Lecture notes

Advanced probability theory

Joonas Turunen

Greatly inspired by earlier notes of Nathanaël Berestycki and Alexander Glazman

University of Vienna

Faculty of Mathematics

Contents

Chapter 0

Introduction

This document comprises the lecture notes of the course *Advanced probability theory* lectured at the University of Vienna, Faculty of Mathematics, in the summer term 2024. The course is part of the basic courses in the *Stochastics* master curriculum. These notes are an updated version of the respective notes for the summer 2023 course, including also corrections of various mistakes and typos. The author acknowledges Nathanael Berestycki and Alexander Glazman for providing their former handwritten notes as a basis of the current set of notes. He also warmly thanks the students for giving valuable feedback and spotting numerous typos.

0.1 Motivation

This is not a course about deep philosophical questions like *"Does probability exist?"*, and we hereby omit considering these straight away; an interested reader is referred to other sources. Neither is this course merely a *toolbox for applications*, and on the contrary, one could admit that it is rather theoretical (as its name suggests). Yet, one of the goals is that the student learns to use modern probability theory in various applications in both pure mathematics and in more applied fields.

The main goal of these notes, however, is to introduce the theory in its rigorous *measure-theoretic* framework. Assuming basic axioms behind the measure theory (such as the axiom of choice) to hold, we can for sure give a positive answer to the first question: *probability exists* as a measure-theoretic construction! These notes will showcast that measure-theoretic approach to probability indeed makes sense, and leads to the same probabilistic notions that correspond to our intuition. In fact, measure theory provides probability with much more, and the reader of these notes would (hopefully) experience and learn to appreciate its power. One should also learn that in particular, probability theory is not just "a theory of finite measures", but rather its own beautiful field of mathematics. Although there is the word *advanced* in the course title, this course is still an introduction to modern probability, providing keys to deeper explorations in the field. In addition, modern measure-theoretic probability has found its applications in neighboring fields of mathematics and science, such as combinatorics, statistical physics, statistics, mathematical finance and biomathematics. More practical applications could be found eg. in weather forecasts, artificial intelligence and financial risk analysis, to name but a few. In these notes, we introduce one particular (albeit theoretical) application, namely *percolation theory*, which has its roots in polymer chemistry and statistical mechanics as a model of porous medium.

0.2 Very brief history

In this brief section, we follow the (partial yet detailed) historic account of [3, Appendix III]. One of the earliest occurrence of probability was a poem called ”De Vetula” from France in 1250 AD, where some calculations including dice were presented. Much later, during the renaissance era and motivated by gambling, probability of dice and cards continued to be mathematically studied. One of the prominent researchers that time was Galileo, and perhaps somewhat less famous Cardano even wrote a probabilistic book called *On games of chance*, which was published in 1663 and written around 100 years earlier. During the 18th century, researchers including Bernoulli, De Moivre and Laplace developed more advanced results like a weak law of large numbers and the central limit theorem. In the same century, many prominent mathematicians, like Euler, Gauss, Lagrange, Legendre and Poisson, made their contributions to the developing theory. However, there was a lack of well-defined axiomatic characterization in probability until the beginning of the 20th century, when Hilbert asked for this characterization in his 6th problem. This lack was addressed by Kolmogorov in 1933 by his axioms of probability. Meanwhile, around 1900, Borel made his significant contributions in the newly born measure theory, and the theory was developing very fast during the first half of the 20th century. Although Kolmogorov's probability theory already included some measure theory, the interplay between probability and measure continued its development throughout the 20th century. An interested reader may see [5] for a detailed exposition of the early connections between measure theory and probability.

Chapter 1

Basic notions of measure-theoretic probability

A very basic idea in probability theory is to consider *events*, which are viewed as subsets of some (general) set, and to find a way to "measure" the probability mass assigned to each event. Roughly speaking, this intuition together with our knowledge of rigorous measure theory leads to measure-theoretic probability theory. We therefore start by recalling some essential definitions and facts from measure theory, mostly without proofs, which will belong to the toolbox of this course; we refer the reader to any book or lecture notes about measure theory for further details.

1.1 Measure theory recap

Definition 1. A *measure space* is a triple $(\Omega, \mathcal{A}, \mu)$ where

- Ω is a set
- $\mathcal{A}$ is a *sigma-algebra*: $\mathcal{A} \subset \mathcal{P}(\Omega)$ such that

  (i) $\emptyset \in \mathcal{A}$

  (ii) $A \in \mathcal{A} \Longrightarrow A^c = \Omega \setminus A \in \mathcal{A}$

  (iii) $A_i \in \mathcal{A}$, $i \in \mathbb{N} \Longrightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$

- μ is a *measure* on Ω: $\mu : \mathcal{A} \to [0, \infty]$ is a function such that

  (i) $\mu(\emptyset) = 0$

  (ii) $A_i \in \mathcal{A}$, $i \in \mathbb{N}$ mutually disjoint sets (i.e. $A_j \cap A_k = \emptyset$ for $j \neq k$)
  $\Longrightarrow \mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu\left(A_i\right)$

Typically, the σ-algebra $\mathcal{A}$ is viewed as the "largest meaningful family" of sets whose size we will measure. Later in the course, we will also see what effect it makes to consider sub-σ-algebras, which intuitively correspond to restricting information. Note that if there is a set $A \in \mathcal{A}$ with $\mu(A) < \infty$, then the first assumption of measure, $\mu(\emptyset) = 0$, is redundant. Namely, by the additivity property of the measure, we have $\mu(A) = \mu(A \cup \emptyset) = \mu(A) + \mu(\emptyset)$ (observe that the empty set is disjoint with any set). For example, if μ is a probability measure, this will always be the case. A set $A \in \mathcal{A}$ will be called a *measurable set*.

Remark 2. It is very useful to notice the following properties of a measure μ:

- $A, B \in \mathcal{A}$, $A \subset B \Longrightarrow \mu(A) \leq \mu(B)$ (monotonicity)
- $A_i \in \mathcal{A}$, $i \in \mathbb{N} \Longrightarrow \mu\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} \mu\left(A_i\right)$ (subadditivity)

Together with $\mu(\emptyset) = 0$, this states that μ is an outer measure (as it should, since it is assumed to be even a measure).

One very important special case of sigma-algebra is the *Borel σ-algebra* when $\Omega = \mathbb{R}^n$, which is defined as follows.

Definition 3 (Borel σ-algebra)**.** The Borel σ-algebra $\mathcal{B}(\mathbb{R}^n)$ on $\Omega = \mathbb{R}^n$ is the σ-algebra *generated* by the open sets of $\mathbb{R}^n$. That is, the smallest σ-algebra of $\mathbb{R}^n$ which contains all the open sets of $\mathbb{R}^n$. When there is no risk of confusion, we simply denote $\mathcal{B}(\mathbb{R}^n) \equiv \mathcal{B}$.

Other generating sets could be chosen. Most obviously, one could use closed sets (by (ii) in the general definition of σ-algebra), or one could even consider multi-intervals (observe that every open set in $\mathbb{R}$ is an union of open intervals, and a generalization to n dimensions is straightforward). As an exercise, the reader is encouraged to show that $\mathcal{B}$ on $\mathbb{R}$ is generated by the semi-infinite closed intervals of the form $(-\infty, a]$.

Next, we consider perhaps the most important example of a measure space.

Example 4 (Lebesgue measure on $\mathbb{R}$)**.** Let $\Omega = \mathbb{R}$ and $\mathcal{A} = \mathcal{B}$. The *Lebesgue measure* μ is the unique measure on $\mathbb{R}$ such that for any $a, b \in \mathbb{R}$ with $a < b$, $\mu((a, b)) = b - a$. For the existence and the uniqueness of such a measure, we refer the reader to any basic course material of measure theory (recall that the open intervals (a, b) generate the Borel σ-algebra).

We now proceed with functions defined on measure spaces, which leads to the following concept of *measurable functions.*

Definition 5 (Measurable function)**.** A function $f : \Omega \to \mathbb{R}$ is measurable if for every $B \in \mathcal{B}$, $f^{-1}(B) \in \mathcal{A}$.

The definition of measurable function extends naturally to a mapping between two arbitrary measure spaces. Next, we would like to consider integrable functions on measure spaces. As a caveat, not every measurable function is integrable, but the definition of integrability with respect to a measure is rather natural generalization of Riemann-integrability at least in an intuitive level. However, in most common applications, the integrals will be well-defined.

Definition 6 (Integrable function)**.** Let $(\Omega, \mathcal{A}, \mu)$ be a measure space. A function $f : \Omega \to \mathbb{R}$ is integrable if it is measurable and $\int_\Omega |f(x)|\, d\mu(x) < \infty$. In particular, the integral $\int_\Omega f(x) d\mu$ is well-defined.

Remark 7. In addition to integrable functions, the integral can be defined for *non-negative* functions, which is very important in probability theory. In that case, the integral may have value ∞.

Remark 8. If $\mathcal{A} = \mathcal{B}$, μ is the Lebesgue measure and $f : \Omega \to \mathbb{R}$ is measurable and *Riemann-integrable*, then its integral can be computed as the usual Riemann integral. In practice, this is the most common way to compute expectations in the case of continuous probability distributions.

Example 9 (A change of measure)**.** Let f be a measurable non-negative function on a measure space $(\Omega, \mathcal{A}, \mu)$. We can define a new measure ν on the same space by setting $\nu(A) := \int_\Omega f(x)\mathbb{1}_{x\in A} d\mu(x) = \int_A f(x)d\mu(x)$ for every $A \in \mathcal{A}$. We leave it as an exercise to the reader to prove that it indeed defines a measure. This is the first example of a *change of measure* in these notes, and can be reformulated using differentials as $d\nu(x) = f(x)d\mu(x)$.

For example, consider $\Omega = [0, 1]$ equipped with the Borel σ-algebra and the Lebesgue measure, and let $f(x) = x(1 - x)$. Then $\nu([0, 1]) = \frac{1}{2} - \frac{1}{3} = \frac{1}{6}$.

Example 10 (Dirac measure)**.** Let $a \in \Omega$, where Ω is equipped with any σ-algebra $\mathcal{A}$. Then the *Dirac* δ-measure for a is defined as

$$\delta_a(A) = \begin{cases} 1 & \text{if} \quad a \in A \\ 0 & \text{if} \quad a \notin A \end{cases} \tag{1.1}$$

for every $A \in \mathcal{A}$. Then for any function $f : \Omega \to \mathbb{R}$, $\int_\Omega f(x)d\delta_a(x) = f(a)$. In particular, every function is integrable with respect to the Dirac measure. In fact, this is the first example of a *probability measure* in this course: $\delta_a(\Omega) = 1$.

More generally, if $p_1 \dots, p_k$ are positive real numbers and $a_1, \dots, a_k \in \Omega$, then the linear combination $\mu(A) = \sum_{i=1}^k p_i\delta_{a_i}(A)$ defines a measure on Ω.

Three fundamental convergence theorems We collect now three important theorems on integral convergence, which will be repeatedly used during these notes. The first two of them use the important concept of *convergence almost everywhere*:

Definition 11 (Convergence almost everywhere)**.** Let $(f_n)_{n\geq 0}$ and f be μ-measurable functions. Then $f_n \xrightarrow[n\to\infty]{} f$ almost everywhere (a.e.), if $\mu\left(\{x \in \Omega : f_n(x) \xrightarrow[n\to\infty]{} f(x)\}\right) = \mu(\Omega)$. Equivalently, $f_n \xrightarrow[n\to\infty]{} f$ almost everywhere (a.e.), if there exists a measurable set E such that $\mu(E) = \mu(\Omega)$ and $f_n(x) \xrightarrow[n\to\infty]{} f(x)$ for all $x \in E$. That is, the convergence holds for all $x \in \Omega$ except possibly in a set with zero μ-measure.

Before we noticed that defining an integral makes sense for non-negative or integrable functions. The following two theorems tell us sufficient conditions when the limit operation and integration can be exchanged in these two respective cases.

