) \left(1 + \sum_{n=1}^{N} \pi_{i,i}^{(n)} \right).
\end{aligned}$$

Finally,

$$1 \geq \sum_{k=1}^{N} P(t_i = k | X_0 = i) \geq \frac{\sum_{n=1}^{N} \pi_{i,i}^{(n)}}{1 + \sum_{n=1}^{N} \pi_{i,i}^{(n)}} \to 1$$

as $N \to +\infty$, which implies

$$P(t_i < +\infty | X_0 = i) = \sum_{k=1}^{+\infty} P(t_i = k | X_0 = i) = 1.$$

That is, i is recurrent.

(2) Consider a transient state i, i.e. $\sum_n \pi_{i,i}^{(n)} < +\infty$. Using (8.2) we have by resummation

$$\begin{aligned}
\sum_{n=1}^{+\infty} \pi_{j,i}^{(n)} &= \sum_{n=1}^{+\infty} \sum_{k=1}^{n} \pi_{i,i}^{(n-k)} P(t_i = k | X_0 = j) \\
&= \sum_{k=1}^{+\infty} P(t_i = k | X_0 = j) \sum_{n=1}^{+\infty} \pi_{i,i}^{(n)} \\
&\leq \sum_{n=1}^{+\infty} \pi_{i,i}^{(n)} < +\infty.
\end{aligned}$$

□

EXAMPLE 8.16. Consider the Markov chain with two states and transition matrix

$$T = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

Thus,

$$T^n = \begin{cases} T, & n \text{ odd} \\ I, & n \text{ even.} \end{cases}$$

It is now easy to check for $i = 1, 2$ that

$$\sum_{n=1}^{+\infty} \pi_{i,i}^{(n)} = +\infty.$$

Both states are recurrent.

REMARK 8.17. If $i \in T$, i.e. $P(t_i = +\infty | X_0 = i) > 0$, then $\tau_i = +\infty$.

So, only recurrent states can have finite mean recurrence time. We will classify them accordingly.

A recurrent state i is *null* iff $\tau_i = +\infty$ ($\tau_i^{-1} = 0$). We write $i \in R_0$. Otherwise it is called *positive*, i.e. $\tau_i \geq 1$ and $i \in R_+$. Hence,

$$R = R_0 \cup R_+.$$

PROPOSITION 8.18. *Let $i \in S$. Then,*

(1) $\tau_i = +\infty$ *iff*
$$\lim_{n\to+\infty} \pi_{i,i}^{(n)} = 0.$$
(2) *If $\tau_i = +\infty$, then for any $j \in S$*
$$\lim_{n\to+\infty} \pi_{j,i}^{(n)} = 0.$$

EXERCISE 8.19. Prove it.

The *period of a state i* is given by

$$Per(i) = gcd\{n \geq 1 \colon \pi_{i,i}^{(n)} > 0\}$$

if this set is non-empty; otherwise the period is not defined. Here *gcd* stands for the greatest common divisor. Furthermore, i is *periodic* iff $Per(i) \geq 2$. It is called *aperiodic* iff $Per(i) = 1$.

REMARK 8.20.

(1) If n is not a multiple of $Per(i)$, then $\pi_{i,i}^{(n)} = 0$.
(2) If $\pi_{i,i} > 0$, then i is aperiodic.

Finally, a state i is said to be *ergodic* iff it is recurrent positive and aperiodic. We denote the set of ergodic states by E, which is a subset of R_+.

EXERCISE 8.21. Given a state $i \in S$, consider the function $N_i \colon \Omega \to \mathbb{N} \cup \{+\infty\}$ that counts the number of times the chain visits its starting point i:

$$N_i = \sum_{n=1}^{+\infty} \mathcal{X}_{X_n^{-1}(\{i\})}.$$

(1) Show that N_i is a random variable.

(2) Compute the distribution of N_i.
(3) What is $P(N_i = +\infty)$ if i is recurrent and if it is transient.

EXERCISE 8.22. * Consider a homogeneous Markov chain on the state space $S = \mathbb{N}$ given by

$$\begin{aligned} P(X_1 = i|X_0 = i) &= r, && i \geq 2, \\ P(X_1 = i-1|X_0 = i) &= 1-r, && i \geq 2, \\ P(X_1 = j|X_0 = 1) &= \frac{1}{2^j}, && j \geq 1. \end{aligned}$$

Classify the states of the chain and find their mean recurence times by computing the probability of first return after n steps, $P(t_i = n|X_0 = i)$ for $i \in S$.

EXERCISE 8.23. Show that

(1) $i \in R_+$ iff $\sum_n \pi_{i,i}^{(n)} = +\infty$ and $\lim_{n\to+\infty} \pi_{i,i}^{(n)} \neq 0$.
(2) $i \in R_0$ iff $\sum_n \pi_{i,i}^{(n)} = +\infty$ and $\lim_{n\to+\infty} \pi_{i,i}^{(n)} = 0$.
(3) $i \in T$ iff $\sum_n \pi_{i,i}^{(n)} < +\infty$ (in particular $\lim_{n\to+\infty} \pi_{i,i}^{(n)} = 0$).

Conclude that $\tau_i = +\infty$ iff $\lim_{n\to+\infty} \pi_{i,i}^{(n)} = 0$.

5. Decomposition of chains

Let $i, j \in S$. We write

$$i \to j$$

whenever there is $n \geq 0$ such that $\pi_{i,j}^{(n)} > 0$. That is, the probability of eventually moving from i to j is positive. Moreover, we use the notation

$$i \longleftrightarrow j$$

if $i \to j$ and $j \to i$ simultaneously.

EXERCISE 8.24. Consider $i \neq j$. Show that $i \to j$ is equivalent to

$$\sum_{n=1}^{+\infty} P(t_j = n|X_0 = i) > 0.$$

PROPOSITION 8.25. $\longleftrightarrow$ *is an equivalence relation on* S.

PROOF. Since $\pi_{i,i}^{(0)} = 1$, we always have $i \longleftrightarrow i$. Moreover, having $i \longleftrightarrow j$ is clearly equivalent to $j \longleftrightarrow i$. Finally, given any three states i, j, k such that $i \longleftrightarrow j$ and $j \longleftrightarrow k$, the probability of moving from i to k is positive because it is greater or equal than the product of the probabilities of moving from i to j and from j to k. In the same way we obtain that $k \to i$. So, $i \longleftrightarrow k$. $\square$

Denote the sets of all states that are equivalent to a given $i \in S$ by

$$[i] = \{j \in S \colon i \longleftrightarrow j\},$$

which is called the *equivalence class of* i. Of course, $[i] = [j]$ iff $i \longleftrightarrow j$. The equivalence classes are also known as *irreducible sets*.

THEOREM 8.26. *If $j \in [i]$, then*

1. *$Per(i) = Per(j)$.*
2. *i is recurrent iff j is recurrent.*
3. *i is null recurrent iff j is null recurrent.*
4. *i is positive recurrent iff j is positive recurrent.*
5. *i is ergodic iff j is ergodic.*

PROOF. We will just prove (2). The remaining cases are similar and left as an exercise.

Notice first that

$$\begin{aligned}
\pi_{i,i}^{(m+n+r)} &= \sum_k \pi_{i,k}^{(m+n)} \pi_{k,i}^{(r)} \\
&\geq \pi_{i,j}^{(m+n)} \pi_{j,i}^{(r)} \\
&= \sum_k \pi_{i,k}^{(m)} \pi_{k,j}^{(n)} \pi_{j,i}^{(r)} \\
&\geq \pi_{i,j}^{(m)} \pi_{j,j}^{(n)} \pi_{j,i}^{(r)}
\end{aligned}$$

Since $i \longleftrightarrow j$, there are $m, r \geq 0$ such that $\pi_{i,j}^{(m)} \pi_{j,i}^{(r)} > 0$. So,

$$\pi_{j,j}^{(n)} \leq \frac{\pi_{i,i}^{(m+n+r)}}{\pi_{i,j}^{(m)} \pi_{j,i}^{(r)}}.$$

This implies that

$$\begin{aligned}
\sum_{n=1}^{+\infty} \pi_{j,j}^{(n)} &\leq \frac{1}{\pi_{i,j}^{(m)} \pi_{j,i}^{(r)}} \sum_{n=1}^{+\infty} \pi_{i,i}^{(m+n+r)} \\
&\leq \frac{1}{\pi_{i,j}^{(m)} \pi_{j,i}^{(r)}} \sum_{n=1}^{+\infty} \pi_{i,i}^{(n)}
\end{aligned}$$

Therefore, if i is transient, i.e.

$$\sum_{n=1}^{+\infty} \pi_{i,i}^{(n)} < +\infty,$$

then j is also transient. $\square$

Consider a subset of the states $C \subset S$. We say that C is *closed* iff for every $i \in C$ and $j \notin C$ we have $\pi_{i,j}^{(1)} = 0$. This means that moving

out of C is an event of probability zero. It does not exclude outside states from moving inside C, i.e. we can have $\pi_{j,i}^{(1)} > 0$.

A closed set C made of only one state is called an *absorving state*.

PROPOSITION 8.27. *If $i \in R$, then $[i]$ is closed.*

PROOF. Suppose that $[i]$ is not closed. Then, there is some $j \notin [i]$ such that $\pi_{i,j}^{(1)} > 0$. That is, $i \to j$ but $j \not\to i$ (otherwise j would be in $[i]$). So,

$$P\left(\bigcap_{n\geq 1}\{X_n \neq i\}|X_0 = i\right) \geq P\left(\{X_1 = j\} \cap \bigcap_{n\geq 2}\{X_n \neq i\}|X_0 = i\right)$$
$$= P\left(X_1 = j|X_0 = i\right) = \pi_{i,j}^{(1)} > 0.$$

Taking the complementary set

$$P\left(\bigcup_{n\geq 1}\{X_n = i\}|X_0 = i\right) = 1 - P\left(\bigcap_{n\geq 1}\{X_n \neq i\}|X_0 = i\right) < 1.$$

This means that $i \in T$. □

The previous proposition implies the following decomposition of the state space.

THEOREM 8.28 (Decomposition). *Any state space S can be decomposed into the union of the set of transient states T and closed recurrent irreducible sets $C_1, C_2, \dots$:*

$$S = T \cup C_1 \cup C_2 \cup \dots.$$

REMARK 8.29.

(1) If $[i]$ is not closed, then $i \in T$.
(2) If X_0 is in C_k, then X_n stays in C_k forever with probability 1.
(3) If X_0 is in T, then X_n stays in T or moves eventually to one of the C_k's. If the state space is finite, it can not stay in T forever.

EXERCISE 8.30. If X_n is an irreducible Markov chain with period d, is $Y_n = X_{nd}$ also an irreducible Markov chain? If yes, what is the period of Y_n?

5.1. Finite closed sets.

PROPOSITION 8.31. *If $C \subset S$ is closed and finite, then*

$$C \cap R = C \cap R_+ \neq \emptyset.$$

Moreover, if C is a irreducible set, then $C \subset R_+$.