Theorem 12 (Monotone convergence theorem)**.** *Let $(f_n)_{n\geq 0}$ and f be non-negative measurable functions such that $f_n \leq f_{n+1}$ for all n and $f_n \nearrow f$ a.e. in Ω as $n \to \infty$. Then*

$$\int_\Omega f_n(x)d\mu(x) \nearrow \int_\Omega f(x)d\mu(x). \tag{1.2}$$

Theorem 13 (Dominated convergence theorem)**.** *Let $(f_n)_{n\geq 0}$, f and g be measurable functions such that $f_n \xrightarrow[n\to\infty]{} f$ a.e. and $|f_n(x)| \leq g(x)$ a.e. with $\int_\Omega g d\mu < \infty$. Then*

$$\int_\Omega f_n(x)d\mu(x) \xrightarrow[n\to\infty]{} \int_\Omega f(x)d\mu(x). \tag{1.3}$$

If we relax the assumption of a converging sequence of functions, we can still consider the lim inf of any sequence, which is meaningful for integrals if the functions stay non-negative. Then the cost is that the equality becomes an inequality, which reflects the fact that without strong control of the function sequence, some mass might be lost at the infinity after exchanging limit and integration:

Theorem 14 (Fatou's lemma). *Let $(f_n)_{n\geq 0}$ be non-negative measurable functions. Then*

$$\int_\Omega \liminf_{n\to\infty} f_n(x)d\mu(x) \leq \liminf_{n\to\infty} \int_\Omega f_n(x)d\mu(x). \tag{1.4}$$

The monotone convergence theorem has the following important consequence:

Lemma 15. *Let $(A_n)_{n=0}^\infty$ be an increasing sequence of sets in $\mathcal{A}$, that is, $A_n \subset A_{n+1}$ for all $n \in \mathbb{N}$. Then $\mu(A_n) \xrightarrow[n\to\infty]{} \mu\left(\bigcup_{n=0}^\infty A_n\right)$.*

Proof. Define

$$f_n(x) = \mathbb{1}_{A_n}(x) = \begin{cases} 1 & \text{if} \quad x \in A_n \\ 0 & \text{if} \quad x \notin A_n \end{cases}.$$

Then obviously f_n is an increasing sequence of functions with $\lim_{n\to\infty} f_n(x) = \mathbb{1}_{\bigcup A_n}(x) =: f(x)$. Using the MCT then yields

$$\mu(A_n) = \int_\Omega f_n(x)d\mu(x) \xrightarrow[n\to\infty]{} \int_\Omega f(x)d\mu(x) = \mu\left(\bigcup_{n=0}^\infty A_n\right).$$

□

For finite measures (including probability measures) we have the following complementary result, also known as the *upper semicontinuity* of μ. We leave its proof as an exercise to the reader.

Corollary 16. *Let μ be a finite measure, and let $(A_n)_{n\geq 0}$ be a decreasing sequence of sets in $\mathcal{A}$, that is, $A_{n+1} \subset A_n$ for all $n \in \mathbb{N}$. Then $\mu(A_n) \xrightarrow[n\to\infty]{} \mu\left(\bigcap_{n=0}^\infty A_n\right)$.*

1.2 Probability and measure

In this section, we make the connection of measure theory to probability precise. At the first glance, it may appear to the reader that we are just considering measure theory with a particular finite measure. However, as we proceed, we will see how the assumption of a probability measure will yield a rich theory in its own interest and its peculiar consequences.

Definition 17. A measure space $(\Omega, \mathcal{F}, \mathbb{P})$ is a *probability space* if $\mathbb{P}(\Omega) = 1$. In this case, the set Ω is called the *sample space*, the σ-algebra $\mathcal{F}$ is the *set of events* and $\mathbb{P}$ is a *probability measure.* A point $\omega \in \Omega$ is called an *outcome* and a set $A \in \mathcal{F}$ an *event.*

Remark 18. In the vast majority of cases, we are not interested in Ω as a point set, but instead focus on $\mathcal{F}$ as a set of events for whom we can compute the probability using the measure $\mathbb{P}$ (the events outside $\mathcal{F}$ may be too complicated to deal with). Thus, we often assume that Ω is implicitly given to us and do not mention it at all. The σ-algebra $\mathcal{F}$ can be viewed as the *information* as a basis of our probability calculations; later we will consider sub σ-algebras and even σ-algebras depending on a time parameter, which naturally arise in various applications of probability. As a rule of thumb, the role of the σ-algebra is more important in probability theory than in general measure theory.

Remark 19. If $A, B \in \mathcal{F}$ are two events, then $A \cap B$ means that both events occur, while $A \cup B$ signifies that *at least* one of them occurs. By induction, this can be generalized to any finite number of events, and using common set operations one can construct abstract events for various purposes.

Example 20. If A, B, C are three possible events, the event corresponding to "only A occurs" is $A \cap B^c \cap C^c = A \cap (B \cup C)^c$.

Next, we give some important examples of probability spaces.

Example 21 (Coin tosses)**.** Let $\Omega = \{0,1\}^n$ (where $n = 1, 2, \ldots$ is fixed), $\mathcal{F} = \mathcal{P}(\Omega)$ and $\mathbb{P}(\omega) = \frac{1}{|\Omega|} = 2^{-n}$ for all $\omega \in \Omega$. This probability space describes all possible tosses of n coins (or sequences of length n of tosses of the same coin). Here $\mathbb{P}$ is the uniform probability measure on the set of n coin tosses, which means that the coin is *fair* (i.e. each side of the coin is always equiprobable). If we encode "heads"$= 0$ and "tails"$= 1$ for a single coin flip, then the event "first toss is heads" would correspond to the event $\{\omega = (\omega_1, \ldots, \omega_n) \in \Omega : \omega_1 = 0\}$.

Example 22 (Uniform distribution)**.** Let $\Omega = [0,1]$, $\mathcal{F} = \mathcal{B}$ the Borel σ-algebra and $\mathbb{P}$ the Lebesgue measure restricted to Ω. Then $\mathbb{P}(\Omega) = 1 - 0 = 1$, so this is a probability measure. It picks samples from $[0,1]$ uniformly at random, hence it is called the *uniform probability measure*. Observe that the probability of an occurrence of any singleton $\{x\}$ is zero, since it is a zero Lebesgue-measurable set. However, if eg. $A = [0, \frac{1}{3}] \cup [\frac{2}{3}, 1]$, then $\mathbb{P}(A) = \frac{1}{3} - 0 + 1 - \frac{2}{3} = \frac{2}{3}$.

Example 23 (Poisson distribution)**.** Let $\Omega = \{0, 1, 2, \ldots\}$, $\mathcal{F} = \mathcal{P}(\Omega)$ and $\mathbb{P}$ defined by $\mathbb{P}(n) = e^{-\lambda}\frac{\lambda^n}{n!}$ for $n \in \Omega$, where $\lambda > 0$ is a fixed parameter. From the Taylor series expansion of the exponential function, we see that $\sum_{n=0}^{\infty} \frac{\lambda^n}{n!} = e^{\lambda}$, hence $\mathbb{P}(\Omega) = \sum_{n=0}^{\infty} \mathbb{P}(n) = 1$. Therefore, $\mathbb{P}$ is a probability measure on Ω. More generally, for any event $A \in \mathcal{P}(\Omega)$, we have $\mathbb{P}(A) = \sum_{n \in A} \mathbb{P}(n)$. The measure $\mathbb{P}$ defines *Poisson distribution*, where the parameter λ is interpreted as a frequency at which some event occurs, and $\mathbb{P}(n)$ is the probability that the event occurs n times. For example, the Poisson distribution might model how many times a lightbulb will likely need replacement within ten years. In that case, λ would be the rate at which the bulb breaks. For this model, some independence and stationarity assumptions for the events in time are required; we will come back to these as the course progresses.

1.3 Random variables

Intuitively, any numerical quantity associated with the outcome of a random experiment is called a random variable. Formally, *random variables* are just measurable functions (or more general measurable mappings) defined on a probability space.

Definition 24 (Random variable)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let $\mathbb{R}$ be equipped with the Borel σ-algebra $\mathcal{B}$. Then a function $X : \Omega \to \mathbb{R}$ is a *random variable* if it is measurable, i.e. $X^{-1}(B) \in \mathcal{F} \quad \forall B \in \mathcal{B}$. More generally, any measurable mapping $X : \Omega \to \mathcal{S}$, where $\mathcal{S}$ is some set (called the *state space*) equipped with a σ-algebra $\mathcal{A}$, is called a random variable.

Remark 25. The definition of a random variable states that for every $B \in \mathcal{B}$, the event $\{\omega \in \Omega : X(\omega) \in B\}$ is measurable. Thus we can define the probability $\mathbb{P}\left(\{\omega : X(\omega) \in B\}\right)$, which we will briefly denote just by $\mathbb{P}(X \in B)$. This sheds some light to the intuition that the state space Ω does usually not play a very important role in probability.

Remark 26. Since the Borel σ-algebra $\mathcal{B}$ can be generated by various families of intervals, we get the following equivalent characterizations for real-valued random variables: for all $a, b \in \mathbb{R}$, $a < b$:

- $X^{-1}\left((a, b)\right) \in \mathcal{F}$
- $X^{-1}\left([a, b]\right) \in \mathcal{F}$
- $X^{-1}\left((-\infty, b]\right) \in \mathcal{F}$

Next, we define generally what is meant by the *probability distribution* or *law* of a random variable. This concept will tell how likely it is for a random variable X to fall in a given set.

Definition 27 (Law of a random variable)**.** Let $X : \Omega \to \mathcal{S}$ be a random variable. Then the *law* of X is the measure μ on the space $(\mathcal{S}, \mathcal{A})$ defined by

$$\mu(A) := \mathbb{P}\left(\{\omega : X(\omega) \in A\}\right) = \mathbb{P}(X \in A)$$

for all $A \in \mathcal{A}$. In other words, the law μ is the *pushforward* measure of $\mathbb{P}$ under the mapping X.

Remark 28. It is a straightforward exercise to show that the law μ is a probability measure on $\mathcal{S}$.

Remark 29. Sometimes, the terms *law* and *distribution* are used interchangeably. In this course, we usually understand the term *distribution* associated with a law of a random variable with real values, so that a (cumulative) *distribution function* can be defined.

Example 30 (The first outcome of a series of coin tosses)**.** We recall our setting in Example 21: That is, $\Omega = \{0, 1\}^n$, $\mathcal{F} = \mathcal{P}(\Omega)$ and $\mathbb{P}(\omega) = \frac{1}{|\Omega|} = 2^{-n}$ for all $\omega \in \Omega$. Now we define a random variable $X : \Omega \to \mathbb{R}$ by setting

$$X(\omega) = X(\omega_1, \ldots, \omega_n) = \omega_1 = \begin{cases} 0 & \text{if} \quad \text{heads} \\ 1 & \text{if} \quad \text{tails} \end{cases} .$$

The law of this random variable is $\mu = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_1$, since each of the two possible outcomes of the random variable X occur with probability $1/2$.

Example 31 (Binomial distribution)**.** We continue from the setting of the previous example, but this time we define the random variable as $X(\omega) = \omega_1 + \cdots + \omega_n$. Assuming again that the coin is fair, an outcome $X = k$ means that tails occur exactly k times in the series of n coin tosses. Therefore, this is the familiar setting of n independent success-failure experiments, thus following the binomial distribution with parameter $p = \frac{1}{2}$ (we will come back to independence formally and in high detail in Chapter 2). To recall, the general binomial distribution $B(n,p)$ with parameters $n \in \mathbb{N}$ and $p \in [0,1]$ is defined by the probability mass function

$$\mathbb{P}(k) = \binom{n}{k} p^k (1-p)^{n-k}, \qquad k = 0, \ldots, n.$$

Consequently, the law of the random variable X is defined as

$$\mathbb{P}(X = k) = \left(\frac{1}{2}\right)^n \binom{n}{k} = \left(\frac{1}{2}\right)^n \frac{n!}{k!(n-k)!}, \qquad k = 0, \ldots, n.$$

That is, $\mu = \sum_{k=0}^{n} \left(\frac{1}{2}\right)^n \binom{n}{k} \delta_k$.

Example 32 (Uniformly distributed random variable)**.** We return to the setting of Example 22. That is, consider $\Omega = [0,1]$, $\mathcal{F} = \mathcal{B}$ and $\mathbb{P}$ the Lebesgue measure on Ω. Define a random variable $U : \Omega \to \mathbb{R}$ by $U(\omega) = \omega$. Then the law of U is given by $\mu(B) = \mathbb{P}(U \in B) = \mathbb{P}(\omega \in B) = \mathbb{P}(B)$ for $B \in \mathcal{B}$. Thus, the law of this random variable is just the Lebesgue measure $d\mu(x) = dx$. In particular, if $B = [0,a]$ for $a \in \Omega$, then $\mu(B) = \mathbb{P}(U \leq a) = \mathbb{P}([0,a]) = a$. This is our first example of a distribution function, which more generally describes a law on $\mathbb{R}$.

Definition 33 (Distribution function)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $X : \Omega \to \mathbb{R}$ a random variable with law μ. The function $F : \mathbb{R} \to [0,1]$, $F(x) = \mathbb{P}(X \leq x) = \mu((-\infty, x])$ is the (cumulative) *distribution function* of X.

Observe that a distribution function can be assigned to any real-valued random variable X. Moreover, since the Borel σ-algebra can be generated by the half-open intervals $(-\infty, a]$, it actually follows that the distribution function fully describes the law of X (Theorem 36 below). Next, we state and prove some key properties of the distribution function.

Proposition 34 (Properties of a distribution function)**.** *Let F be a distribution function of a real-valued random variable X with law μ. Then the following properties hold:*

(i) F is an increasing function (i.e. $x \leq y \Longrightarrow F(x) \leq F(y)$).

(ii) $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$.

(iii) F is a càdlàg *function, i.e. it is right-continuous with left limits: for every $a \in \mathbb{R}$, $\lim_{x\to a+} F(x) = F(a)$ and $\lim_{x\to a-} F(x)$ exists.*

Proof. (i) Let $x \leq x'$. Then $(-\infty, x] \subset (-\infty, x']$, and hence by monotonicity of the measure μ, we deduce $F(x) = \mu((-\infty, x]) \leq \mu((-\infty, x']) = F(x')$.

(ii) First, let $(x_n)_{n=0}^{\infty}$ be a strictly increasing sequence of real numbers such that $x_n \nearrow \infty$ as $n \to \infty$. Define $A_n := (-\infty, x_n]$. Then $A_n \subset A_{n+1}$ for all $n \in \mathbb{N}$. We note that $\bigcup_{n=0}^{\infty} A_n = \mathbb{R}$, and thus Lemma 15 yields $F(x_n) = \mu(A_n) \xrightarrow[n\to\infty]{} \mu(\mathbb{R}) = 1$ since μ is a probability measure. Then, let $(x_n)_{n=0}^{\infty}$ be a strictly decreasing sequence of real numbers such that $x_n \searrow -\infty$ as $n \to \infty$. Now, the sequence of sets $(A_n)_{n=0}^{\infty}$ is decreasing with $\bigcap_{n=0}^{\infty} A_n = \emptyset$, and thus we deduce applying Corollary 16 that $F(x_n) = \mu(A_n) \xrightarrow[n\to\infty]{} \mu(\emptyset) = 0$.

(iii) For the right-continuity, let us develop the idea we just used to show $\lim_{x\to-\infty} F(x) = 0$. That is, fix $x \in \mathbb{R}$ and let $x_n \searrow x$ where $(x_n)_{n=0}^{\infty}$ is strictly decreasing. Consider again the sets $A_n := (-\infty, x_n]$. Now $\bigcap_{n=0}^{\infty} A_n = (-\infty, x]$, and an application of Corollary 16 gives $F(x_n) = \mu(A_n) =\xrightarrow[n\to\infty]{} \mu((-\infty, x]) = F(x)$. Existence of the left-limits is a bit more delicate. Let us anyway continue with the same approach. Thus, let $x_n \nearrow x$ as $n \to \infty$, where $(x_n)_{n=0}^{\infty}$ is strictly increasing, and $A_n := (-\infty, x_n]$. Now observe that $\bigcup_{n=0}^{\infty} A_n = (-\infty, x)$. Thus, again by Lemma 15, $F(x_n) \xrightarrow[n\to\infty]{} \mu((-\infty, x))$. Hence, the left limit exists, but if $\mathbb{P}(X = x) = \mu(\{x\}) > 0$, then $\lim_{x_n\to x-} F(x_n) < F(x)$ and there is thus a positive jump at x. □

Example 35 (Distribution function of a coin toss)**.** We continue the setting of Example 30: $\Omega = \{0,1\}^n$, $\mathcal{F} = \mathcal{P}(\Omega)$ and $\mathbb{P}(\omega) = \frac{1}{|\Omega|} = 2^{-n}$, and define

$$X(\omega) = \omega_1 = \begin{cases} 0 & \text{if} \quad \text{heads} \\ 1 & \text{if} \quad \text{tails} \end{cases} .$$

We already showed that the law of X is $\mu = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_1$. Then the distribution function of X is

$$F(x) = \mu((-\infty, x]) = \begin{cases} 0 & \text{if} \quad x < 0 \\ \frac{1}{2} & \text{if} \quad 0 \le x < 1 \\ 1 & \text{if} \quad x \ge 1. \end{cases}$$

We see that this function is increasing, having two jumps at $x = 0$ and $x = 1$ respectively, and outside these points constant. In particular, the function is continuous outside the two aforementioned points.

Remarkably, a converse result for Proposition 34 holds. That is, if a function satisfying the conditions of Proposition 34 is given, it is always a distribution function of some random variable.

Theorem 36. *Let $F : \mathbb{R} \to [0,1]$ be a function satisfying the conditions (i)-(iii) of Proposition 34. Then there exists a unique measure μ on $(\mathbb{R}, \mathcal{B})$ such that $F(x) = \mu((-\infty, x])$. Moreover, there exists a random variable X on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ such that $F(x) = \mathbb{P}(X \le x)$.*

Proof. Given F, we would like to define an "inverse" of F which satisfies $u \le F(x)$ if and only if $F^{-1}(u) \le x$ for all $u \in [0,1]$. However, since F is not necessarily *strictly* increasing, we cannot just invert it in a traditional sense. Instead, let us define a generalized "càdlàg inverse" by setting $F^{-1}(u) := \sup\{y \in \mathbb{R} : F(y) < u\}$ for all $u \in [0,1]$. Let us show that this function satisfies the above inversion property. Thus, let $x \in \mathbb{R}$ and $u \in [0,1]$ such that $u \le F(x)$. Now, since F is increasing, we also have $u \le F(y)$ for every $y \ge x$. This implies by definition that $F^{-1}(u) \le x$.

To show the reverse implication, assume on the contrary that u and x are such that $u > F(x)$. Now by right-continuity of F, there exists an $\epsilon > 0$ such that $F(x + \epsilon) < u$.

Then, we obtain $F^{-1}(u) = \sup\{y \in \mathbb{R} : F(y) < u\} \geq x + \epsilon > x$. Hence, $F^{-1}(u) \leq x$ implies $u \leq F(x)$.

Now the construction of the measure μ and the random variable X is just a simple matter of observations. First, let $\Omega = [0,1]$ with the Borel σ-algebra $\mathcal{B}$ and let $\mathbb{P}$ be the Lebesgue measure restricted to Ω. Now $u \mapsto F^{-1}(u)$ defines a random variable $X : \Omega \to \mathbb{R}$: for any $x \in \mathbb{R}$, $\{u : X(u) \leq x\} = \{u : F^{-1}(u) \leq x\} = \{u : u \leq F(x)\} = [0, F(x)] \in \mathcal{B}$. Moreover, $\mathbb{P}(X \leq x) = \mathbb{P}(F^{-1}(u) \leq x) = \mathbb{P}(u \leq F(x)) = F(x)$. That is, F is the distribution function of X. Let μ be its law.

It remains to show that μ is the unique measure on $(\mathbb{R}, \mathcal{B})$ satisfying $\mu((-\infty, x]) = F(x)$. If there is another measure μ' satisfying $\mu'((-\infty, x]) = F(x)$, then in particular $\mu((-\infty, x]) = \mu'((-\infty, x])$ for all $x \in \mathbb{R}$. Recall that $\mathcal{B}$ is generated by the set of intervals $\{(-\infty, x] : x \in \mathbb{R}\}$, which form a π-system, i.e. it is closed under intersections (see Definition 67 or the exercises of the course). Thus it follows from Dynkin's lemma (Proposition 72 stated and proven later) that μ and μ' agree on every Borel set, showing the claim. A similar argument is used in detail later in Remark 79. □

Next, we study the important special case of continuous distribution.

Definition 37. Let X be a real-valued random variable and μ its law. Then X has *continuous distribution* if $d\mu(x) = f(x)dx$ for some function $f : \mathbb{R} \to \mathbb{R}$, called a (probability) *density function*. That is, for all $B \in \mathcal{B}$, $\mu(B) = \int_B f(x)dx$. One also says that X is a *continuous random variable*.

Remark 38. From the definition of the density function, it follows that $f \geq 0$ and $\int_{\mathbb{R}} f(x)dx = \mu(\mathbb{R}) = 1$. Moreover, if f is continuous, then the distribution function F is differentiable and $F' = f$. This is a consequence of the fundamental theorem of calculus, since we may write $F(a) = \mu((-\infty, a]) = \int_{-\infty}^{a} f(x)dx$. Conversely, if F has a continuous derivative, then F' is a density function. Note that, in general, we do not assume f is continuous. The reason why we call the *distribution* continuous is rather that, in this case, the measure μ is absolutely continuous with respect to the Lebesgue measure, and also the cumulative distribution function F is absolutely continuous. Hence F has a derivative almost everywhere, which in turn coincides with f a.e.

Example 39 (Square of a uniform random variable)**.** Let U be a uniform random variable on $[0,1]$ and define $X = U^2$. We want to compute the law of the random variable X. Denote the law of X by μ and the distribution function by F. For $x \in [0,1]$, we have $F(x) = \mathbb{P}(X \leq x) = \mathbb{P}(-\sqrt{x} \leq U \leq \sqrt{x}) = \mathbb{P}(U \leq \sqrt{x}) = \sqrt{x}$, which is differentiable when $x > 0$. Now $F'(x) = \frac{1}{2\sqrt{x}}$, which is continuous for $x > 0$. Thus, X has a density $F'(x)$, and its law is $d\mu(x) = \frac{1}{2\sqrt{x}} \mathbb{1}_{x \in (0,1]} dx$.

1.4 Expectation

Intuitively, the *expectation* of a random variable corresponds to the ”average value” or the ”expected result” of a random variable or experiment, respectively. From earlier studies in probability, the reader is probably familiar with the following special cases.

- The expectation of a *discrete* random variable X is $\mathbb{E}(X) = \sum_x x\mathbb{P}(X = x)$, where the summation goes over all the possible (countable) values of the random variable X.

- The expectation of a *continuous* random variable X is $\mathbb{E}(X) = \int_{-\infty}^{\infty} xf(x)dx$, where f is a density function for X.

Using measure theory, one definition suffices:

Definition 40 (Expectation)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let X be a random variable defined on that space. Then the *expectation* of X is

$$\mathbb{E}(X) := \int_{\Omega} X(\omega) d\mathbb{P}(\omega),$$

provided the integral of X is well-defined (that is, when $X \geq 0$ or $\int_{\Omega} |X| \, d\mathbb{P} < \infty$).

Let us now list without proof some basic properties of expectation, some of which are direct consequences from the definition, some are general properties of function spaces, and others will be proven later in these notes.

Proposition 41 (Properties of the expectation)**.** *The following properties hold for the expectation.*

(i) $\mathbb{E}(\mathbb{1}_A) = \mathbb{P}(A)$ *for any* $A \in \mathcal{F}$.

(ii) The expectation is linear*: for random variables* X, Y *and for* $a, b \in \mathbb{R}$*, we have* $\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y)$.

(iii) If $X \geq 0$*, then also* $\mathbb{E}(X) \geq 0$*. Moreover, if* $X \geq 0$ *and* $\mathbb{E}(X) = 0$*, then* $X = 0$ almost surely*, i.e.* $\mathbb{P}$*-almost everywhere.*

(iv) $\mathbb{E}(|XY|)^2 \leq \mathbb{E}(X^2)\mathbb{E}(Y^2)$ *for any two real-valued random variables* X, Y *(Cauchy-Schwarz inequality).*

(v) More generally, let $\|X\|_p = \mathbb{E}\left(|X|^p\right)^{\frac{1}{p}}$ *be the* L^p*-norm of* X*. Then* $\mathbb{E}(|XY|) \leq \|X\|_p \|Y\|_q$ *whenever* $\frac{1}{p} + \frac{1}{q} = 1$ *with* $p, q \in [1, \infty]$ *(Hölder's inequality). Here* $\|X\|_\infty = \inf\{M : \mathbb{P}(X > M) = 0\}$ *is the essential supremum of* X.