Proof. Suppose that all states are transient. Then, for any $i, j \in C$ we have $\pi_{j,i}^{(n)} \to 0$ as $n \to +\infty$ by Proposition 8.14. Moreover, for any $j \in C$ we have

$$\sum_{i \in C} \pi_{j,i}^{(n)} = 1.$$

So, for any $\varepsilon > 0$ there is $N \in \mathbb{N}$ such that for any $n \geq N$ we have $\pi_{j,i}^{(n)} < \varepsilon$. Therefore,

$$1 = \sum_{i \in C} \pi_{j,i}^{(n)} < \varepsilon \, \#C,$$

which implies for any ε that $\#C < 1/\varepsilon$. That is, C is infinite.

Assume now that there is $i \in R_0 \cap C$. So, by Proposition 8.18 we have for any $j \in C$ that $\pi_{j,i}^{(n)} \to 0$ as $n \to +\infty$. As before,

$$\sum_{j \in C} \pi_{j,i}^{(n)} = 1$$

and the limit of the left hand side is zero unless C is infinite.

Finally, if C is irreducible all its states have the same recurrence property. Since at least one is in R_+, then all are in R_+. □

Remark 8.32. The previous result implies that if $[i]$ is finite and closed, then $[i] \subset R_+$. In particular, if S is finite and irreducible (notice that it is always closed), then $S = R_+$.

Example 8.33. Consider the finite state space $S = \{1, 2, 3, 4, 5, 6\}$ and the transition probabilities matrix

$$T = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 \\ \frac{1}{4} & \frac{3}{4} & 0 & 0 & 0 & 0 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & 0 & 0 \\ \frac{1}{4} & 0 & \frac{1}{4} & \frac{1}{4} & 0 & \frac{1}{4} \\ 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix}.$$

It is simple to check that $1 \longleftrightarrow 2$, $3 \longleftrightarrow 4$ and $5 \longleftrightarrow 6$. We have that $[1] = \{1, 2\}$ and $[5] = \{5, 6\}$ are irreducible closed sets, while $[3] = \{3, 4\}$ is not closed. So, the states in $[1]$ and $[5]$ are positive recurrent and in $[3]$ are transient.

6. Stationary distributions

Consider a homogeneous Markov chain X_n on a state space S. Given an initial distribution α of X_0, we have seen that the distribution of X_n is given by $\alpha_n = \alpha T^n$, $n \in \mathbb{N}$. A special case is when the distribution stays the same for all times n, i.e. $\alpha_n = \alpha$. So, a distribution α on S is called *stationary* iff

$$\alpha T = \alpha.$$

EXAMPLE 8.34. Consider a Markov chain with $S = \mathbb{N}$ and for any $i \in S$

$$P(X_1 = 1|X_0 = i) = \frac{1}{2}, \qquad P(X_1 = i+1|X_0 = i) = \frac{1}{2}.$$

A stationary distribution has to satisfy

$$P(X_0 = i) = P(X_1 = i), \quad i \in S.$$

So,

$$P(X_0 = i) = \sum_j P(X_1 = i|X_0 = j)P(X_0 = j).$$

If $i = 1$, this implies that

$$P(X_0 = 1) = \frac{1}{2}\sum_j P(X_0 = j) = \frac{1}{2}.$$

If $i \geq 2$,

$$\begin{aligned} P(X_0 = i) &= P(X_1 = i|X_0 = i-1)P(X_0 = i-1) \\ &= \frac{1}{2}P(X_0 = i-1). \end{aligned}$$

So,

$$P(X_0 = i) = \frac{1}{2^i}.$$

In the case of a finite state space a stationary distribution α is a solution of the linear equation:

$$(T^\top - I)\alpha^\top = 0.$$

It can also be computed as an eigenvector of $T^\top$ (the transpose matrix of T) corresponding to the eigenvalue 1. Notice that it must satisfy $\alpha_i \geq 0$ and $\sum_i \alpha_i = 1$. Moreover, if T does not have an eigenvalue 1 (T and $T^\top$ share the same eigenvalues), then there are no stationary distributions.

EXAMPLE 8.35. Consider the Markov chain with transition matrix

$$T = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{4} & \frac{3}{4} \end{bmatrix}.$$

The eigenvalues of T are 1 and $\frac{1}{4}$. Furthermore, an eigenvector of $T^\top$ associated to the unit eigenvalue is $(1, 2) \in \mathbb{R}^2$. Therefore, the eigenvector which corresponds to a distribution is $\alpha = (\alpha_1, 2\alpha_1)$ with $\alpha_1 \geq 0$ and $3\alpha_1 = 1$. That is, $\alpha = (\frac{1}{3}, \frac{2}{3})$.

EXERCISE 8.36. Find a stationary distribution for the Markov chain in Example 8.33.

THEOREM 8.37. *Consider an irreducible S. Then, $S = R_+$ iff there is a unique stationary distribution, in which case it is given by $\alpha_i = \tau_i^{-1}$.*

The proof of the above theorem is contained in section 6.1.

REMARK 8.38. Recall that if S is finite and irreducible then $S = R_+$. So, in this case there is a unique stationary distribution.

EXERCISE 8.39. Find the unique stationary distribution for the Markov chain with transition matrix:

$$T = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix}.$$

6.1. Proof of Theorem 8.37. A measure μ on S is stationary iff $\mu T = \mu$. Notice that it is not required that μ is a probability measure as in the case of a stationary distribution (when $\mu(S) = 1$). In other words, a stationary measure is a generalization of a stationary distribution.

EXERCISE 8.40. Show that any stationary measure $\nu = (\nu_1, \dots)$ on an irreducible S verifies $0 < \nu_i < +\infty$ for every $i \in S$.

In the following we always assume S to be irreducible.

PROPOSITION 8.41. *If $S = T$ then there are no stationary measures.*

PROOF. As for any $i, j \in S = T$ we have $\sum_n \pi_{i,j}^{(n)} < +\infty$, thus $\pi_{i,j}^{(n)} \to 0$ as $n \to +\infty$. Therefore $\mu T^n \to 0$, implying that μT can not be equal to μ unless $\mu = 0$ which is not a measure. $\square$

PROPOSITION 8.42. *If $S = R$, then for each $i \in S$ the measure $\mu^{(i)}$ given by*

$$\mu_j^{(i)} := \mu^{(i)}(\{j\}) = \sum_{n \geq 1} P(X_n = j, t_i \geq n | X_0 = i), \quad j \in S,$$

is stationary. Moreover, $\mu^{(i)}(S) = \tau_i$.

PROOF. Fix $i \in S$ and let

$$N_j = \sum_{n \geq 1} \mathcal{X}_{\{X_n = j, t_i \geq n\}}$$

be the random variable that counts the number of visits to state j until time t_i. That is, the chain visits the state j for N_j times until it reaches i. Notice that $N_i = 1$.

The mean of N_j starting at $X_0 = i$ is

$$\rho_j = E(N_j | X_0 = i).$$

Clearly, $\rho_i = 1$. Considering the simple functions

$$\varphi_m = \sum_{n=1}^{m} P(X_n = j, t_i \geq n | X_0 = i)$$

so that $\varphi_m \nearrow N_j$ as $m \to +\infty$, we can use the monotone convergence theorem to get

$$\rho_j = \sum_{n \geq 1} P(X_n = j, t_i \geq n | X_0 = i).$$

Furthermore,

$$\begin{aligned}
\rho_j &= \pi_{i,j} + \sum_{n\geq 2} \sum_{k \neq i} P(X_n = j, X_{n-1} = k, t_i \geq n | X_0 = i) \\
&= \pi_{i,j} + \sum_{n\geq 2} \sum_{k \neq i} \pi_{k,j} P(X_{n-1} = k, t_i \geq n | X_0 = i) \\
&= \pi_{i,j} + \sum_{k \neq i} \pi_{k,j} \sum_{n \geq 1} P(X_n = k, t_i \geq n+1 | X_0 = i).
\end{aligned}$$

Notice that for $k \neq i$ we have

$$\begin{aligned}
\{X_n = k, t_i \geq n+1\} &= \{X_1 \neq i, \dots, X_{n-1} \neq i, X_n = k\} \\
&= \{X_n = k, t_k \geq n\}.
\end{aligned}$$

So, since $\rho_i = 1$,

$$\rho_j = \pi_{i,j}\rho_i + \sum_{k \neq i} \pi_{k,j}\rho_j = \sum_{k \in S} \pi_{k,j}\rho_j.$$

That is,

$$\rho = \rho T$$

where $\rho = (\rho_1, \rho_2, \dots)$. We therefore take $\mu^{(i)}(\{j\}) = \rho_j$.

The sum of all the N_j's is equal to t_i. Indeed,

$$\sum_{j \in S} N_j = \sum_{n \geq 1} \sum_{j \in S} \mathcal{X}_{\{X_n = j, t_i \geq n\}} = \sum_{n \geq 1} \mathcal{X}_{\{t_i \geq n\}} = t_i.$$

Again by the monotone convergence theorem,

$$\mu_i(S) = \sum_{j \in S} \rho_j = E\left(t_i | X_0 = i\right) = \tau_i.$$

□

Exercise 8.43. Show that $\mu_i^{(i)} = 1$.

Exercise 8.44. Show that

$$\mu_j^{(i)} = \pi_{i,j} + \sum_{n \geq 1} \sum_{k_1, \dots, k_n \neq i} \pi_{i,k_1}\pi_{k_1,k_2} \dots \pi_{k_{n-1},k_n}\pi_{k_n,j}.$$

Proposition 8.45. *If $S = R$ and $\nu = (\nu_1, \dots)$ a stationary measure, then for any $i \in S$ we have $\nu = \nu_i \mu^{(i)}$.*

PROOF. Given $j \in S$ there is n such that $\pi_{j,i}^{(n)} > 0$ by the irreducibility of S. Using also the stationarity property of the measures ($\nu T^n = \nu$ and $\mu^{(i)} T^n = \mu^{(i)}$),

$$\sum_{k \in S} \nu_k \pi_{k,i}^{(n)} = \nu_i \quad \text{and} \quad \sum_{k \in S} \mu_k^{(i)} \pi_{k,i}^{(n)} = \mu_i^{(i)} = 1.$$

So, from

$$0 = \sum_{k \in S} (\nu_k - \nu_i \mu_k^{(i)}) \pi_{k,i}^{(n)} \geq (\nu_j - \nu_i \mu_j^{(i)}) \pi_{j,i}^{(n)}$$

we obtain $\nu_j \leq \nu_i \mu_j^{(i)}$.