(vi) For all $p \in [1, \infty]$*,* $\|X + Y\|_p \leq \|X\|_p + \|Y\|_p$ *(Minkowski's inequality).*

(vii) Fundamental convergence theorems of the general measure theory hold with their usual assumptions. For example, the dominated convergence theorem reads: if $X_n \xrightarrow[n\to\infty]{} X$ *a.s.,* $|X_n| \leq Y$ *and* $\mathbb{E}(Y) < \infty$*, then* $\mathbb{E}(X_n) \xrightarrow[n\to\infty]{} \mathbb{E}(X)$*. In particular, if* $(X_n)_{n\geq 0}$ *is a bounded sequence, the DCT can be applied by the finiteness of the measure* $\mathbb{P}$.

(viii) Let $\varphi : \mathbb{R} \to \mathbb{R}$ *be a convex function. Then* $\varphi\left(E(X)\right) \leq \mathbb{E}\left(\varphi(X)\right)$ *(Jensen's inequality).*

Jensen's inequality has the following important corollary.

Corollary 42. *Let* $0 < p < q$ *and assume* $E\left(|X|^q\right) < \infty$*. Then also* $E\left(|X|^p\right) < \infty$.

Proof. Since $q/p > 1$, the function $\varphi(x) = x^{\frac{q}{p}}$ is convex for $x \geq 0$. Thus by Jensen´s inequality,

$$\mathbb{E}\left(|X|^p\right)^{\frac{q}{p}} = \varphi(\mathbb{E}\left(|X|^p\right)) \leq \mathbb{E}\left(\varphi(|X|^p)\right) = \mathbb{E}(|X|^q) < \infty,$$

from which the claim follows. □

Example 43. Let X be a $\mathbb{N}$-valued random variable, whose law is given by $\mathbb{P}(X = n) = \frac{1}{C}\frac{1}{1+n^2}$, where $C > 0$ is a normalizing constant, more precisely $C := \sum_{n=0}^{\infty}\frac{1}{1+n^2}$. Then $\mathbb{E}(X) = \sum_{n=0}^{\infty} n\mathbb{P}(X = n) = \infty$. On the other hand, for any $0 \leq p < 1$, $\mathbb{E}(X^p) = \sum_{n=0}^{\infty} n^p\mathbb{P}(X = n) < \infty$.

Let us now show one of the simplest, yet an important, tail bound for a law of a random variable, given it has the expectation.

Proposition 44 (Markov's inequality)**.** *Let $X \geq 0$ be a random variable with expectation $\mathbb{E}(X) < \infty$, and $M > 0$ a constant. Then $\mathbb{P}(X \geq M) \leq \frac{\mathbb{E}(X)}{M}$.*

Proof. We simply write $\mathbb{E}(X) = \mathbb{E}\left(X\mathbb{1}_{X<M}\right)+\mathbb{E}\left(X\mathbb{1}_{X\geq M}\right) \geq \mathbb{E}\left(X\mathbb{1}_{X\geq M}\right) \geq M\mathbb{E}\left(\mathbb{1}_{X\geq M}\right) = M\mathbb{P}(X \geq M)$, where we used the monotonicity and the linearity of expectation. □

The reader might wonder why the expectation of a random variable can be written as an integral w.r.t. its law, or how to express the expectation of a random variable using its law in general. In the general setup, this involves a change of variable, which is detailed by the following proposition. In principle, the following very important result is the most general formula to compute or estimate expectations, which encompasses all the well-known "simple" cases.

Proposition 45 (Change of measure and variable)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $\mathcal{A}$ a σ-algebra on a set $\mathcal{S}$ and $X : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathcal{S}, \mathcal{A})$ a random variable with law μ. Furthermore, let $g : \mathcal{S} \to \mathbb{R}$ be measurable with respect to the sigma-algebras $\mathcal{B}$ and $\mathcal{A}$. Then:*

(a) $\mathbb{E}(|g(X)|) < \infty$ if and only if $\int_{\mathcal{S}} |g(s)|\, d\mu(s) < \infty$.

(b) If either of the properties in (a) hold, then $\mathbb{E}(g(X)) = \int_{\mathcal{S}} g(s)d\mu(s)$.

Proof. The proof strategy follows approximation of g by step functions, which is typical in the measure theory literature. The proof has four steps.

1) Let $g = \mathbb{1}_A$ for some $A \in \mathcal{A}$. Then $\mathbb{E}(g(X)) = \mathbb{E}(\mathbb{1}_{X\in A}) = \mathbb{P}(X \in A) = \mu(A) = \int_{\mathcal{S}} \mathbb{1}_A d\mu = \int_{\mathcal{S}} g d\mu < \infty$. Hence, both (a) and (b) hold for this choice of g.

2) Assume $g \geq 0$ is a step function, that is, $g = \sum_{k=0}^{n} a_k\mathbb{1}_{A_k}$ for some $n \in \mathbb{N}$, $a_k \in \mathbb{R}_+$ and $A_k \in \mathcal{A}$. Now using the linearity of expectation, we deduce $\mathbb{E}(g(X)) = \sum_{k=0}^{n} a_k\mathbb{E}(\mathbb{1}_{X\in A_k}) = \sum_{k=0}^{n} a_k \int_{\mathcal{S}} \mathbb{1}_{A_k}d\mu = \int_{\mathcal{S}} \sum_{k=0}^{n} a_k\mathbb{1}_{A_k}d\mu = \int_{\mathcal{S}} g d\mu < \infty$. We exploited step 1) above.

3) Assume $g \geq 0$ is measurable. Now we approximate g using the following step functions: let $g_n := \sum_{k=0}^{n2^2-1} \frac{k}{2^n}\mathbb{1}_{\frac{k}{2^n}\leq g\leq\frac{k+1}{2^n}} + n\mathbb{1}_{g\geq n}$. It is easy to see that this sequence of functions is increasing, and in fact, $g_n \nearrow g$ as $n \to \infty$ (the latter is proven in most classical courses of measure theory, those in doubt may try to prove it). Now the sequence of functions $(g_n)_{n\geq 0}$ satisfies the assumptions of the monotone convergence theorem, and hence we deduce $\mathbb{E}(g(X)) = \lim_{n\to\infty} \mathbb{E}(g_n(X))$. On the other hand, from 2) we deduce that $\mathbb{E}(g_n(X)) = \int_{\mathcal{S}} g_n d\mu < \infty$, and again by the MCT, $\lim_{n\to\infty} \int_{\mathcal{S}} g_n d\mu = \int_{\mathcal{S}} g d\mu$. Hence, $\mathbb{E}(g(X)) = \int_{\mathcal{S}} g d\mu$, which gives (b) and (a).

4) Finally, assume g is a measurable function. Define the positive and the negative part of g, respectively, by setting $g_+ = g\mathbb{1}_{g\geq 0}$ and $g_- = -g\mathbb{1}_{g<0}$. Observe that both g_+ and g_- are measurable, since due to the measurability of g, the sets $\{s \in \mathcal{S} : g(s) \geq 0\}$ and $\{s \in \mathcal{S} : g(s) < 0\}$ are measurable, and so are their indicator functions. Moreover, $g_+ \geq 0$ and $g_- \geq 0$, and $g = g_+ - g_-$ as well as $|g| = g_+ + g_-$. The final claims (a) and (b) then follow by applying 3) to g_+ and g_-, and finally linearity.

□

Remark 46. The formulae for the expectations of the discrete and continuous random variables, respectively, follow now from the above proposition:

- If X is discrete, its law is given by a (possible countably infinite) linear combination of Dirac masses, $\mu = \sum_a \delta_a \mathbb{P}(X = a)$. Therefore,

  $$\mathbb{E}(X) = \int_{\mathbb{R}} x d\mu(x) = \int_{\mathbb{R}} x \sum_a \mathbb{P}(X = a) d\delta_a = \sum_a \mathbb{P}(X = a) \int_{\mathbb{R}} x d\delta_a = \sum_a a\mathbb{P}(X = a).$$

  Above we used the Fubini's theorem, which will be stated later, to exchange the sum and the integration. As a rule of thumb, in sums (or more general integrals) with positive terms, the order of summation/ integration can be exchanged freely.

- If X is continuous with density f, then X has law $d\mu(x) = f(x)dx$, so

  $$\mathbb{E}(X) = \int_{\mathbb{R}} x d\mu(x) = \int_{\mathbb{R}} x f(x) dx.$$

Proposition 45 often makes computations of expectations of measurable transforms of random variables more straightforward, as the following simple example demonstrates.

Example 47. Let U be uniformly distributed on $[0, 1]$. Let us compute its expectation $\mathbb{E}(U^2)$. We recall from Example 39 that U^2 has density $f(x) = \frac{1}{2\sqrt{x}} \mathbb{1}_{(0,1]}$. Thus, $\mathbb{E}(U^2) = \int_{-\infty}^{\infty} x f(x) dx = \int_0^1 \frac{\sqrt{x}}{2} dx = \frac{1}{3}$.

However, Proposition 45 provides a way where we do not need to solve the density of U^2. Indeed, let h be a density of U, for which we can choose $h(x) = \mathbb{1}_{(0,1]}$. Then $\mathbb{E}(U^2) = \int_{-\infty}^{\infty} s^2 d\mu(s) = \int_0^1 s^2 ds = \frac{1}{3}$.

1.5 Variance and covariance

In the previous example, we computed the *second moment* of a uniform random variable, which is closely related to its *variance*. While the expectation $m := \mathbb{E}(X)$ represents the "average" of a random variable X, its variance describes its fluctuations around m.

Definition 48 (Variance)**.** Let X be a real-valued random variable with finite second moment, i.e. $\mathbb{E}(X^2) < \infty$. Then the *variance* of X is $\mathrm{Var}(X) := \mathbb{E}\left((X - m)^2\right)$, where $m := \mathbb{E}(X)$ (which is well-defined since $\mathbb{E}(|X|) < \infty$ by Corollary 42).

Remark 49. We have $\mathrm{Var}(X) := \mathbb{E}\left((X - m)^2\right) = \mathbb{E}(X^2) - 2m\mathbb{E}(X) + m^2 = \mathbb{E}(X^2) - \mathbb{E}(X)^2$ which is useful in practice to compute the variance. In particular, $\mathrm{Var}(X) < \infty$ if and only if $\mathbb{E}(X^2) < \infty$.

Remark 50 (Standard deviation)**.** The *standard deviation* of X is defined as $\sigma := \sqrt{Var(X)}$ provided the variance exists.

Next, we give an important consequence of Markov's inequality. While the former uses only knowledge of the expectation, the following one is useful in when the random variable has finite second moment.

Lemma 51 (Chebyshev's inequality)**.** *Let X be a random variable with $\mathbb{E}(X^2) < \infty$. Then for any $a > 0$,*

$$\mathbb{P}(|X - m| \geq a) \leq \frac{\operatorname{Var}(X)}{a^2}$$

where $m := \mathbb{E}(X)$.

Equivalently, for any $k > 0$, $\mathbb{P}(|X - m| \geq k\sigma) \leq \frac{1}{k^2}$.

Proof. By Markov's inequality, we deduce $\mathbb{P}(|X - m| \geq a) = \mathbb{P}((X-m)^2 \geq a^2) \leq \frac{\mathbb{E}\big((X-m)^2\big)}{a^2}$. □

Next, we generalize the notion of variance to quantify how two random variables are *correlated.*

Definition 52 (Covariance)**.** Let X and Y be real random variables, defined on the same probability space, such that $\mathbb{E}(X^2) < \infty$ and $\mathbb{E}(Y^2) < \infty$. Denote $m_X := \mathbb{E}(X)$ and $m_Y := \mathbb{E}(Y)$. The *covariance* of X and Y is $\operatorname{Cov}(X, Y) := \mathbb{E}\left((X - m_X)(Y - m_Y)\right)$.

Remark 53. In the spirit of Remark 49, we find a practical formula for the covariance: $\operatorname{Cov}(X, Y) := \mathbb{E}\left((X - m_X)(Y - m_Y)\right) = \mathbb{E}(XY) - m_X\mathbb{E}(Y) - m_Y\mathbb{E}(X) + m_X m_Y = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$. Note that by Cauchy-Schwarz inequality, $\mathbb{E}(|XY|) \leq \sqrt{\mathbb{E}(X^2)\mathbb{E}(Y^2)}$ which implies that the covariance is finite if the second moments are finite.

Remark 54. The *correlation coefficient* of X and Y is defined by

$$\rho_{XY} := \frac{\operatorname{Cov}(X, Y)}{\sqrt{\operatorname{Var}(X)\operatorname{Var}(Y)}},$$

and by Cauchy-Schwarz, $\rho_{XY} \in [-1, 1]$. Based on the value of the covariance, we distinguish the following terminology.

- If $\operatorname{Cov}(X, Y) \geq 0$, then X and Y are *positively correlated.* That is, on average, $(X - m_X)(Y - m_Y) \geq 0$.
- If $\operatorname{Cov}(X, Y) \leq 0$, then X and Y are *negatively correlated.*
- If $\operatorname{Cov}(X, Y) = 0$, then X and Y are *uncorrelated.*

The most important case of uncorrelated random variables occur when X and Y are *independent.*

Chapter 2

Independence

Definition 55 (Conditional probability)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $B \in \mathcal{F}$ such that $\mathbb{P}(B) > 0$. Then the *conditional probability* of A given B, or the probability of A conditional on B, is defined by $\mathbb{P}(A|B) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$ for all $A \in \mathcal{F}$.

Remark 56. It is an easy exercise to show that $A \mapsto \mathbb{P}(A|B)$ defines a probability measure on $(\Omega, \mathcal{F})$ for a given B such that $\mathbb{P}(B) > 0$.

Example 57 (Memorylessness of an exponential random variable)**.** Let $X \sim \mathrm{Exp}(\lambda)$, i.e. X is an *exponentially distributed* random variable with a parameter $\lambda > 0$. That is, X is a positive valued (continuous) random variable with law given by $\mathbb{P}(X > x) = e^{-\lambda x}$ for all $x \geq 0$, or equivalently, with distribution function $F(x) = (1 - e^{-\lambda x})\mathbb{1}_{x \geq 0}$. Let us fix $t > 0$ and condition on the set $B = \{X > t\}$, which has a positive measure by definition. The claim is that $X - t \sim \mathrm{Exp}(\lambda)$ conditional on B. Indeed,

$$\mathbb{P}(X - t > x | X > t) = \frac{\mathbb{P}(X > t + x, X > t)}{\mathbb{P}(X > t)} = \frac{\mathbb{P}(X > t + x)}{\mathbb{P}(X > t)} = \frac{e^{-\lambda(t+x)}}{e^{-\lambda t}} = e^{-\lambda x}.$$

This is commonly referred as the *memoryless property* of the exponential distribution.

Definition 58 (Independence of events)**.**

- Two events A and B are *independent* if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. If $\mathbb{P}(B) > 0$, this is equivalent to $\mathbb{P}(A|B) = \mathbb{P}(A)$.
- More generally, events $A_1, A_2, \ldots, A_n$ are independent if for every $k \in \mathbb{N}$ and distinct $i_1, \ldots, i_k \in \{1, \ldots, n\}$,
$$\mathbb{P}\left(\bigcap_{j=1}^{k} A_{i_j}\right) = \prod_{j=1}^{k} \mathbb{P}(A_{i_j}).$$

Example 59. If $A_1, A_2, \ldots, A_n$ are independent, then also $A_1^c, A_2^c, \ldots, A_n^c$ are independent. For example, when $n = 2$, we have $\mathbb{P}(A^c \cap B^c) = \mathbb{P}((A \cup B)^c) = 1 - \mathbb{P}(A \cup B) = 1 - (\mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)) = 1 - \mathbb{P}(A) - \mathbb{P}(B) + \mathbb{P}(A)\mathbb{P}(B) = (1 - \mathbb{P}(A))(1 - \mathbb{P}(B)) = \mathbb{P}(A^c)\mathbb{P}(B^c)$. The general case requires some more writing and is left as an additional exercise.

The very rough meaning of independence is that, knowing an event B occurs does not give information on another event A. Much more generally, this is formalized in the following definition via σ-algebras, which can be viewed as the general notion of information in probability theory.

Definition 60 (Independence of σ-algebras)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $\mathcal{F}_i \subset \mathcal{F}$ sub-σ-algebras indexed by a (possibly infinite) set I. Then the sigma-algebras $(\mathcal{F}_i)_{i \in I}$ are independent if for all $k \in \mathbb{N}$, all distinct indices $i_1, \dots, i_k \in I$ and all $A_1 \in \mathcal{F}_{i_1}$, $A_2 \in \mathcal{F}_{i_2}, \dots$, $A_k \in \mathcal{F}_{i_k}$, it holds

$$\mathbb{P}\left(\bigcap_{j=1}^{k} A_j\right) = \prod_{j=1}^{k} \mathbb{P}(A_j). \tag{2.1}$$

Remark 61 (Independence of an infinite collection of events)**.** The previous definition extends the definition of the independence of events in a natural way as follows: Let $(A_i)_{i \in I}$ be events indexed by an infinite set I. Then the events A_i are independent if for all $k \in \mathbb{N}$ and all distinct indices $i_1, \dots, i_k \in I$,

$$\mathbb{P}\left(\bigcap_{j=1}^{k} A_j\right) = \prod_{j=1}^{k} \mathbb{P}(A_j).$$

Remark 62. It is a straightforward additional exercise to show that two sets A and B are independent if and only if the σ-algebras $\{\emptyset, A, A^c, \Omega\}$ and $\{\emptyset, B, B^c, \Omega\}$ are independent. More generally, given n events $A_1, \dots, A_n$, then the events are independent if and only if the σ-algebras $\mathcal{F}_j := \{\emptyset, A_j, A_j^c, \Omega\}$ are independent. In particular, independence of sigma-algebras is a much more general property than independence of sets.

Finally, we study *independent random variables.* We are interested in the information available to us when observing random variables X_i, which then brings us back to the definition of independence of σ-algebras which these random variables generate.

Definition 63 (Sigma-algebra generated by a random variable)**.** Let $X : \Omega \to \mathcal{S}$ be a random variable, where $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space and $\mathcal{A}$ is a σ-algebra on $\mathcal{S}$. Then the *sigma-algebra generated by* X is defined as $\sigma(X) := X^{-1}(\mathcal{A}) = \{A' \in \mathcal{F} : A' = X^{-1}(A) \text{ for some } A \in \mathcal{A}\}$.

Now we define independence of real-valued random variables using the above concept, since it is the most important and widely used case. Generalizations to other cases are straightforward.

Definition 64 (Independence of random variables)**.** Let $(X_i)_{i \in I}$ be a family of real-valued random variables (where $\mathbb{R}$ is equipped with the Borel σ-algebra $\mathcal{B}$). We say that the random variables X_i are *independent* if the sigma-algebras generated by them are independent. That is, $\sigma(X_i) := X_i^{-1}(\mathcal{B}) = \{A \in \mathcal{F} : A = X^{-1}(B) \text{ for some } B \in \mathcal{B}\}$ are independent σ-algebras.

By combining Definitions 60 and 64, we find an alternative definition for independence of random variables, which is usually more practical for applications.

Definition 65 (Independence of random variables, 2nd version)**.** Let $(X_i)_{i \in I}$ be a family of real-valued random variables (where $\mathbb{R}$ is equipped with the Borel σ-algebra $\mathcal{B}$). We say the the random variables X_i are *independent* if for all $k \in \mathbb{N}$, all distinct indices $i_1, \dots, i_k \in I$ and all $B_1, \dots, B_k \in \mathcal{B}$,

$$\mathbb{P}\left(X_{i_1} \in B_1, \dots, X_{i_k} \in B_k\right) = \prod_{j=1}^{k} \mathbb{P}(X_{i_j} \in B_j).$$

Example 66 (Independence of two coin tosses)**.** We continue our study of coin tosses as in Example 35 and earlier. This time, let us fix $n = 2$, so we are only tossing the coin twice. Then $\Omega = \{(0,0),(0,1),(1,0),(1,1)\}$ and $\mathbb{P}$ is a probability measure on Ω. For $\omega = (\omega_1, \omega_2) \in \Omega$, define now $X_1(\omega) = \omega_1$ and $X_2(\omega) = \omega_2$. The sigma-algebra generated by X_1 is $\sigma(X_1) = \{X_1^{-1}(\emptyset), X_1^{-1}(\{0\}), X_1^{-1}(\{1\}), X_1^{-1}(\{0,1\})\} = \{\emptyset, \{(0,0),(0,1)\}, \{(1,0),(1,1)\}, \Omega\}$, and similarly, $\sigma(X_2) = \{\emptyset, \{(0,0),(1,0)\}, \{(0,1),(1,1)\}, \Omega\}$. Now the random variables X_1 and X_2 are independent, if for all $A_1 \in \sigma(X_1)$ and $A_2 \in \sigma(X_2)$, $\mathbb{P}(A_1 \cap A_2) = \mathbb{P}(A_1)\mathbb{P}(A_2)$. Let us focus on the most non-trivial subsets consisting of $\{(0,0),(0,1)\}, \{(1,0),(1,1)\}$ and $\{(0,0),(1,0)\}, \{(0,1),(1,1)\}$, respectively. For example, if $A_1 = \{(0,0),(0,1)\}$ and $A_2 = \{(0,1),(1,1)\}$, we have $\mathbb{P}(A_1) = \mathbb{P}(A_2) = \frac{1}{2}$ and $\mathbb{P}(A_1 \cap A_2) = \mathbb{P}((0,1)) = \frac{1}{4}$. Thus, we have $\mathbb{P}(A_1 \cap A_2) = \mathbb{P}(A_1)\mathbb{P}(A_2)$. The other cases are similar, hence the random variables X_1 and X_2 are independent by definition.

Note an important issue above: in fact, the definition of the uniform distribution $\mathbb{P}$ on Ω included the hidden assumptions that the coin was fair and the outcome of a coin toss was not affected by another tosses or the environment. These assumptions correspond to independence, so in fact, we *assumed* the independence, which was then rather an intrinsic property of the model which we rigorously verified *à posteriori*. This "intrinsic independence" of the (continuous) uniform distribution will be formalized in Example 83.

2.1 Sufficient conditions for independence

So far, we have given rather abstract definitions of independence, and the reader may wonder whether there are any simpler and more practical *sufficient conditions* for it. Recall that the Borel σ-algebra $\mathcal{B}$ is generated by semi-infinite closed intervals of the form $(-\infty, x]$. Next, we study whether identities of type $\mathbb{P}(X_1 \leq x_1, \ldots, X_n \leq x_n) = \mathbb{P}(X_1 \leq x_1) \cdots \mathbb{P}(X_n \leq x_n)$ in some sense characterize independence. This leads us back to an auxiliary result from measure theory called Dynkin's lemma.

Definition 67 (π-system)**.** Let Ω be a set and $\mathcal{A}$ a collection of its subsets. If $\mathcal{A}$ is closed under intersections, i.e. $A, B \in \mathcal{A} \Rightarrow A \cap B \in \mathcal{A}$, then it is called a *$\pi$-system.*

Example 68. The semi-infinite closed intervals of the form $(-\infty, x]$, $x \in \mathbb{R}$, form a π-system on $\Omega = \mathbb{R}$, see the exercises of the course. As an additional (simple) exercise, the reader is encouraged to show that the collection of open intervals (a, b) (with $a < b$) and the collection of finite unions of half-closed intervals of the form $(a_1, b_1] \cup \cdots \cup (a_n, b_n]$ are also π-systems, respectively.

Remark 69. Typically a π-system is a collection of "nice sets", on which it is sufficient to test some probabilistic property, such as independence.

Next, we present a slight generalization of σ-algebra, which together with π-systems form the basis of Dynkin's lemma.