Now, for $j \in S$, again by the stationarity of ν,

$$\nu_j = \nu_i \pi_{i,j} + \sum_{k_1 \neq i} \nu_{k_1} \pi_{k_1,j}$$

Using the same relation for ν_{k_1} we obtain

$$\nu_j = \nu_i \pi_{i,j} + \nu_i \sum_{k_1 \neq i} \pi_{i,k_1} \pi_{k_1,j} + \sum_{k_1,k_2 \neq i} \nu_{k_2} \pi_{k_2,k_1} \pi_{k_1,j}.$$

Repeating this indefinitely, we get

$$\begin{aligned} \nu_i^{-1} \nu_j &\geq \pi_{i,j} + \sum_{n \geq 1} \sum_{k_1,\dots,k_n \neq i} \pi_{i,k_1} \pi_{k_1,k_2} \dots \pi_{k_{n-1},k_n} \pi_{k_n,j} \\ &= \mu_j^{(i)}. \end{aligned}$$

□

EXERCISE 8.46. Complete the proof of Theorem 8.37.

7. Limit distributions

Recall that the distribution of a Markov chain at time n is given by

$$P(X_n = j) = \sum_{i \in S} \pi_{i,j}^{(n)} P(X_0 = i), \quad j \in S.$$

We are now interested in determining the convergence in distribution of X_n.

THEOREM 8.47. *Let S be irreducible and aperiodic. Then,*

$$\lim_{n \to +\infty} \pi_{i,j}^{(n)} = \frac{1}{\tau_j}, \qquad i, j \in S,$$

and

$$\lim_{n \to +\infty} P(X_n = j) = \frac{1}{\tau_j}.$$

See the proof in section 7.1.

THEOREM 8.48. *Let S be irreducible and aperiodic.*

(1) *If $S = T$ or $S = R_0$, then X_n diverges in distribution.*
(2) *If $S = R_+ = E$, then X_n converges in distribution to the unique stationary distribution.*

PROOF. Recall that if S is transient or null recurrent, then $\tau_j = +\infty$ for all $j \in S$. So, $\pi_{i,j}^{(n)} \to 0$ for all $i, j \in S$. So, according to Theorem 8.47, $\lim_{n\to+\infty} P(X_n = j) = 0$ for all j. Hence, it can not be a distribution. In these cases there are no limit distributions.

If the $S = R_+$, its unique stationary distribution is

$$P(X_n = j) = \frac{1}{\tau_j}, \qquad n \geq 0, \quad j \in S.$$

Using again Theorem 8.47, the limit distribution is equal to the stationary distribution. $\square$

REMARK 8.49.

(1) When S is aperiodic and positive recurrent is said to be ergodic. That is why the previous theorem is usually called the ergodic theorem.
(2) Since the limit distribution for irreducible ergodic Markov chains does not depend on the initial distribution, we say that these chains forget their origins.

EXAMPLE 8.50. Consider the transition matrix

$$T = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix}.$$

The chain is irreducible and ergodic. Thus, there is a unique stationary distribution which is the limit distribution. From the fact that $T^n = T$ we have $\pi_{i,j}^{(n)} = \frac{1}{2} \to \frac{1}{2}$ we know that $\alpha_1 = \alpha_2 = \frac{1}{2}$ and $\tau_1 = \tau_2 = 2$.

EXAMPLE 8.51. Let now

$$T = \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix}.$$

The chain is irreducible and finite, hence $S = R_+$. The period is 2 and $\alpha = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ is the unique stationary distribution. On the other hand, $\pi_{1,2}^{(n)}$ is zero iff n is even. This implies for example that $\lim_{n\to+\infty} \pi_{1,2}^{(n)}$ does not exist. The aperiodicity condition imposed in Theorem 8.47 is therefore essential.

7.1. Proof of Theorem 8.47. Fix $j \in S$ and consider the successive visits to j given for each $k \geq 1$ by

$$R_k = \min\{n > R_{k-1} \colon X_n = j\} \quad \text{and} \quad R_0 = \min\{n \geq 0 \colon X_n = j\}.$$

Notice that $R_0 = 0$ means that $X_0 = j$. Otherwise, $R_0 = t_j$. Also, $X_n = j$ is equivalent to the existence of some $k \in \mathbb{N}$ satisfying $R_k = n$. Moreover, there is only a finite number of visits to j iff $R_k = \infty$ for some $k \in \mathbb{N}_0$.

So,

$$\pi_{j,j}^{(n)} = P\left(\bigcup_k \{R_k = n\} | R_0 = 0\right).$$

On the other hand, we have

$$\pi_{i,j}^{(n)} = \sum_{k=1}^{n} P(t_j = k | X_0 = i) \pi_{j,j}^{(n-k)}.$$

EXERCISE 8.52. Show that $\lim_{n\to+\infty} \pi_{i,j}^{(n)} = \lim_{n\to+\infty} \pi_{j,j}^{(n)}$

EXERCISE 8.53. Show that R_k is a homogeneous Markov chain with transition probabilities

$$f_n := P(R_1 = m + n | R_0 = m) = P(t_j = n | X_0 = j)$$

for any $n \in \mathbb{N} \cup \{\infty\}$, which is the same for every $m \in \mathbb{N}_0$.

Notice that

$$\tau_j = E(t_j | X_0 = j) = \sum_{n\geq 1} n f_n + \infty \cdot f_\infty.$$

Suppose that

$$f_\infty = P(R_1 = \infty | R_0 = m) = P(t_j = \infty | X_0 = j) > 0,$$

so that j is transient for the chain X_n with mean recurrence time $\tau_j = \infty$. Also,

$$\lim_{n\to+\infty} \pi_{j,j}^{(n)} = 0,$$

which means that it is equal to $1/\tau_j$.

It remains to be proved that for $f_\infty = 0$ we have

$$\lim_{n\to+\infty} \pi_{j,j}^{(n)} = \frac{1}{\tau_j}.$$

Take first the sup limit

$$a_0 = \limsup \pi_{j,j}^{(n)}.$$

By considering a subsequence for which the limit is a_0, we use the diagonalization argument to have a subsequence for which there exists

$$a_k = \lim \pi_{j,j}^{(k_n - k)} \leq a_0$$

for any $k \in \mathbb{N}$. Here we make the assumption that $\pi_{j,j}^{(k)} = 0$ for any $k \leq -1$.

Recall that

$$\pi_{j,j}^{(n)} = \sum_{m=1}^{n} f_m \pi_{j,j}^{(n-m)}.$$

Taking the limit along the sequence k_n, by the dominated convergence theorem, we get

$$a_0 = \sum_{m=1}^{+\infty} f_m a_m.$$

Since $\sum_k f_k = 1$ and $a_k \leq a_0$, then $a_k = a$ for every $k \in D = \{n \in \mathbb{N} \colon f_n > 0\}$. Similarly,

$$a_k = \lim \sum_{m=1}^{k_n - k} f_m \pi_{j,j}^{(k_n - k - m)} = \sum_{m=1}^{+\infty} f_m a_{k+m}$$

and $a_k = a$ for all $k \in D \oplus D = \{n_1 + n_2 \in \mathbb{N} \colon n_1, n_2 \in D\}$. Proceeding by induction and using the fact that the gcd of D is 1, we get that $a_k = a$ for k sufficiently large. This implies that $a_k = a$ for all k.

Exercise 8.54. Show that

$$\sum_{k=1}^{n} \sum_{i=k-1}^{+\infty} f_i \pi_{j,j}^{(n-k+1)} = \sum_{k=1}^{n} f_k.$$

Taking the limit along k_n the above equality becomes

$$\limsup \pi_{j,j}^{(n)} \sum_{k=1}^{+\infty} \sum_{i=k-1}^{+\infty} f_i = 1$$

Exercise 8.55. Show that

$$\sum_{k=1}^{+\infty} \sum_{i=k-1}^{+\infty} f_i = \tau_j.$$

So,

$$\limsup \pi_{i,j}^{(n)} = \frac{1}{\tau_j}.$$

The same idea can be used for the lim inf proving that the limit exists. This completes the proof of Theorem 8.47.

CHAPTER 9

Martingales

1. The martingale strategy

In the 18th century there was a popular strategy to guarantee a profit when gambling in a casino. We assume that the game is fair, for simplicity it is the tossing of a fair coin. Starting with an initial capital K_0 a gambler bets a given amount b. Winning the game means that the capital is now $K_1 = K_0 + b$ and there is already a profit. A loss implies that $K_1 = K_0 - b$. The martingale strategy consists in repeating the game until we get a win, while doubling the previous bet at each time. That is, if there is a first loss, at the second game we bet $2b$. If we win it then $K_2 = K_0 - b + 2b = K_0 + b$ and there is a profit. If it takes n games to obtain a win, then

$$K_n = K_{n-1} + 2^{n-1}b = K_0 - \sum_{i=1}^{n-1} 2^{i-1}b + 2^{n-1}b = K_0 + b$$

and there is a profit.

In conclusion, if we wait long enough until getting a win (and it is quite unlikely that one would obtain only losses in a reasonable fair game), then we will obtain a profit of b. It seems a great strategy, without risk. Why everybody is not doing it? What would happen if all players were doing it? What is the catch?

The problem with the strategy is that the capital K_0 is finite. If it takes too long to obtain a win (say n times), then

$$K_{n-1} = K_0 - (2^{n-1} - 1)b.$$

Bankruptcy occurs when $K_{n-1} \leq 0$, i.e. waiting n steps with

$$n \geq \log_2(K_0/b + 1) + 1.$$

For example, if we start with the Portuguese GDP in 2015[1]:

$$K_0 = €198\ 920\ 000\ 000$$

and choosing $b =$€1, then we can afford to loose 38 consecutive times.

On the other hand, if we assume that getting 10 straight losses in a row is definitely very rare and are willing to risk, then we need to assemble an initial capital of $K_0 =$€$511b$.

[1] cf. PORDATA http://www.pordata.pt/

We can formulate the probabilistic model in the following way. Consider τ to be the first time we get a win. We call it stopping time and denote it by

$$\tau = \min\{n \geq 1 \colon Y_n = 1\},$$

where the Y_n's are iid random variables with distribution

$$P(Y_n = 1) = \frac{1}{2} \quad \text{and} \quad P(Y_n = -1) = \frac{1}{2}$$

(the tossing of a coin). Notice that if $Y_n = -1$ for every $n \in \mathbb{N}$, then we set $\tau = +\infty$.

EXERCISE 9.1. Show that $\tau \colon \Omega \to \mathbb{N} \cup \{+\infty\}$ is a random variable.