Definition 70 (D-system)**.** Let Ω be a set and $\mathcal{D}$ a collection of its subsets. The collection $\mathcal{D}$ is called a *d-system* (or a Dynkin system or a λ-system) if the following three conditions hold:

(i) $\Omega \in \mathcal{D}$

(ii) $A, B \in \mathcal{D}$ and $A \subset B \Rightarrow B \setminus A \in \mathcal{D}$.

(iii) $A_n \in \mathcal{D}$ $(n = 1, 2, \dots)$ such that $A_n \subset A_{n+1} \Rightarrow \bigcup_{n=1}^{\infty} A_n \in \mathcal{D}$ (that is, $\mathcal{D}$ is closed under monotone set limits).

Remark 71. A σ-algebra is always a d-system. This is easily seen by noting that $B \setminus A = B \cap A^c$ and $\Omega = \emptyset^c$, whereas the third condition trivially holds for sets belonging to a sigma-algebra.

Now we finally present Dynkin's lemma (also known as the $\pi - \lambda$-theorem). Since the focus is in its probabilistic applications, the proof is omitted in these notes. We refer to measure theory materials for a proof.

Proposition 72 (Dynkin's lemma)**.** *Let $\mathcal{A}$ be a π-system and $\mathcal{D}$ a d-system such that $\mathcal{A} \subset \mathcal{D}$. Then $\sigma(\mathcal{A}) \subset \mathcal{D}$, where $\sigma(\mathcal{A})$ is the smallest σ-algebra containing $\mathcal{A}$.*

Let us now show some powerful applications of Dynkin's lemma to independence.

Proposition 73. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $\mathcal{A}_1, \dots, \mathcal{A}_n \subset \mathcal{F}$ π-systems such that $\Omega \in \mathcal{A}_i$ for all $i = 1, \dots, n$, and for all $A_i \in \mathcal{A}_i$, $(i = 1, \dots n)$,*

$$\mathbb{P}(A_1 \cap \dots \cap A_n) = \prod_{i=1}^{n} \mathbb{P}(A_i). \tag{2.2}$$

Then $\sigma(\mathcal{A}_1), \dots, \sigma(\mathcal{A}_n)$ are independent.

Proof. The proof is by induction. The first step is to define

$$\mathcal{D}_1 := \Big\{ A_1 \in \mathcal{F} : \forall\ A_2 \in \mathcal{A}_2, \dots, A_n \in \mathcal{A}_n, \quad \mathbb{P}(A_1 \cap A_2 \cap \dots \cap A_n) = \prod_{i=1}^{n} \mathbb{P}(A_i) \Big\}.$$

We claim that $\mathcal{D}_1$ is a d-system. Let us prove the three points in the definition of a d-system.

(i) By the assumption of the proposition, $\Omega \in \mathcal{A}_1 \subset \mathcal{D}_1$.

(ii) Let $A_1 \in \mathcal{D}_1$ and $B_1 \in \mathcal{D}_1$ such that $A_1 \subset B_1$. Then if $A_2 \in \mathcal{A}_2, \dots, A_n \in \mathcal{A}_n$, we have

$$\begin{aligned}\mathbb{P}((B_1 \setminus A_1) \cap A_2 \cap \dots \cap A_n) &= \mathbb{P}(B_1 \cap A_2 \cap \dots \cap A_n) - \mathbb{P}(A_1 \cap A_2 \cap \dots \cap A_n) \\ &= \mathbb{P}(B_1)\mathbb{P}(A_2) \dots \mathbb{P}(A_n) - \mathbb{P}(A_1)\mathbb{P}(A_2) \dots \mathbb{P}(A_n) \\ &= (\mathbb{P}(B_1) - \mathbb{P}(A_1))\mathbb{P}(A_2) \dots \mathbb{P}(A_n) \\ &= \mathbb{P}(B_1 \setminus A_1)\mathbb{P}(A_2) \dots \mathbb{P}(A_n).\end{aligned}$$

Hence, $B_1 \setminus A_1 \in \mathcal{D}_1$.

(iii) Let $A^{(k)} \in \mathcal{D}_1$ $(k = 1, 2, \dots)$ such that $A^{(k)} \subset A^{(k+1)}$. Now $\mathbb{P}(A^{(k)} \cap A_2 \cap \dots \cap A_n) = \mathbb{P}(A^{(k)})\mathbb{P}(A_2) \dots \mathbb{P}(A_n)$. We can then apply Lemma 15 to both of the sides of this equation, which gives $\mathbb{P}\left(\bigcup_{k=1}^{\infty} A^{(k)} \cap A_2 \cap \dots \cap A_n\right) = \mathbb{P}\left(\bigcup_{k=1}^{\infty} A^{(k)}\right) \mathbb{P}(A_2) \dots \mathbb{P}(A_n)$, showing that $\bigcup_{k=1}^{\infty} A^{(k)} \in \mathcal{D}_1$.

Now, since $\mathcal{A}_1$ is a π-system and $\mathcal{D}_1$ a d-system with $\mathcal{A}_1 \subset \mathcal{D}_1$, Dynkin's lemma implies that $\sigma(\mathcal{A}_1) \subset \mathcal{D}_1$. The next step is to repeat the above steps for

$$\mathcal{D}_2 := \Big\{ A_2 \in \mathcal{F} : \forall\ A_1 \in \sigma(\mathcal{A}_1), A_3 \in \mathcal{A}_3, \dots, A_n \in \mathcal{A}_n, \quad \mathbb{P}(A_1 \cap A_2 \cap \cdots \cap A_n) = \prod_{i=1}^n \mathbb{P}(A_i) \Big\},$$

and more generally,

$$\mathcal{D}_k :=$$
$$\Big\{ A_k : \forall\ A_1 \in \sigma(\mathcal{A}_1), \dots, A_{k-1} \in \sigma(\mathcal{A}_{k-1}), A_{k+1} \in \mathcal{A}_{k+1}, \dots, A_n \in \mathcal{A}_n,\ \mathbb{P}(\bigcap_{j=1}^n A_j) = \prod_{i=1}^n \mathbb{P}(A_i) \Big\}.$$

By induction, the claim then easily follows. □

Corollary 74. *Let $X_1, \dots, X_n$ be random variables such that for all $x_i \in \mathbb{R}$, $i = 1, \dots, n$,*

$$\mathbb{P}(X_1 \le x_1, \dots, X_n \le x_n) = \mathbb{P}(X_1 \le x_1) \cdots \mathbb{P}(X_n \le x_n). \tag{2.3}$$

Then $X_1, \dots, X_n$ are independent.

Proof. For $i = 1, \dots, n$, let $\mathcal{A}_i$ be the collection of events $\{X_i \le x\}$ for all $x \in \mathbb{R} \cup \{\infty\}$, and observe $\{X_i \le \infty\} = \Omega$. Observe also that (2.3) holds for this extension since $\mathbb{P}(X_i \le \infty) = 1$. Moreover, for $x, y \in \mathbb{R} \cup \{\infty\}$, we have $\{X_i \le x\} \cap \{X_i \le y\} = \{X_i \le \min(x, y)\} \in \mathcal{A}_i$, and therefore each $\mathcal{A}_i$ is a π-system. Now by Proposition 73 the sigma-algebras $\sigma(\mathcal{A}_1), \dots, \sigma(\mathcal{A}_n)$ are independent, so it is enough to show that $\sigma(\mathcal{A}_i) = \sigma(X_i)$ for all $i = 1, \dots, n$. But since $\{X_i \le x\} = X_i^{-1}(-\infty, x]$ for all $x \in \mathbb{R}$ and the intervals $(-\infty, x]$ generate the Borel σ-algebra, the claim follows by the definition of $\sigma(X_i)$. □

Remark 75. Equation (2.3) tells that independence of real random variables $X_1, \dots, X_n$ is in fact equivalent to the fact that the distribution function of the random vector $(X_1, \dots, X_n)$ factorizes into a product of the distribution functions of the marginals X_i. Observe that the converse implication of Corollary 74 holds by definition if the sigma-algebras $\sigma(X_i)$ are independent.

2.2 Characterizations for independence

As mentioned in the previous remark, Corollary 74 gives in some sense a characterization of independence via distribution functions. However, the distribution function only exists for $\mathbb{R}$-valued random variables and is sometimes also hard to compute, thus having its own limitations: Therefore, we would also like to explore other important characterizations of independence, which involve law or even density function. That leads us to the concept of *product measure*, and Dynkin's lemma will again play a role.

Definition 76 (Product σ-algebra)**.** Let $n \in \mathbb{N}$ and $(\mathcal{S}_1, \mathcal{A}_1), \dots, (\mathcal{S}_n, \mathcal{A}_n)$ be sets equipped with σ-algebras. The *product sigma-algebra* $\mathcal{P}$ on $\mathcal{S} := \mathcal{S}_1 \times \cdots \times \mathcal{S}_n$ is defined as the σ-algebra generated by the sets $A_1 \times \cdots \times A_n$ where $A_i \in \mathcal{A}_i$ for $i = 1, \dots, n$. That is, $\mathcal{P}$ is the smallest σ-algebra on $\mathcal{S}_1 \times \cdots \times \mathcal{S}_n$ containing the aforementioned sets $A_1 \times \cdots \times A_n$.

Example 77 (Product σ-algebra of Borel sets)**.** Let $n = 2$ and $(\mathcal{S}_i, \mathcal{A}_i) = (\mathbb{R}, \mathcal{B})$, $i = 1, 2$. Then the product-σ-algebra is generated by eg. the rectangles $(a, b) \times (a', b')$, $a, a', b, b' \in \mathbb{R}$, which generate the Borel sigma-algebra $\mathcal{B}(\mathbb{R}^2)$. The generalization to $\mathbb{R}^n$ is straightforward.

Definition 78 (Product measure)**.** Let $n \in \mathbb{N}$ and $\mu_1, \ldots, \mu_n$ be probability measures on $(\mathcal{S}_1, \mathcal{A}_1), \ldots, (\mathcal{S}_n, \mathcal{A}_n)$. Then the *product measure* $\bigotimes_{k=1}^n \mu_k = \mu_1 \otimes \cdots \otimes \mu_n$ is the unique measure on $(\mathcal{S}, \mathcal{P})$ such that $\bigotimes_{k=1}^n \mu_k(A_1 \times \cdots \times A_n) = \prod_{k=1}^n \mu_k(A_k)$ for all $A_1 \in \mathcal{A}_1, \ldots, A_n \in \mathcal{A}_n$.

Remark 79. To check that the product measure $\bigotimes_{k=1}^n \mu_k$ is well-defined, we need to show its existence and uniqueness. For the existence part, define

$$\mu(A) = \int_{\mathcal{S}_1} \left(\int_{\mathcal{S}_2} \cdots \left(\int_{\mathcal{S}_n} \mathbb{1}_A(x_1, \ldots, x_n) d\mu_n(x_n) \right) \ldots d\mu_2(x_2) \right) d\mu_1(x_1) \tag{2.4}$$

where $A \in \mathcal{P}$. It is not hard to see that μ is a measure on $(\mathcal{S}, \mathcal{P})$, and in fact is just a generalization of the setting in Example 9. Moreover, since $\mathbb{1}_{A_1 \times \cdots \times A_n} = \prod_{k=1}^n \mathbb{1}_{A_k}$, it follows that $\mu(A_1 \times \cdots \times A_n) = \prod_{k=1}^n \mu(A_k)$.

For the uniqueness part, let μ' be another measure on $(\mathcal{S}, \mathcal{P})$ satisfying the conditions of the definition, and define $\mathcal{D} := \{A \in \mathcal{P} : \mu(A) = \mu(A')\}$. It is not hard to check that the collection $\mathcal{A} := \{A_1 \times \cdots \times A_n : A_1 \in \mathcal{A}_1, \ldots, A_n \in \mathcal{A}_n\}$ is a π-system and $\mathcal{D}$ is a d-system, and obviously $\mathcal{A} \subset \mathcal{D}$. Hence by Dynkin's lemma, $\sigma(\mathcal{A}) \subset \mathcal{D}$ and thus μ and μ' agree on every measurable set, showing the uniqueness.

Observe that we could have exchanged the order of integrations in Equation 2.4, which by uniqueness would define the same product measure. This is an instance of Fubini's theorem, which is perhaps the most practically useful result related to product measures. We present this theorem without proof, and refer an interested reader eg. to [2].

Theorem 80 (Fubini's theorem)**.** *Let $n \in \mathbb{N}$ and $\mu_1, \ldots, \mu_n$ be probability measures on $(\mathcal{S}_1, \mathcal{A}_1), \ldots, (\mathcal{S}_n, \mathcal{A}_n)$ and μ their product measure on $\mathcal{S} = \mathcal{S}_1 \times \cdots \times \mathcal{S}_n$. Let f be a μ-measurable function. Then if $f \geq 0$ or $\int_{\mathcal{S}} |f| \, d\mu < \infty$,*

$$\begin{aligned} \int_{\mathcal{S}} f(x) d\mu(x) &= \int_{\mathcal{S}_1} \left(\int_{\mathcal{S}_2} \cdots \left(\int_{\mathcal{S}_n} f(x_1, \ldots, x_n) d\mu_n(x_n) \right) \ldots d\mu_2(x_2) \right) d\mu_1(x_1) \\ &= \int_{\mathcal{S}_n} \left(\int_{\mathcal{S}_2} \cdots \left(\int_{\mathcal{S}_1} f(x_1, \ldots, x_n) d\mu_1(x_1) \right) \ldots d\mu_2(x_2) \right) d\mu_n(x_n). \end{aligned}$$

Remark 81. The definition of the product measure and the statement of Fubini's theorem still hold if one replaces the probability measures by σ-finite measures. The version for $f \geq 0$ is often referred as Tonelli's theorem.

Now we are ready to give a characterization of independence via product measure.

Theorem 82 (Independence and product measure)**.** *Let $X_1, \ldots, X_n$ be real-valued random variables on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with laws $\mu_1, \ldots, \mu_n$, respectively. Then the following are equivalent.*

(i) $X_1, \ldots, X_n$ are independent.

(ii) The law μ of the random vector $X := (X_1, \dots, X_n)$ on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ is given by the product measure: $\mu = \bigotimes_{k=1}^n \mu_k$.

Proof. Let us start by proving $(ii) \Rightarrow (i)$. Thus, let X have the law $\mu = \bigotimes_{k=1}^n \mu_k$ and let $A_1, \dots, A_n \in \mathcal{B}(\mathbb{R})$ and $A := A_1 \times \cdots \times A_n$. Then, since μ is the product measure, we have $\mathbb{P}(X_1 \in A_1, \dots, X_n \in A_n) = \mathbb{P}(X \in A) = \mu(A) = \prod_{k=1}^n \mu_k(A_k) = \prod_{k=1}^n \mathbb{P}(X_k \in A_k)$. Since the sets A_k were arbitrary (and possibly $A_k = \Omega$), the sigma-algebras $\sigma(X_1), \dots, \sigma(X_n)$ are independent, which was to be shown.

Then, let us show $(i) \Rightarrow (ii)$. Observe that the sets of the form $A = A_1 \times \cdots \times A_n$, where $A_k \in \mathcal{B}(\mathbb{R})$, generate the Borel sigma-algebra $\mathcal{B}(\mathbb{R}^n)$ on $\mathbb{R}^n$. Now since $X_1, \dots, X_n$ are independent, we have by definition $\mu(A) = \mathbb{P}(X \in A) = \mathbb{P}(X_1 \in A_1, \dots, X_n \in A_n) = \prod_{k=1}^n \mathbb{P}(X_k \in A_k) = \prod_{k=1}^n \mu_k(A_k)$. Thus, by the definition and the uniqueness of the product measure, $\mu = \bigotimes_{k=1}^n \mu_k$. □

Example 83 (Uniform distribution and independence)**.** Consider $A = [x_1, x_2]$ and $B = [y_1, y_2]$, and let U be a uniform point in $A \times B$. That is, $\Omega = A \times B$, $\mathcal{F} = \mathcal{B}(\mathbb{R}^2)_{|A \times B}$ and $\mathbb{P} = \frac{1}{(x_2 - x_1)(y_2 - y_1)} \mu$, where μ is the Lebesgue measure on $A \times B$, and moreover U is the random variable $U(\omega) = \omega$ on $(\Omega, \mathcal{F}, \mathbb{P})$. For $\omega = (\omega_1, \omega_2) \in \Omega$, let $X_i(\omega) = \omega_i$ be the ith coordinate mapping $(i = 1, 2)$. We leave it as an exercise to show that X_1 and X_2 are independent. Conversely, let $D \subset \mathbb{R}^2$ be bounded and open and assume $U = (X_1, X_2)$ is uniform in D. Then X_1 and X_2 are independent if and only if D is an open rectangle. A similar results holds if D is closed, in which case independence corresponds to D being a closed rectangle.

Example 84 (Ratio of two independent exponential random variables)**.** We consider the setting in Example 57, this time with two independent exponential random variables X, Y with parameter $\lambda > 0$. In particular, both of the random variables have density $f(x) = \lambda e^{-\lambda x} \mathbb{1}_{x>0}$. Let us compute the law of Y/X. By Theorem 82, the joint law of (X, Y) is the product measure of the marginal laws of X and Y. Therefore, for $t > 0$, we find

$$\begin{aligned}
\mathbb{P}\left(\frac{Y}{X} > t\right) &= \mathbb{P}(Y > tX) = \int_{\mathbb{R}^2} \mathbb{1}_{y>tx} d(\mu_X \otimes \mu_Y)(x, y) = \int_{\mathbb{R}^2} \mathbb{1}_{y>tx} d\mu_X(x) d\mu_Y(y) \\
&= \int_0^\infty \left(\int_{tx}^\infty \lambda e^{-\lambda y} dy\right) \lambda e^{-\lambda x} dx = \int_0^\infty e^{-\lambda tx} \lambda e^{-\lambda x} dx = \int_0^\infty \lambda e^{-\lambda(t+1)x} dx \\
&= \frac{1}{1+t}.
\end{aligned}$$

The distribution function is then given by $F_{Y/X}(t) = \mathbb{P}(Y/X \le t) = \left(1 - \frac{1}{1+t}\right) \mathbb{1}_{t>0} = \frac{t}{1+t} \mathbb{1}_{t>0}$, and by differentiating we find a density $f_{Y/X}(t) = \frac{1}{(1+t)^2} \mathbb{1}_{t>0}$. Note that

$$\mathbb{E}(Y/X) = \int_{\mathbb{R}} t d\mu_{\frac{Y}{X}}(t) = \int_{\mathbb{R}} t f_{\frac{Y}{X}}(t) dt = \int_0^\infty \frac{t}{(1+t)^2} dt = \infty$$

where we applied Lemma 45 and Definition 37.

We continue now exploring the relationship of independence to expectation. To start with, we need one more abstract measure-theoretic result, which is essential to regroup random variables.

Lemma 85 (Regrouping lemma)**.** *Let $n \in \mathbb{N}$ and, for all $i = 1, \dots, n$, assign $m(i) \in \mathbb{N}$. Let $\{\mathcal{F}_{ij} : i \in \{1, \dots, n\}, j \in \{1, \dots, m(i)\}\}$ be a family of independent sub-σ-algebras on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. For all $i = 1, \dots, n$, let $\mathcal{A}_i := \sigma\left(\bigcup_{j=1}^{m(i)} \mathcal{F}_{ij}\right)$. Then $\mathcal{A}_1, \dots, \mathcal{A}_n$ are independent.*

Proof. For each $i = 1, \dots, n$, define $\mathcal{P}_i := \{A_{i1} \cap \dots \cap A_{im(i)} : A_{ij} \in \mathcal{F}_{ij} \quad \forall\, j = 1, \dots, m(i)\}$. The following properties follow immediately:

(i) $\mathcal{P}_i$ is a π-system.

(ii) $\Omega \in \mathcal{P}_i$.

(iii) $A_1 \in \mathcal{P}_1, \dots, A_n \in \mathcal{P}_n \Rightarrow \mathbb{P}\left(\bigcap_{i=1}^n A_i\right) = \prod_{i=1}^n \mathbb{P}(A_i)$.

It follows from Proposition 73 that the sigma-algebras $\sigma(\mathcal{P}_1), \dots, \sigma(\mathcal{P}_n)$ are independent, and by definition, $\sigma(\mathcal{P}_i) = \mathcal{A}_i$. □

Corollary 86. *Let $\{X_{ij} : i \in \{1, \dots, n\}, j \in \{1, \dots, m(i)\}\}$ be a family of independent random variables on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. For each $i = 1, \dots, n$, let $f_i : \mathbb{R}^{m(i)} \to \mathbb{R}$ be a measurable function. Then the random variables $f_i(X_{i1}, \dots, X_{im(i)})$, $i = 1, \dots, n$, are independent.*

Proof. We denote $\mathcal{F}_{ij} := \sigma(X_{ij})$ and $\mathcal{A}_i := \sigma\left(\bigcup_{j=1}^{m(i)} \mathcal{F}_{ij}\right)$. Then by the previous lemma, the sigma-algebras $\mathcal{A}_i$ are independent. Moreover, obviously $\sigma(f_i(X_{i1}, \dots, X_{im(i)})) \subset \mathcal{A}_i$ for all $i = 1, \dots, n$. Hence the claim follows. □

Example 87. The previous corollary is useful in order to determine independence of random variables under measurable transformation. For example, if $X_1, \dots, X_n$ are independent random variables, so are X_1 and the product $X_2 \cdots X_n$, or the sum $X_1 + \dots + X_{n-1}$ and X_n^2.

We apply the previous results to expectations of random variables. One remarkable property is that, under mild conditions, the expectation of a product of independent random variables factorizes into the product of expectations of the random variables.

Theorem 88. *Let $X_1, \dots, X_n$ be independent random variables such that at least one of the following holds:*

(i) $X_i \geq 0$ for all $i = 1, \dots, n$.

(ii) $\mathbb{E}(|X_i|) < \infty$ for all $i = 1, \dots, n$.

Then $\mathbb{E}(X_1 \cdots X_n) = \mathbb{E}(X_1) \cdots \mathbb{E}(X_n)$.

Proof. The proof is by induction on n. First, let $n = 2$, and X_1, X_2 be independent random variables with laws μ_1 and μ_2, respectively. By independence, the law of (X_1, X_2) is given by the product measure $\mu_1 \otimes \mu_2$. Thus,

$$\begin{aligned}\mathbb{E}(X_1 X_2) &= \int_{\mathbb{R}^2} xy d(\mu_1 \otimes \mu_2)(x, y) = \int_{\mathbb{R}} \int_{\mathbb{R}} xy d\mu_1(x) d\mu_2(y) \\ &= \int_{\mathbb{R}} x d\mu_1(x) \cdot \int_{\mathbb{R}} y d\mu_2(y) = \mathbb{E}(X_1)\mathbb{E}(X_2).\end{aligned}$$

Observe that we applied Fubini's theorem in the last row, which is justified by assumptions (i) or (ii).

Let us now assume $k \geq 3$ is such that the claim holds for $k-1$ independent random variables satisfying (i) or (ii). Then we have $\mathbb{E}(X_2 \cdots X_k) = \mathbb{E}(X_2) \cdots \mathbb{E}(X_k)$ by the induction assumption. Moreover, if $X_i \geq 0$ for all i, then $X_2 \cdots X_k \geq 0$. By Corollary 86, the random variables X_1 and $X_2 \cdots X_k$ are independent. Hence by the $n = 2$ case, we now find $\mathbb{E}(X_1 X_2 \cdots X_k) = \mathbb{E}(X_1)\mathbb{E}(X_2 \cdots X_k) = \mathbb{E}(X_1)\mathbb{E}(X_2) \cdots \mathbb{E}(X_k)$. The final claim follows then by induction. If

$\mathbb{E}(|X_i|) < \infty$ for all $i = 1$, then $\mathbb{E}(|X_2 \cdots X_k|) = \mathbb{E}(|X_2|) \cdots \mathbb{E}(|X_k|) < \infty$, since $|X_1|, \ldots, |X_n|$ are independent by Corollary 86. The claim follows as before. □

Remark 89. An alternative proof could be obtained by showing the claim $n = 2$ using the traditional measure-theoretic construction of integral, which is also present in the proof of Proposition 45. That is, the claim is first shown to simple functions, then passed to non-negative functions using the monotone convergence theorem, and finally to general integrable functions by the decomposition to positive and negative parts.