At time n the capital K_n is thus the random variable

$$K_n = K_0 + b \sum_{i=1}^{\tau \wedge n} 2^{i-1} Y_i,$$

where

$$\tau \wedge n = \min\{\tau, n\}.$$

The probability of winning in finite time is

$$\begin{aligned} P(\tau < +\infty) &= P\left(\bigcup_{n=1}^{+\infty} \{Y_n = 1\}\right) \\ &= 1 - P\left(\bigcap_{n=1}^{+\infty} \{Y_n = -1\}\right) \\ &= 1 - \prod_{n=1}^{+\infty} P(Y_i = -1) \\ &= 1. \end{aligned}$$

So, with full probability the gambler eventually wins. Since

$$P(\tau = n) = P(Y_1 = -1, \dots, P_{n-1} = -1, Y_n = 1) = \frac{1}{2^n},$$

we easily determine the mean time of getting a win is

$$E(\tau) = \sum_{n \in \mathbb{N}} n \, P(\tau = n) = \sum_{n \in \mathbb{N}} \frac{n}{2^n} = 2.$$

So, on average it does not take too long to get a win. However, what matters to avoid ruin is the mean capital just before a win. Whilst

$E(K_\tau) = K_0 + b$, we have

$$\begin{aligned}
E(K_{\tau-1}) &= K_0 - E\left(b\sum_{i=1}^{\tau-1} 2^{i-1}\right) \\
&= K_0 - bE(2^{\tau-1} - 1) \\
&= K_0 - b\sum_{n=1}^{+\infty} P(\tau = n)(2^{n-1} - 1) \\
&= K_0 - b\sum_{n=1}^{+\infty} \frac{1}{2^n}(2^{n-1} - 1) \\
&= -\infty.
\end{aligned}$$

That is, the mean value for the capital just before winning is $-\infty$.

Notice also that $E(K_1) = K_0$. In general, for any n, since $K_{n+1} = K_n + 2^n bY_{n+1}$ and Y_{n+1} is independent of K_n (K_n is a sum involving only $Y_1, \dots, Y_n$ and the sequence Y_n is independent) we have

$$E(K_{n+1}|K_n) = K_n + 2^n bE(Y_{n+1}|K_n) = K_n.$$

The martingale strategy is therefore an example of a fair game in the sense that knowing your capital at time n, the best prediction of K_{n+1} is actually K_n. There is therefore no advantage but only risk.

2. General definition of a martingale

Let $(\Omega, \mathcal{F}, P)$ be a probability space. An increasing sequence of σ-subalgebras

$$\mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots \subset \mathcal{F}.$$

is called a *filtration*.

A stochastic process X_n is a *martingale with respect to a filtration* $\mathcal{F}_n$ if for every $n \in \mathbb{N}$ we have that

(1) X_n is $\mathcal{F}_n$-measurable (we say that X_n is adapted to the filtration $\mathcal{F}_n$)
(2) X_n is integrable (i.e. $E(|X_n|) < +\infty$).
(3) $E(X_{n+1}|\mathcal{F}_n) = X_n$, P-a.e.

Remark 9.2.

(1) Given a stochastic process X_n, the sequence of σ-algebras

$$\mathcal{F}_n = \sigma(X_1, \dots, X_n)$$

is a filtration and it is called the natural filtration. Notice that

$$\sigma(X_1, \dots, X_n) \subset \sigma(X_1, \dots, X_{n+1}).$$

(2) It is simple to check that the expected value of the random variables X_n in a martingale is constant:

$$E(X_n) = E(E(X_{n+1}|\mathcal{F}_n)) = E(X_{n+1}).$$

(3) Since X_n is $\mathcal{F}_n$-measurable we have that $E(X_n|\mathcal{F}_n) = X_n$. So, $E(X_{n+1}|\mathcal{F}_n) = X_n$ is equivalent to $E(X_{n+1} - X_n|\mathcal{F}_n) = 0$.

A sub-martingale is defined whenever

$$X_n \leq E(X_{n+1}|\mathcal{F}_n), \qquad P\text{-a.e.}$$

and a super-martingale requires that

$$X_n \geq E(X_{n+1}|\mathcal{F}_n), \qquad P\text{-a.e.}$$

So, $E(X_n)$ decreases for sub-martingales and it increases for super-martingales.

In some contexts a martingale is known as a fair game, a sub-martingale is a favourable game and a super-martingale is an unfair game. The interpretation of a martingale as a fair game (there is risk, there is no arbitrage) is very relevant in the application to finance.

3. Examples

Example 9.3. Consider a sequence of independent and integrable random variables Y_n, and the natural filtration $\mathcal{F}_n = \sigma(Y_1, \dots, Y_n)$.

(1) Let

$$X_n = \sum_{i=1}^{n} Y_i.$$

Then X_n is $\mathcal{F}_n$-measurable since any Y_i is $\mathcal{F}_i$-measurable and $\mathcal{F}_i \subset \mathcal{F}_n$ for $i \leq n$. Furthermore,

$$E(|X_n|) \leq \sum_{i=1}^{n} E(|Y_i|) < +\infty,$$

i.e. X_n is integrable. Since all the Y_n's are independent,

$$E(X_{n+1} - X_n|\mathcal{F}_n) = E(Y_{n+1}|\mathcal{F}_n) = E(Y_{n+1}).$$

Therefore, X_n is a martingale iff $E(Y_n) = 0$ for every $n \in \mathbb{N}$.

(2) Let

$$X_n = Y_1 Y_2 \dots Y_n.$$

It is also simple to check that X_n is $\mathcal{F}_n$-measurable for each $n \in \mathbb{N}$. In addition, because the Y_n's are independent,

$$E(|X_n|) = E(|Y_1|) \dots E(|Y_n|) < +\infty.$$

Now,

$$E(X_{n+1} - X_n|\mathcal{F}_n) = X_n E(Y_{n+1} - 1).$$

Thus, X_n is a martingale iff $E(Y_n) = 1$ for every $n \in \mathbb{N}$.

(3) Consider now the stochastic process

$$X_n = \left(\sum_{i=1}^{n} Y_n\right)^2$$

assuming that Y_n^2 is also integrable. Clearly X_n is $\mathcal{F}_n$-measurable for each $n \in \mathbb{N}$. It is also integrable since

$$E(|X_n|) \leq n \sum_{i=1}^{n} E(|Y_i|^2) < +\infty.$$

where we have use the Cauchy-Schwarz inequality. Finally,

$$E(X_{n+1} - X_n|\mathcal{F}_n) = E(2\sqrt{X_n}Y_{n+1}|\mathcal{F}_n) + E(Y_{n+1}^2) \geq 0.$$

So, X_n is a sub-martingale if $E(Y_n) = 0$ for every $n \in \mathbb{N}$.

Example 9.4 (Doob's process). Consider an integrable random variable Y and a filtration $\mathcal{F}_n$. Let

$$X_n = E(Y|\mathcal{F}_n).$$

By definition of the conditional expectation X_n is $\mathcal{F}_n$-measurable. It is also integrable since

$$E(|X_n|) = E(|E(Y|\mathcal{F}_n)|) \leq E(E(|Y|\,|\mathcal{F}_n)) = E(|Y|).$$

Finally,

$$E(X_{n+1} - X_n|\mathcal{F}_n) = E(E(Y|\mathcal{F}_{n+1}) - E(Y|\mathcal{F}_n)|\mathcal{F}_n) = E(Y - Y|\mathcal{F}_n) = 0.$$

That is, X_n is a martingale.

4. Stopping times

Let $(\Omega, \mathcal{F}, P)$ be a probability space and $\mathcal{F}_n$ a filtration. A function $\tau\colon \Omega \to \mathbb{N} \cup \{+\infty\}$ is a *stopping time* iff $\{\tau = n\} \in \mathcal{F}_n$ for every $n \in \mathbb{N}$.

Proposition 9.5. *The following propositions are equivalent:*

(1) $\{\tau = n\} \in \mathcal{F}_n$ *for every* $n \in \mathbb{N}$.
(2) $\{\tau \leq n\} \in \mathcal{F}_n$ *for every* $n \in \mathbb{N}$.
(3) $\{\tau > n\} \in \mathcal{F}_n$ *for every* $n \in \mathbb{N}$.

Exercise 9.6. Prove it.

Proposition 9.7. τ *is a random variable ($\mathcal{F}$-measurable).*

Exercise 9.8. Prove it.

EXAMPLE 9.9. Let X_n be a stochastic process, $B \in \mathcal{B}(\mathbb{R})$ and $\mathcal{F}_n$ is a filtration such that for each $n \in \mathbb{N}$ we have that X_n is $\mathcal{F}_n$-measurable. Consider the time of the first visit to B given by

$$\tau = \min\{n \in \mathbb{N}\colon X_n \in B\}.$$

Notice that $\tau = +\infty$ if $X_n \notin B$ for every $n \in \mathbb{N}$. Hence,

$$\begin{aligned}\{\tau = n\} &= \{X_1 \notin B, \dots, X_{n-1} \notin B, X_n \in B\} \\ &= \{X_n \in B\} \cap \bigcap_{i=1}^{+\infty} \{X_i \in B^c\} \in \mathcal{F}_n.\end{aligned}$$

That is, τ is a stopping time.

PROPOSITION 9.10.

$$E(\tau) = \sum_{n=1}^{+\infty} P(\tau \geq n).$$

EXERCISE 9.11. Prove it.

5. Stochastic processes with stopping times

Let X_n be a stochastic process and $\mathcal{F}_n$ is a filtration such that for each $n \in \mathbb{N}$ we have that X_n is $\mathcal{F}_n$-measurable. Given a stopping time τ with respect to $\mathcal{F}_n$, we define the sequence X_n stopped at τ by

$$Z_n = X_{\tau \wedge n},$$

where $\tau \wedge n = \min\{\tau, n\}$.

EXERCISE 9.12. Show that

$$Z_n = \sum_{i=1}^{n-1} X_i \mathcal{X}_{\{\tau = i\}} + X_n \mathcal{X}_{\{\tau \geq n\}}.$$

PROPOSITION 9.13.

$$E(Z_{n+1} - Z_n | \mathcal{F}_n) = E(X_{n+1} - X_n | \mathcal{F}_n) \mathcal{X}_{\{\tau \geq n+1\}}.$$

EXERCISE 9.14. Prove it.

REMARK 9.15. From the above result we can conclude that:

(1) If X_n is a martingale, then Z_n is also a martingale.
(2) If X_n is a submartingale, then Z_n is also a submartingale.
(3) If X_n is a supermartingale, then Z_n is also a supermartingale.

Consider now the term in the sequence X_n corresponding to the stopping time τ,

$$X_\tau = \sum_{i=1}^{+\infty} X_n \mathcal{X}_{\{\tau = n\}}.$$

Clearly, it is a random variable.

THEOREM 9.16 (Optional stopping). *Let X_n be a martingale. If*

(1) $P(\tau < +\infty) = 1$
(2) *there is an integrable function $g \geq 0$ such that $|X_{\tau\wedge n}| \leq g$ for every $n \in \mathbb{N}$ (domination),*

then $E(X_\tau) = E(X_1)$.

PROOF. Since $P(\tau < +\infty) = 1$ we have that $\lim_{n\to+\infty} X_{\tau\wedge n} = X_\tau$ P-a.e. Hence, by the dominated convergence theorem using the domination by g,

$$E(X_\tau) = E\left(\lim_{n\to+\infty} X_{\tau\wedge n}\right) = \lim_{n\to+\infty} E(X_{\tau\wedge n}) = E(X_{\tau\wedge 1}) = E(X_1),$$

where we have used the fact that $X_{\tau\wedge n}$ is also a martingale. □

The domination condition that is required in the optional stopping theorem above is implied by other conditions that might be simpler to obtain.

PROPOSITION 9.17. *If any of the following holds:*

(1) *there is $k \in \mathbb{N}$ such that $P(\tau \leq k) = 1$*
(2) *$E(\tau) < +\infty$ and there is $M > 0$ such that for any $n \in \mathbb{N}$*

$$E(|X_{n+1} - X_n||\mathcal{F}_n) \leq M,$$

then $X_{\tau\wedge n}$ is dominated.

EXERCISE 9.18. Prove it.

A related result (not requiring to have a martingale) is the following.

THEOREM 9.19 (Wald's equation). *Let Y_n be a sequence of integrable iid random variables, $X_n = \sum_{i=1}^n Y_i$, $\mathcal{F}_n = \sigma(Y_1, \dots, Y_n)$ and τ a stopping time with respect to $\mathcal{F}_n$. If $E(\tau) < +\infty$, then*

$$E(X_\tau) = E(X_1)\, E(\tau).$$

PROOF. Recall that

$$X_\tau = \sum_{n=1}^{+\infty} Y_n \mathcal{X}_{\{\tau\geq n\}}.$$

So,

$$E(X_\tau) = \sum_{n=1}^{+\infty} E(Y_n \mathcal{X}_{\{\tau\geq n\}}).$$

Recall that $\mathcal{X}_{\{\tau\geq n\}} = 1 - \mathcal{X}_{\{\tau\leq n-1\}}$, which implies that

$$\sigma(\mathcal{X}_{\{\tau\geq n\}}) \subset \mathcal{F}_{n-1}.$$

Since $\mathcal{F}_{n-1}$ and $\sigma(Y_n)$ are independent, it follows that Y_n and $\mathcal{X}_{\{\tau \geq n\}}$ are independent random variables.

Finally,

$$E(X_\tau) = E(Y_1) \sum_{n=1}^{+\infty} P(\tau \geq n) = E(X_1)\, E(\tau).$$

$\square$

APPENDIX A

Things that you should know before starting

1. Notions of mathematical logic

1.1. Propositions. A *proposition* is a statement that can be qualified either as true (T) or else as false (F) – there is no third way.

EXAMPLE A.1.

(1) $p =$“Portugal is bordered by the Atlantic ocean” (T)
(2) $q =$“zero is an integer number” (T)
(3) $r =$“Sevilla is the capital city of Spain” (F)

REMARK A.2. There are statements that can not be qualified as true or false. For instance, “This sentence is false”. If it is false, then it is true (contradiction). On the other hand, if it is true, then it is false (again contradiction). This kind of statements are not considered to be propositions since they leads us to contradiction (simultaneously true and false). Therefore, they will not be the object of our study.

The goal of the mathematical logic is to relate propositions through their logical symbols: T or F. We are specially interested in those that are T.

1.2. Operations between propositions. Let p and q be propositions. We define the following operations between propositions. The result is still a proposition.

- $\sim p$, not p (p is not satisfied).
- $p \wedge q$, p and q (both propositions are satisfied).
- $p \vee q$, p or q (at least one of the propositions is satisfied).
- $p \Rightarrow q$, p implies q (if p is satisfied, then q is also satisfied).
- $p \Leftrightarrow q$, p is equivalent to q (p is satisfied iff q is satisfied).

EXAMPLE A.3. Using the propositions p, q and r in Example A.1,

(1) $\sim p =$“Portugal is not bordered by the Atlantic Ocean” (F)
(2) $p \wedge q =$“Portugal is bordered by the Atlantic Ocean and zero is an integer number” (T)
(3) $p \vee r =$“Portugal is bordered by the Atlantic Ocean or Sevilla is the capital city of Spain” (T)

EXAMPLE A.4.

(1) "If Portugal is bordered by the Atlantic Ocean, then Portugal is bordered by the sea" (T)
(2) "$x = 0$ iff $|x| = 0$" (T)

The logic value of the proposition obtained by operations between propositions is given by the following table:

p	q	$\sim p$	$p \wedge q$	$p \vee q$	$p \Rightarrow q$	$\sim q \Rightarrow \sim p$	$\sim p \vee q$	$p \Leftrightarrow q$	$(\sim p \wedge \sim q) \vee (p \wedge q)$
T	T	F	T	T	T	T	T	T	T
T	F	F	F	T	F	F	F	F	F
F	T	T	F	T	T	T	T	F	F
F	F	T	F	F	T	T	T	T	T

EXERCISE A.5. Show that the following propositions are true

(1) $\sim (\sim p) \Leftrightarrow p$
(2) $(p \Rightarrow q) \Leftrightarrow (\sim q \Rightarrow \sim p)$
(3) $\sim (p \wedge q) \Leftrightarrow (\sim p) \vee (\sim q)$
(4) $\sim (p \vee q) \Leftrightarrow (\sim p) \wedge (\sim q)$
(5) $((p \Rightarrow q) \wedge (q \Rightarrow p)) \Leftrightarrow (p \Leftrightarrow q)$
(6) $p \wedge (q \vee r) \Leftrightarrow ((p \wedge q) \vee (p \wedge r))$
(7) $p \vee (q \wedge r) \Leftrightarrow ((p \vee q) \wedge (p \vee r))$
(8) $(p \Leftrightarrow q) \Leftrightarrow ((\sim p \wedge \sim q) \vee (p \wedge q))$

EXAMPLE A.6. Consider the following propositions:

- $p =$"Men are mortal"
- $q =$"Dogs live less than men"
- $r =$"Dogs are not imortal"

So, the relation $((p \wedge q) \Rightarrow r) \Leftrightarrow (\sim r \Rightarrow (\sim p \vee \sim q))$ can be read as:

> Saying that "if men are mortal and dogs live less than men, then dogs are mortal ", is the same as saying that "if dogs are imortal, then men are imortal or dogs live more than men".

1.3. Symbols. In the mathematical writing it is used frequently the following symbols:

- $\forall$ for all.
- $\exists$ there is.
- $:$ such that.
- , usually means "and".

EXAMPLE A.7.

(1) $\forall_{x\geq 0}\exists_{y\geq 1}\colon x+y\geq 1$. "For any x non-negative there is y greater or equal to 1 such that $x+y$ is greater or equal than 1". (T)
(2) $\forall y$ multiple of 4 $\exists x\geq 0\colon -\frac{1}{2}<x+y<\frac{1}{2}$. "For any y multiple of 4 there is $x\geq 0$ such that $x+y$ is strictly between $-\frac{1}{2}$ and $\frac{1}{2}$". (F)

We can apply the $\sim$ operator

$$\sim\exists_x p(x)\Leftrightarrow\forall_x\sim p(x)$$

where p is a proposition that depends on x.

1.4. Mathematical induction. Let $p(n)$ be a proposition that depends on a number n that can be $1,2,3,\dots$. We want to show that $p(n)$ is T for any n. The *mathematical induction principle* is a method that allows to prove for any such n in just two steps:

(1) Show that $p(1)$ is T.
(2) Suppose that $p(m)$ is T for a fixed m, then show that the next proposition $p(m+1)$ is also T.

This method works because if it is T for $n=1$ and for the consecutive propostion of any that it is T, then is T for $n=2,3,\dots$.

Example A.8. Consider the propositions $p(n)$ given for each n by

$$1+2+\cdots+n=\frac{(n+1)n}{2}.$$

For $n=1$, we have that $p(1)$ reduces simply to $1=1$ that is clearly T. Suppose now that $p(m)$ for a fixed m. I.e. assume that $1+2+\cdots+m=\frac{(m+1)m}{2}$. Thus,

$$1+\cdots+m+(m+1)=\frac{(m+1)m}{2}+(m+1)=\frac{(m+1)(m+2)}{2}.$$

That is, we have just showed that $p(m+1)$ is T. Therefore, $\forall_n p(n)$ is T.

This is one of the more popular methods in all sub-areas of mathematics, in computacional sciences, in economics, in finance and most sciences that use quantitative methods. A professional mathematician has the obligation to master it.

1.5. Mathematical writing and proofs. What really distinguishes Mathematics from any other science are 'proofs'. These are logical constructions to show a proposition beyond any doubt. This reliability is unique to Mathematics.

The mathematical literature is in general based on presenting definitions and then demonstrating propositions yielding several consequences and properties that can be useful. A proposition is called

either a Lemma, a Proposition or a Theorem by increasing order of importance (but in many cases it mostly depends on the subjective choice of the writer). A Corollary is a simple consequence of a Theorem. A Conjecture is just a guess that can turn into a proposition if proved to be true.

2. Set theory notions

2.1. Sets. A *set* is a collection of elements represented in the form:

$$\Omega = \{a, b, c, \dots\}$$

where $a, b, c, \dots$ are the elements of Ω. A subset A of Ω is a set whose elements are also in Ω, and we write $A \subset \Omega$. We also write $\Omega \supset A$ to mean the same thing.

Instead of naming all the elements of a set (an often impossible task), sometimes it is necessary to define a set through a given property that we want satisfied. So, we also use the following representation for a set:

$$\Omega = \{x \colon p(x)\}$$

where $p(x)$ is a proposition that depend on x. This can be read as "Ω is the set of all x such that $p(x)$ holds".

We write

$$a \in A$$

to mean that a is an element of A (a is in A). Sometimes it is convenient to write instead $A \ni a$. If a is not in A we write $a \notin A$.

EXAMPLE A.9.

(1) $1 \in \{1\}$
(2) $\{1\} \notin \{1\}$
(3) $\{1\} \in \{\{1\}, \{1, 2\}, \{1, 2, 3\}\}$.

A subset A of Ω corresponding to all the elements x of Ω that satisfy a proposition $q(x)$ is denoted by

$$A = \{x \in \Omega \colon q(x)\}.$$

If a set has a finite number of elements it is called finite. Otherwise, it is an infinite set. The set with zero elements is called the empty set and it is denoted by $\{\}$ or $\emptyset$.

EXAMPLE A.10.

(1) $A = \{0, 1, 2, \dots, 9\}$ is finite (it has 10 elements).
(2) The set of natural numbers

$$\mathbb{N} = \{1, 2, 3, \dots\}$$

is infinite.

(3) The set of integers
$$\mathbb{Z} = \{\dots, -2, -1, 0, 1, 2, \dots\}$$
is infinite.

(4) The set of rational numbers (ratio between integers)
$$\mathbb{Q} = \left\{\frac{p}{q} \colon p \in \mathbb{Z}, q \in \mathbb{N}\right\}$$
is infinite.

(5) The set of real numbers $\mathbb{R}$ consists of numbers of the form
$$a_0.a_1a_2a_2\dots$$
where $a_0 \in \mathbb{Z}$ and $a_i \in \{0, 1, 2, \dots, 9\}$ for any $i \in \mathbb{N}$, is also infinite.

2.2. Relation between sets. Let A and B be any two sets.

- $A = B$ (A equals B) iff $(x \in A \Leftrightarrow x \in B) \vee (A = \emptyset \wedge B = \emptyset)$.
- $A \subset B$ (A is contained in B) iff $(x \in A \Rightarrow x \in B) \vee A = \emptyset)$.

Example A.11. $\mathbb{N} \subset \mathbb{Z} \subset \mathbb{Q} \subset \mathbb{R}$.

Properties A.12.

(1) $\emptyset \subset A$
(2) $(A = B \wedge B = C) \Rightarrow A = C$
(3) $A \subset A$
(4) $(A \subset B \wedge B \subset A) \Rightarrow A = B$
(5) $(A \subset B \wedge B \subset C) \Rightarrow A \subset C$

Remark A.13. If
$$A = \{x \colon p(x)\} \quad \text{and} \quad B = \{x \colon q(x)\}, \tag{A.1}$$
then
$$A = B \Leftrightarrow \forall_x (p(x) \Leftrightarrow q(x)) \quad \text{and} \quad A \subset B \Leftrightarrow \forall_x (p(x) \Rightarrow q(x)).$$

2.3. Operations between sets. Let $A, B \subset \Omega$.

- $A \cap B = \{x \colon x \in A \wedge x \in B\}$ is the *intersection* between A and B.
- $A \cup B = \{x \colon x \in A \vee x \in B\}$ is the *union* between A and B.

Representing the sets as in (A.1), we have
$$A \cap B = \{x \colon p(x) \wedge q(x)\} \quad \text{and} \quad A \cup B = \{x \colon p(x) \vee q(x)\}.$$

Example A.14. Let $A = \{x \in \mathbb{R} \colon |x| \leq 1\}$ and $B = \{x \in \mathbb{R} \colon x \geq 0\}$. Thus, $A \cap B = \{x \in \mathbb{R} \colon 0 \leq x \leq 1\}$ and $A \cup B = \{x \in \mathbb{R} \colon x \geq -1\}$.

Properties A.15. Let $A, B, C \subset \Omega$. Then,

(1) $A \cap B = B \cap A$ and $A \cup B = B \cup A$ (commutativity)

(2) $A \cap (B \cap C) = (A \cap B) \cap C$ and $A \cup (B \cup C) = (A \cup B) \cup C$ (associativity)
(3) $A \cap (B \cup C) = (A \cap B) \cup (B \cap C)$ and $A \cup (B \cap C) = (A \cup B) \cap (B \cup C)$ (distributivity)
(4) $A \cap A = A$ and $A \cup A = A$ (idempotence)
(5) $A \cap (A \cup B) = A$ and $A \cup (A \cap B) = A$ (absortion)

Let $A, B \subset \Omega$.

- $A \setminus B = \{x \in \Omega \colon x \in A \wedge x \notin B\}$ is the *difference* between A and B (A minus B).
- $A^c = \{x \in \Omega \colon x \notin A\}$ is the *complementary set* of A in Ω.

As in (A.1) we can write:

$$A \setminus B = \{x \colon p(x) \wedge \sim q(x)\} \quad \text{and} \quad A^c = \{x \colon \sim p(x)\}.$$

Properties A.16.

(1) $A \setminus B = A \cap B^c$
(2) $A \cap A^c = \emptyset$
(3) $A \cup A^c = \Omega$.

It is also possible to define without difficulties the intersection and union of infinitely many sets. Let I to be a set, which we will call index set. This corresponds to the indices of a family of sets $A_\alpha \subset \Omega$ with $\alpha \in I$. Hence,

$$\bigcap_{\alpha \in I} A_\alpha = \{x \colon \forall_{\alpha \in I} x \in A_\alpha\} \quad \text{and} \quad \bigcup_{\alpha \in I} A_\alpha = \{x \colon \exists_{\alpha \in I} x \in A_\alpha\}.$$

Example A.17.

(1) Let $A_n = [n, n+1] \subset \mathbb{R}$, with $n \in \mathbb{N}$ (notice that $I = \mathbb{N}$). Then

$$\bigcap_{n \in \mathbb{N}} A_n = \emptyset, \quad \bigcup_{n \in \mathbb{N}} A_n = [1, +\infty[.$$

(2) Let $A_\alpha = [0, |\sin \alpha|]$, $\alpha \in I = \mathbb{R}$. Then

$$\bigcap_{\alpha \in \mathbb{R}} A_\alpha = \{0\}, \quad \bigcup_{\alpha \in \mathbb{R}} A_\alpha = [0, 1].$$

Proposition A.18 (Morgan laws).

(1)

$$\left(\bigcap_{\alpha \in I} A_\alpha\right)^c = \bigcup_{\alpha \in I} A_\alpha^c$$

(2)

$$\left(\bigcup_{\alpha \in I} A_\alpha\right)^c = \bigcap_{\alpha \in I} A_\alpha^c$$

EXERCISE A.19. Prove it.

If two sets do not intersect, i.e. $A \cap B = \emptyset$, we say that they are *disjoint*. A family of sets A_α, $\alpha \in I$, is called *pairwise disjoint* if for any $\alpha, \beta \in I$ such that $\alpha \neq \beta$ we have $A_\alpha \cap A_\beta = \emptyset$ (each pair of sets in the family is disjoint).

3. Function theory notions

Given two sets A and B, a *function* f is a correspondence between each $x \in A$ to *one and only one* $y = f(x) \in B$. It is also called a map or a mapping.

Representation:

$$\begin{aligned} f\colon A &\to B \\ x &\mapsto y = f(x). \end{aligned}$$

Notation:

- A is the domain of f.
- $f(C) = \{f(x) \in B\colon x \in C\}$ is the image of $C \subset A$.
- $f^{-1}(D) = \{x \in A\colon f(x) \in D\}$ is the pre-image of $D \subset B$.

EXAMPLE A.20.

(1) Let $A = \{a, b, c, d\}$, $B = \mathbb{N}$ and f a function $f\colon A \to B$ defined by the following table:

x	a	b	c	d
$f(x)$	3	5	7	9

Then, $f(\{b, c\}) = \{5, 7\}$, $f^{-1}(\{1\}) = \emptyset$, $f^{-1}(\{3, 5\}) = \{a, b\}$, $f^{-1}(\{n \in \mathbb{N}\colon \frac{n}{2} \in \mathbb{N}\}) = \emptyset$.

(2) A function whose domain is $\mathbb{N}$ is called a **sequence**. For example, consider $u\colon \mathbb{N} \to \{-1, 1\}$ given by $u_n = u(n) = (-1)^n$. Then, $u(\mathbb{N}) = \{-1, 1\}$, $u^{-1}(\{1\}) = \{2n\colon n \in \mathbb{N}\}$, $u^{-1}(\{-1\}) = \{2n - 1\colon n \in \mathbb{N}\}$.

(3) Let $f\colon \mathbb{R} \to \mathbb{R}$,

$$f(x) = \begin{cases} |x|, & x \leq 1 \\ 2, & x > 1. \end{cases}$$

Thus, $f(\mathbb{R}) = \mathbb{R}_0^+$, $f([1, +\infty[) = \{1, 2\}$, $f^{-1}([2, +\infty[) =]-\infty, -2] \cup]1, +\infty[$, $f^{-1}(\{1, 2\}) = \{-1, -2\} \cup [1, +\infty[$.

(4) For any set ω, the **identity function** is $f\colon \Omega \to \Omega$ with $f(x) = x$. We use the notation $f = \mathrm{Id}$.

(5) Let $A \subset \Omega$. The **indicator function** is $\mathcal{X}_A \colon \Omega \to \mathbb{R}$ with

$$\mathcal{X}_A(x) = \begin{cases} 1, & x \in A \\ 0, & x \notin A. \end{cases}$$

The pre-image behaves nicely with the union, intersection and complement of sets. Let I to be the set of indices of $A_\alpha \subset A$ and $B_\alpha \subset B$ with $\alpha \in I$.

Proposition A.21.

(1)
$$f\left(\bigcup_{\alpha\in I} A_\alpha\right) = \bigcup_{\alpha\in I} f(A_\alpha)$$

(2)
$$f\left(\bigcap_{\alpha\in I} A_\alpha\right) \subset \bigcap_{\alpha\in I} f(A_\alpha)$$

(3)
$$f^{-1}\left(\bigcup_{\alpha\in I} B_\alpha\right) = \bigcup_{\alpha\in I} f^{-1}(B_\alpha)$$

(4)
$$f^{-1}\left(\bigcap_{\alpha\in I} B_\alpha\right) = \bigcap_{\alpha\in I} f^{-1}(B_\alpha)$$

(5)
$$f^{-1}\left(B_\alpha^c\right) = f^{-1}(B_\alpha)^c$$

(6)
$$f(f^{-1}(B_\alpha) \subset B_\alpha$$

(7)
$$f^{-1}(f(A_\alpha)) \supset A_\alpha$$

Exercise A.22. Prove it.

3.1. Injectivity e surjectivity. According to the definition of a function $f \colon A \to B$, to each x in the domain it corresponds a unique $f(x)$ in the image. Notice that nothing is said about the possibility of another point x' in the domain to have the same image $f(x') = f(x)$. This does not happen for injective functions. On the other hand, there might be that B is different from $f(A)$. This does not happen for surjective functions.

- f is injective (one-to-one) iff $f(x_1) = f(x_2) \Rightarrow x_1 = x_2$.
- f is surjective (onto) iff $f(A) = B$.
- f is a bijection iff it is injective and surjective.

3.2. Composition of functions. After computing $g(x)$ as the image of x by a function g, in many situations we want to apply yet another function (ot the same) to $g(x)$, i.e. $f(g(x))$. It is said that we are composing two functions. Let $g\colon A \to B$ and $f\colon C \to D$. We define the *composition function* in the following way

$$f \circ g\colon g^{-1}(C) \to D$$
$$(f \circ g)(x) = f(g(x))$$

(read as f composed with g or f after g).

EXAMPLE A.23. Let $g\colon \mathbb{R} \to \mathbb{R}$, $g(x) = 1-2x$ and $f\colon [1,+\infty[\to \mathbb{R}$, $f(x) = \sqrt{x-1}$. We have that $g^{-1}([1,+\infty[) = \mathbb{R}_0^-$. So, $f\circ g\colon]-\infty,0] \to \mathbb{R}$, $f \circ g(x) = f(1-2x) = \sqrt{-2x}$.

An injective function $f\colon A \to B$ is also called invertible because we can find its *inverse function* $f^{-1}\colon f(A) \to A$ such that

$$\forall_{x\in A} f^{-1}(f(x)) = x \quad \text{and} \quad \forall_{y\in f(A)} f(f^{-1}(y)) = y.$$

EXAMPLE A.24.

(1) $f\colon \mathbb{R} \to \mathbb{R}$, $f(x) = x^2$. Is not invertible since, e.g. $f(1) = f(-1)$. However, if we restrict the domain to $\mathbb{R}_0^+$, it becomes invertible. I.e. $g\colon \mathbb{R}_0^+ \to \mathbb{R}$, $g(x) = x^2$ is invertible and $g(\mathbb{R}_0^+) = \mathbb{R}_0^+$. From $y = x^2 \Leftrightarrow x = \sqrt{y}$, we write $g^{-1}\colon \mathbb{R}_0^+ \to \mathbb{R}_0^+$, $g(x) = \sqrt{x}$.
(2) Let $\sin\colon \mathbb{R} \to \mathbb{R}$ be the function sine. This function is invertible if restricted to certain sets. For example, $\sin\colon [-\frac{\pi}{2},\frac{\pi}{2}] \to \mathbb{R}$ is invertible. Notice that $\sin([-\frac{\pi}{2},\frac{\pi}{2}]) = [-1,1]$. Then, we define the function arc-sine $\arcsin\colon [-1,1] \to [-\frac{\pi}{2},\frac{\pi}{2}]$ that to each $x \in [-1,1]$ corresponds the angle in $[-\frac{\pi}{2},\frac{\pi}{2}]$ whose sine is x. Finally, we have that $\arcsin(\sin x) = \sin(\arcsin x) = x$.
(3) When restricted to $[0,\pi]$ the cosine can also be inverted. The arc-cosine function $\arccos\colon [-1,1] \to [0,\pi]$ at $x \in [-1,1]$ is the angle whose cosine is x. Consequently, $\arccos(\cos x) = \cos(\arccos x) = x$.
(4) The tangent function in $]-\frac{\pi}{2},\frac{\pi}{2}[$ has the inverse given by the arc-tangent function $\operatorname{arctg}\colon \mathbb{R} \to]-\frac{\pi}{2},\frac{\pi}{2}[$ such that $\operatorname{arctg}(\operatorname{tg} x) = \operatorname{tg}(\operatorname{arctg} x) = x$.
(5) The exponential function $f\colon \mathbb{R} \to \mathbb{R}$, $f(x) = e^x$ is invertible and $f(\mathbb{R}) = \mathbb{R}^+$. Its inverse is the logarithm function $f^{-1}\colon \mathbb{R}^+ \to \mathbb{R}$, $f^{-1}(x) = \log x$.

PROPOSITION A.25. *If f and g are invertible, then $f\circ g$ is invertible and*

$$(f \circ g)^{-1} = g^{-1} \circ f^{-1}.$$

PROOF.

- If $f \circ g(x_1) = f \circ g(x_2) \Leftrightarrow f(g(x_1)) = f(g(x_2))$, then as f is invertible, $g(x_1) = g(x_2)$. In addition, as g is invertible, it implies that $x_1 = x_2$. So, $f \circ g$ is invertible.
- We can now show that $g^{-1} \circ f^{-1}$ is the inverse of $f \circ g$:
  - $g^{-1} \circ f^{-1}(f \circ g(x)) = g^{-1}(f^{-1}(f(g(x)))) = g^{-1}(g(x)) = x$,
  - $f \circ g(g^{-1} \circ f^{-1}(x)) = f(g(g^{-1}(f^{-1}(x)))) = f(f^{-1}(x)) = x$,

  where we have used the fact that f^{-1} and g^{-1} are the inverse functions of f and g, respectively.
- It remains to show that the inverse is unique. Suppose that there is another inverse function of $f \circ g$ namely u different from $g^{-1} \circ f^{-1}$. Hence, $f \circ g(u(x)) = x$. If we apply the function $g^{-1} \circ f^{-1}$, then $g^{-1} \circ f^{-1}(f \circ g(u(x))) = g^{-1} \circ f^{-1}(x) \Leftrightarrow u(x) = g^{-1} \circ f^{-1}(x)$.

□

3.3. Countable and uncountable sets. As seen before, we can classify sets in terms of its number of elements, either finite or infinite. There is a particular case of an infinite set: the set of natural numbers $\mathbb{N}$. This set can be counted in the sense that we can have an ordered sequence of its elements: given any element we know what is the next one.

A set A is *countable* if there is a one-to-one function $f: A \to \mathbb{N}$. Countable sets can be either finite (f only takes values in a finite subset of $\mathbb{N}$) or infinite (like $\mathbb{N}$). A set which is not countable is called *uncountable*.

Exercise A.26. Let $A \subset B$. Show that

(1) If B is countable, then A is also countable.
(2) If A is uncountable, then B is also uncountable.

Example A.27. The following are countable sets:

(1) $\mathbb{Q}$, by choosing a sequence that covers all rational numbers. Find one that works.
(2) $\mathbb{Z}$, because $\mathbb{Z} \subset \mathbb{Q}$.

Example A.28. The following are uncountable sets:

(1) $[0, 1]$, by the following argument. Suppose that $[0, 1]$ is countable. This implies the existence of a sequence that covers all the points in $[0, 1]$. Write the sequence as $x_n = 0.a_{n,1}a_{n,2}\dots$ where $a_{n,i} \in \{0, 1, 2, \dots, 9\}$. Take now $x \in [0, 1]$ given by $x = 0.b_1b_2\dots$ where $b_i \neq a_{i,i}$ for every $i \in \mathbb{N}$. In order to avoid the cases of the type $0.1999\dots = 0.2$, whenever $a_{i,i} = 9$ we choose $b_i \neq 9$ also. Thus, x is different from every point in the

sequence (it is different from every x_n because $b_n \neq a_{n,n}$). So, $[0,1]$ can not be countable.
(2) $\mathbb{R}$, because $[0,1] \subset \mathbb{R}$.

PROPOSITION A.29. *Let A and B to be any two sets and $h\colon A \to B$ a bijection between them. Then,*

(1) *A is finite iff B is finite.*
(2) *A is countable iff B is countable.*
(3) *A is uncountable iff B is uncountable.*

EXERCISE A.30. Prove it.

Consider an index set I and a family of sets A_α with $\alpha \in I$. If I is finite, we say that

$$\bigcap_{\alpha \in I} A_\alpha$$

is a finite intersection. If I is infinite but countable, the above is a countable intersection. Otherwise, whenever I is uncountable, it is called an uncountable intersection. Similarly, we use the same type of nomenclature for unions.

4. Topological notions in ℝ

4.1. Distance. The usual *distance* between two points $x, y \in \mathbb{R}$ is given by

$$d(x,y) = |x - y|. \tag{A.2}$$

We can easily deduce the following properties.

PROPERTIES A.31. For all $x, y, z \in \mathbb{R}$,

(1) $d(x,y) \geq 0$
(2) $d(x,y) = 0 \Leftrightarrow x = y$
(3) $d(x,y) = d(y,x)$ (symmetry)
(4) $d(x,z) \leq d(x,y) + d(y,z)$ (triangular inequality).

In fact, we could have defined distance[1] only using the above properties, since they are the relevant ones. An example of another distance d on $\mathbb{R}$ satisfying the same properties is:

$$d(x,y) = \frac{|x-y|}{1+|x-y|}.$$

Notice that with this distance we have $d(0,1) = d(1,2) = \frac{1}{2}$ and that $d(0,2) = \frac{2}{3}$. On the other hand, there are no points whose distance between each other is more than 1.

[1]In some literature it is called a metric.

We will restrict our study to the usual distance in (A.2). However, with some care we could have developped our study for a generic distance.

4.2. Neighbourhood. One of the main consequences of the ability to measure distances is the notion of proximity. Let $a \in \mathbb{R}$ and $\varepsilon > 0$. An *ε-neighbourhood* of a is the set of points which are at a distance less than ε from a. That is,

$$V_\varepsilon(a) = \{x \in \mathbb{R} \colon d(x, a) < \varepsilon\}.$$

For the usual distance we obtain

$$V_\varepsilon(a) =]a - \varepsilon, a + \varepsilon[.$$

Proposition A.32.

(1) *If $0 < \delta < \varepsilon$, then $V_\delta(a) \subset V_\varepsilon(a)$ and $V_\delta(a) \cap V_\varepsilon(a) = V_\delta(a)$.*
(2) *$\bigcap_{\varepsilon>0} V_\varepsilon(a) = \{a\}$ is not a neighbourhood of a.*
(3) *If $a \neq b$, then $V_\delta(a) \cap V_\varepsilon(a) = \emptyset \Leftrightarrow \delta + \varepsilon \leq |b - a|$.*

4.3. Interior, boundary and exterior. With the notion of neighbourhood of a point, we can distinguish the points that are in the "interior" of a set. Let $A \subset \mathbb{R}$ and $a \in \mathbb{R}$.

- a is an *interior point* of A iff there is a neighbourhood of a contained in A, i.e.

$$\exists_{\varepsilon>0} V_\varepsilon(a) \subset A.$$

- a is an *exterior point* of A iff it is an interior point of A^c.
- a is a *boundary point* of A iff it is neither interior nor exterior.

The set of interior points of A is denoted by $\operatorname{int} A$, the exterior by $\operatorname{ext} A$ and the boundary by $\operatorname{front} A$. So,

$$\mathbb{R} = \operatorname{int} A \cup \operatorname{front} A \cup \operatorname{ext} A.$$

Example A.33.

(1) $\operatorname{int}[0,1[=]0,1[$, $\operatorname{front}[0,1[= \{0,1\}$, $\operatorname{ext}[0,1[= \mathbb{R} \setminus [0,1]$.
(2) $\operatorname{int} \emptyset = \operatorname{front} \emptyset = \operatorname{ext} \mathbb{R} = \operatorname{front} \mathbb{R} = \emptyset$, $\operatorname{ext} \emptyset = \operatorname{int} \mathbb{R} = \mathbb{R}$.
(3) $\operatorname{int} \mathbb{Q} = \operatorname{ext} \mathbb{Q} = \operatorname{int}(\mathbb{R} \setminus \mathbb{Q}) = \operatorname{ext}(\mathbb{R} \setminus \mathbb{Q}) = \emptyset$, $\operatorname{front} \mathbb{Q} = \operatorname{front}(\mathbb{R} \setminus \mathbb{Q}) = \mathbb{R}$.

Proposition A.34.

(1) $\operatorname{int}(A^c) = \operatorname{ext} A$.
(2) $\operatorname{ext}(A^c) = \operatorname{int} A$.
(3) $\operatorname{front}(A^c) = \operatorname{front} A$.
(4) $\operatorname{int} A \subset A$.
(5) $\operatorname{ext} A \subset A^c$.

4.4. Open and closed sets. The *closure* of $A \subset \mathbb{R}$ is

$$\overline{A} = \operatorname{int} A \cup \operatorname{front} A.$$

Therefore, $\mathbb{R} = \overline{A} \cup \operatorname{ext} A$ and

$$\operatorname{int} A \subset A \subset \overline{A}.$$

In cases one has equalities, we label the set A as:

- *open* iff $\operatorname{int} A = A$.
- *closed* iff $A = \overline{A}$.

Example A.35.

(1) $]0,1[$ is open, $[0,1]$ is closed, $]0,1]$ is neither open nor closed, $]-\infty,1]$ is closed.
(2) $\mathbb{N}$ and $\mathbb{Z}$ are closed, $\mathbb{Q}$ and $\mathbb{R}\setminus\mathbb{Q}$ are neither open nor closed.
(3) $\emptyset$ and $\mathbb{R}$ are open and closed.

Proposition A.36.

(1) *$A \subset \mathbb{R}$ is open iff A^c is closed.*
(2) *If $A, B \subset \mathbb{R}$ are open, then $A \cap B, A \cup B$ are open.*
(3) *If $A, B \subset \mathbb{R}$ are closed, then $A \cap B, A \cup B$ are closed.*
(4) *If $A \neq \emptyset$ is bounded and closed (compact), then it has a maximum and a minimum.*

Proof.

(1) A open $\Leftrightarrow \operatorname{front} A \subset A^c \Leftrightarrow \operatorname{front} A^c \subset A^c \Leftrightarrow A^c$ closed.
(2) Let $a \in A \cap B$. Then, as A and B are open, there are $\varepsilon_1, \varepsilon_2 > 0$ such that
$$V_{\varepsilon_1}(a) \subset A \quad \text{e} \quad V_{\varepsilon_2}(a) \subset B.$$
Choosing $\varepsilon = \min\{\varepsilon_1, \varepsilon_2\}$, we have that
$$V_\varepsilon(a) \subset V_{\varepsilon_1}(a) \cap V_{\varepsilon_2}(a) \subset A \cap B.$$
Same idea for $A \cup B$.
(3) A and B closed $\Leftrightarrow A^c$ and B^c open $\Rightarrow A^c \cup B^c$ open $\Leftrightarrow (A \cap B)^c$ open $\Leftrightarrow A \cap B$ closed. Same idea for $A \cup B$.
(4) A bounded from above $\Rightarrow \sup A \in \operatorname{front} A$. As A is closed, i.e. $\operatorname{front} A \subset A$, we have that $\sup A \in A$. Thus, $\max A = \sup A$. Same idea for inf and min.

$\square$

Remark A.37. The following example

$$\bigcap_{n=1}^{+\infty} \left]-\frac{1}{n}, \frac{1}{n}\right[= \{0\}$$

shows that the infinite intersection of open sets might not be an open set In the previous proposition it is only proved that the finite interesection of open sets is an open set. The infinite union of closed sets might not also be a closed set. For example,

$$\bigcup_{n=1}^{+\infty}\left[-1+\frac{1}{n},1-\frac{1}{n}\right]=]-1,1[.$$

Proposition A.38. *Any open set is a countable union of pairwise disjoint open intervals.*

Proof. Let $U\subset\mathbb{R}$ be an open set and $x\in U$. Then, there is a neighbourhood V of x contained in U. Write $I=]a,b[$ where

$$a=\inf\{\alpha\colon\,]\alpha,x[\subset U\}\quad\text{and}\quad b=\sup\{\beta\colon\,]x,\beta[\subset U\}.$$

So, I_r is the maximal interval in U containing x. Choose a rational number r in each I and denote it by I_r. Notice that for any other rational number $r'\in I_r$ we have $I_{r'}=I_r$. In addition, if $I_r\neq I_{r'}$, then $I_r\cap I_{r'}=\emptyset$.

Since any $x\in U$ is inside some I_r with $r\in\mathbb{Q}\cap U$,

$$U\subset\bigcup_{r\in\mathbb{Q}\cap U}I_r.$$

Moreover, $I_r\subset U$ for every $r\in\mathbb{Q}\cap U$, so

$$U=\bigcup_{r\in\mathbb{Q}\cap U}I_r.$$

□

4.5. Accumulation points. Note first that for $A\subset\mathbb{R}$ and $a\in\mathbb{R}$:

$$a\in\overline{A}\Leftrightarrow\forall_{\varepsilon>0}V_\varepsilon(a)\cap A\neq\emptyset.$$

I.e. a point a belongs to the closure of A (in its interior or at the boundary) iff any neighbourhood of a intersects A.

We are now interested in the closure points of A having for sure closeby other points of the set. In other words, that are not isolated points:

- $a\in A$ is an **isolated point** of A iff $\exists_{\varepsilon>0}V_\varepsilon(a)\cap(A\setminus\{a\})=\emptyset$.

So, we define

- a is an **accumulation point** of A iff $\forall_{\varepsilon>0}V_\varepsilon(a)\cap(A\setminus\{a\})\neq\emptyset$.

Accumulation points are thus elements of the closure of A minus the isolated ones. The set of accumulation points is denoted by A'.

Example A.39.

(1) $([0,1[)' = [0,1]$.
(2) $(\{0,1\})' = \emptyset$.
(3) $(\{\frac{1}{n} \colon n \in \mathbb{N}\})' = \{0\}$.
(4) $\mathbb{Q}' = \mathbb{R}$, $(\mathbb{R} \setminus \mathbb{Q})' = \mathbb{R}$.

4.6. Numerical sequences and convergence. A numerical sequence (or simply sequence for short) is a real valued function defined on $\mathbb{N}$, i.e. $u \colon \mathbb{N} \to \mathbb{R}$. Its expression is usually written as u_n instead of $u(n)$. Each n is said to be an order of the sequence and the respective value u_n is the term of order n.

Whenever we have a strictly increasing sequence k_n we write $k_n \nearrow \infty$. This allows us to define a subsequence of u_n by u_{k_n}.

A sequence u_n is said to converge to $b \in \mathbb{R}$ (or b is the limit of u_n) iff

$$\forall_{\varepsilon>0} \exists_{p\in\mathbb{N}} \forall_{n\geq p}\, u_n \in V_\varepsilon(b).$$

In this case we write $u_n \to b$ as $n \to +\infty$ or $\lim_{n\to+\infty} u_n = b$.

A sequence might not have a limit. In that case there is still a possibility that subsequences have limits. The sublimits of u_n are the limits of its subsequences. The infimum of this set is denoted by $\liminf u_n$, and its supremum is $\limsup u_n$. These are also computable by

$$\liminf u_n = \sup_{n\geq 1} \inf_{k\geq n} u_n = \lim_{n\to+\infty} \inf_{k\geq n} u_n$$

and

$$\limsup u_n = \inf_{n\geq 1} \sup_{k\geq n} u_n = \lim_{n\to+\infty} \sup_{k\geq n} u_n.$$

Consider now the sum of the first n terms of a sequence u_n given by a new sequence

$$S_n = \sum_{i=1}^{n} u_i.$$

A numerical series is the limit of S_n as $n \to +\infty$ and it is denoted by

$$\sum_{i=1}^{+\infty} u_i = \lim_{n\to+\infty} \sum_{i=1}^{n} u_i.$$

4.6.1. *Diagonal argument.* Consider a sequence $u_{n,m}$ that depends on two variables $n, m \in \mathbb{N}$.

THEOREM A.40 (Diagonal argument). *If there is $M > 0$ such that $|u_{n,m}| < M$ for all $n, m \in \mathbb{N}$, then there is $k_n \nearrow \infty$ with $k_n \in \mathbb{N}$ such that*

$$a_m := \lim_{n\to+\infty} u_{k_n,m}$$

exists for every $m \in \mathbb{N}$.

PROOF. Notice that $u_{n,1}$ is a bounded sequence, thus there exists a subsequence $u_{k_n^{(1)},1}$ which converges to a_1. Choose now $k_n^{(2)}$ to be a subsequence of $k_n^{(1)}$ so that $u_{k_n^{(2)},2}$ converges to a_2. By induction we obtain a subsequence $k_n^{(j)}$ of $k_n^{(j-1)}$ so that $u_{k_n^{(j)},j}$ converges to a_j. Therefore, $u_{k_n^{(j)},i}$ converges for every $i \leq j$.

Finally, take $k_n = k_n^{(n)}$ which is called the diagonal subsequence. Hence, $u_{k_n,m}$ converges for every m.

$\square$

5. Notions of differentiable calculus on $\mathbb{R}$

Consider a function $f\colon A \to \mathbb{R}$ where $A \subset \mathbb{R}$. We say that the limit of f at $a \in A$ is b iff

$$\forall_{\varepsilon>0}\exists_{\delta>0}\forall_{x\in V_\delta(a)\cap A} f(x) \in V_\varepsilon(b).$$

In this case we write $f(x) \to b$ as $x \to a$ or $\lim_{x\to a} f(x) = b$. This is also equivalent to say that for any sequence u_n with values in A that converges to a we have $f(u_n)$ converging to b.

We say that f is continuous at a iff $\lim_{x\to a} f(x) = f(a)$. Moreover, f is continuous in A if it is continuous at every point of A.

Now, if $a \in A$ and A is an open set, let

$$f'(a) = \lim_{x\to a} \frac{f(x) - f(a)}{x - a}.$$

If the above limit exists we say that f is differentiable at a and it is called the derivative of f at a. The function is differentiable in A if it is differentiable at every point of A. The function $f'\colon A \to \mathbb{R}$ is the derivative of f and it can also be differentiable. In this case we can compute the second derivative f'' and so on. If performing the derivative k times we write $f^{(k)}$ as the k-th derivative of f. The 0-derivative corresponds to f itself.

The set of all continuous functions in A is denoted by $C^0(A)$. The functions that are differentiable in A whose derivatives are continuous form the set $C^1(A)$. More generally, $C^k(A)$ is the set of k-times differentiable functions whose k-th derivative is continuous. Finally, $C^\infty(A)$ corresponds to the set of all functions which are infinitely times differentiable.

Any $f \in C^{k+1}(A)$ can be approximated by its Taylor polynomial around $x_0 \in A$:

$$P(x) = \sum_{i=0}^{k} \frac{f^{(i)}(x_0)}{n!}(x - x_0)^i,$$

with an error given by

$$f(x) - P(x) = \frac{f^{(k+1)}(\xi)}{n!}(x - x_0)^{k+1}$$

for some ξ between x and x_0. If $k = 0$ and $f \in C^1(A)$ we can write

$$f(x) - f(x_0) = f'(\xi)\,(x - x_0),$$

which is known as the mean value theorem.

6. Greek alphabet

Letter	lower case	upper case
Alpha	α	A
Beta	β	B
Gamma	γ	Γ
Delta	δ	Δ
Epsilon	$\epsilon\,\varepsilon$	E
Zeta	ζ	Z
Eta	η	E
Theta	$\theta\,\vartheta$	Θ
Iota	ι	I
Kappa	κ	K
Lambda	λ	Λ
Mu	μ	M
Nu	ν	N
Xi	ξ	Ξ
Omicron	o	O
Pi	$\pi\,\varpi$	Π
Rho	$\rho\,\varrho$	R
Sigma	$\sigma\varsigma$	Σ
Tau	τ	T
Upsilon	υ	Υ
Phi	$\phi\,\varphi$	Φ
Chi	χ	X
Psi	ψ	Ψ
Omega	ω	Ω

Bibliography

[1] M. Capinski and E. Kopp. *Measure, Integral and Probability* Springer, 2005.
[2] H. L. Royden. *Real Analysis*. Macmillan, 3rd Ed, 1988.
[3] W. Rudin. *Real and Complex Analysis*. McGraw-Hill, 3rd Ed, 1987.
[4] S. R. S. Varadhan. *Probability Theory* AMS, 2001.