The above theorem has very important implications to variance and covariance:

Corollary 90. *Let $X_1, \ldots, X_n$ be independent random variables with $\mathbb{E}(X_i^2) < \infty$ for all $i = 1, \ldots, n$. Then $\operatorname{Var}(X_1 + \cdots + X_n) = \operatorname{Var}(X_1) + \cdots + \operatorname{Var}(X_n)$, and $\operatorname{Cov}(X_i, X_j) = 0$ for all $i \neq j$.*

Proof. First by independence, $\operatorname{Cov}(X_i, X_j) = \mathbb{E}(X_i X_j) - \mathbb{E}(X_i)\mathbb{E}(X_j) = \mathbb{E}(X_i)\mathbb{E}(X_j) - \mathbb{E}(X_i)\mathbb{E}(X_j) = 0$. Then, denote $m_i := \mathbb{E}(X_i)$. Using the definition of variance, we then obtain

$$\begin{aligned}
\operatorname{Var}\left(\sum_{i=1}^n X_i\right) &= \mathbb{E}\left(\left(\sum_{i=1}^n X_i - \mathbb{E}\left(\sum_{i=1}^n X_i\right)\right)^2\right) = \mathbb{E}\left(\left(\sum_{i=1}^n (X_i - m_i)\right)^2\right) \\
&= \mathbb{E}\left(\sum_{i,j=1}^n (X_i - m_i)(X_j - m_j)\right) \\
&= \sum_{i=1}^n \mathbb{E}\left((X_i - m_i)^2\right) + \sum_{i \neq j} \mathbb{E}\left((X_i - m_i)(X_j - m_j)\right) \\
&= \sum_{i=1}^n \operatorname{Var}(X_i) + \sum_{i \neq j} \operatorname{Cov}(X_i, X_j) = \sum_{i=1}^n \operatorname{Var}(X_i).
\end{aligned}$$

□

Example 91 (Sum of independent random variables and convolution)**.** Let X and Y be independent real random variables. It is left as an exercise to the reader to find the law of the sum $X + Y$. Assuming X has law μ and Y's law is described with distribution function G, one can find quite a neat expression for the distribution function of $X + Y$. Moreover, one can show that if either of X or Y has a density, so does the sum $X + Y$. If both of them have densities, say f and g respectively, then a density of the sum is given by the *convolution* $f \star g(x) = \int_{\mathbb{R}} f(t)g(x-t)dt$. The reader is encouraged to prove this, as well as to apply it to the sum of two independent exponential or uniform random variables, respectively.

The law of a sum of two discrete random variables is also given by a convolution. For instance, recall from Example 23 that the probability mass function $\mathbb{P}(n) = e^{-\lambda}\frac{\lambda^n}{n!}$ for $n \in \{0, 1, 2, \dots\}$ defines the Poisson distribution with parameter $\lambda > 0$. Now, let $X \sim \text{Poisson}(\lambda)$ and $Y \sim \text{Poisson}(\mu)$ be independent Poisson-distributed random variables. Then

$$\begin{aligned}\mathbb{P}(X+Y=n) &= \sum_{k=0}^{n} \mathbb{P}(X=k)\mathbb{P}(Y=n-k) = \sum_{k=0}^{n} e^{-\lambda}\frac{\lambda^k}{k!}e^{-\mu}\frac{\mu^{n-k}}{(n-k)!} \\ &= \frac{e^{-(\lambda+\mu)}}{n!}\sum_{k=0}^{n}\binom{n}{k}\lambda^k\mu^{n-k} = e^{-(\lambda+\mu)}\frac{(\lambda+\mu)^n}{n!}.\end{aligned}$$

Therefore, $X + Y \in \text{Poisson}(\lambda + \mu)$.

The reader might wonder if a converse statement for Theorem 88 holds. This is not true as such, but instead needs a more general statement, which takes into account a large class of transformations of the random variables.

Theorem 92. *Let $X_1, \dots, X_n$ be real random variables. Then the following are equivalent.*

(i) $X_1, \dots, X_n$ are independent.

(ii) For any bounded Borel-measurable functions $g_i : \mathbb{R} \to \mathbb{R}$, $(i = 1, \dots, n)$,

$$\mathbb{E}(g_1(X_1)\cdots g_n(X_n)) = \mathbb{E}(g_1(X_1))\cdots\mathbb{E}(g_n(X_n)). \tag{2.5}$$

(iii) For any bounded continuous functions $g_i : \mathbb{R} \to \mathbb{R}$, $(i = 1, \dots, n)$, 2.5 holds.

Proof. Let us first show $(i) \Rightarrow (ii)$. That is, assume $X_1, \dots, X_n$ are independent and let $g_i : \mathbb{R} \to \mathbb{R}$, $(i = 1, \dots, n)$ be bounded Borel functions. Then by Corollary 86, the random variables $g_1(X_1), \dots, g_n(X_n)$ are independent. Moreover, since the functions g_i are bounded, we have $\mathbb{E}(|f_i(X_i)|) < \infty$ for all $i = 1, \dots, n$. Hence by Theorem 88, the claim (ii) follows. Clearly $(ii) \Rightarrow (iii)$ holds, since continuous functions are measurable.

Although redundant in principle, let us then show $(ii) \Rightarrow (i)$, which gives some insight and a computation for $(iii) \Rightarrow (i)$ which is useful in practice. For this, let $B_1 \dots, B_n \in \mathcal{B}$ and define $g_i := \mathbb{1}_{B_i}$. Obviously the indicator functions are bounded and measurable. Hence by (2.5), we have

$$\begin{aligned}\mathbb{P}(X_1 \in B_1, \dots, X_n \in B_n) &= \mathbb{E}(\mathbb{1}_{X_1 \in B_1} \dots \mathbb{1}_{X_n \in B_n}) = \mathbb{E}(g_1(X_1) \dots g_n(X_n)) \\ &= \prod_{i=1}^{n} \mathbb{E}(g_i(X_i)) = \prod_{i=1}^{n} \mathbb{E}(\mathbb{1}_{X_i \in B_i}) = \prod_{i=1}^{n} \mathbb{P}(X_i \in B_i).\end{aligned}$$

The implication $(iii) \Rightarrow (i)$ is somewhat more cumbersome. For simplicity of notation, let us prove it for $n = 2$, and note that the general case is a straightforward generalization. Thus, let X and Y be random variables such that for all continuous and bounded functions f and g, we have $\mathbb{E}(f(X)g(Y)) = \mathbb{E}(f(X))\mathbb{E}(g(Y))$. Let ν and μ denote the laws of X and Y, respectively. Let $A, B \in \mathcal{B}$, and choose $f = \mathbb{1}_A$ and $g = \mathbb{1}_B$. Note that f and g are in general not continuous, yet they are step functions and in particular bounded. Observe also that any powers of bounded functions are integrable with respect to the laws ν and μ (since they are probability measures), and hence $f, g \in L^p$ for all $p \in [1, \infty]$

(where L^∞ is the space of essentially bounded functions). That is, $\int |f|^p \, d\nu < \infty$, and same holds for the pair (g, μ). By a general theorem on function spaces, continuous functions are dense in L^p when $1 \leq p < \infty$. Hence, the functions f and g can be approximated by sequences of continuous L^p functions $(f_n)_{n=0}^\infty$ and $(g_n)_{n=0}^\infty$ such that $f_n \xrightarrow[n\to\infty]{} f$ and $g_n \xrightarrow[n\to\infty]{} g$ in the L^p-norm. For the rest of the proof, let us choose $p = 2$, so in particular, we could apply the Cauchy-Schwarz inequality. Without loss of generality, since f, g have values in $\{0, 1\}$, the functions f_n and g_n can be assumed to be bounded. Namely, if this is not the case, we could just set their values to constants 0 and 1 in those sets where the functions attain values negative or greater that 1, respectively. Now by the assumption, we have $\mathbb{E}(f_n(X)g_n(Y)) = \mathbb{E}(f_n(X))\mathbb{E}(g_n(Y))$ for all $n \in \mathbb{N}$. It would be tempting to try dominated convergence theorem at this point, but observe that we only have convergence of the functions f_n and g_n in the L^2-norm, which is weaker than almost sure convergence. Let us anyway show $\mathbb{E}(f_n(X)) \xrightarrow[n\to\infty]{} \mathbb{E}(f(X))$ (which then holds when f is replaced by g and X by Y as well), and $\mathbb{E}(f_n(X)g_n(Y)) \xrightarrow[n\to\infty]{} \mathbb{E}(f(X)g(Y))$. Instead of the DCT, the idea is to apply the Cauchy-Schwarz inequality. First,

$$|\mathbb{E}(f_n(X)) - \mathbb{E}(f(X))| = |\mathbb{E}(f_n(X) - f(X))| \leq \mathbb{E}(|f_n(X) - f(X)|) \leq \sqrt{\mathbb{E}\left((f_n(X) - f(X))^2\right)}.$$

Recall (Proposition 45) that $\mathbb{E}\left((f_n(X) - f(X))^2\right) = \int_{\mathbb{R}} (f_n(x) - f(x))^2 d\nu(x)$, which converges to 0 since $f_n \xrightarrow[n\to\infty]{} f$ in L^2. The claim for g_n and g is similar. Finally, observe that by writing $f_n(X)g_n(Y) - f(X)g(Y) = f_n(X)g_n(Y) - f(X)g_n(Y) + f(X)g_n(Y) - f(X)g(Y)$, and applying the triangle inequality as well as Cauchy-Schwarz, we find

$$\begin{aligned}
|\mathbb{E}(f_n(X)g_n(Y) - f(X)g(Y))| &= |\mathbb{E}(g_n(Y)(f_n(X) - f(X)) + f(X)(g_n(Y) - g(Y))| \\
&\leq \mathbb{E}(|g_n(Y)(f_n(X) - f(X)|) + \mathbb{E}(|f(X)(g_n(Y) - g(Y))|) \\
&\leq \sqrt{\mathbb{E}\left(g_n(Y)^2\right)\mathbb{E}\left((f_n(X) - f(X))^2\right)} \\
&+ \sqrt{\mathbb{E}(f(X)^2)\mathbb{E}\left((g_n(Y) - g(Y))^2\right)}.
\end{aligned}$$

Since $f_n \xrightarrow[n\to\infty]{} f$ and $g_n \xrightarrow[n\to\infty]{} g$ in L^2, it follows that $\mathbb{E}(f_n(X)g_n(Y)) \xrightarrow[n\to\infty]{} \mathbb{E}(f(X)g(Y))$. By uniqueness of limit, we thus have $\mathbb{E}(f(X)g(Y)) = \mathbb{E}(f(X))\mathbb{E}(g(Y))$, and the final claim (i) follows from the computation for $(ii) \Rightarrow (i)$. □

Remark 93. In the above theorem, it is easy to see that we could also choose non-negative Borel-measurable functions instead of bounded ones. It also works if the boundedness is relaxed to functions which keep the random variables integrable.

Chapter 3

Sequences of events and random variables

So far, we have mostly considered finite collections of events or random variables. Now we relax the finiteness assumption, which yields some interesting results which could, broadly speaking, be classified under the umbrella term *zero-one laws.* The latter term means that under some natural (and often mild) conditions, certain events always have either probability zero or one. That is very powerful, because then computing the probability of such events reduce to showing if some pre-condition holds or not, or whether the estimated probability is positive or not. As the reader might already anticipate, independence assumptions are crucial for many of such results.

3.1 Borel-Cantelli lemmas

In this section, we discuss two very important results, which are both simple to state and prove, but which turn out to be extremely useful to determine whether a given event occurs infinitely often or not.

Definition 94 (Limit superior and limit inferior of a sequence of sets)**.** Let $A_1, A_2, \dots$ be an infinite sequence of events (with possibly only a finite number of $A_k \neq \emptyset$). Then the *limit superior* and *limit inferior* of the sequence $A_1, A_2, \dots$ are the events

$$\limsup_n A_n := \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k$$

and

$$\liminf_n A_n := \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} A_k.$$

Remark 95. Since all complements and countable unions of measurable sets are measurable (i.e. belong to the given σ-algebra), the sets $\limsup_n A_n$ and $\liminf_n A_n$ are indeed measurable, i.e. events. Note that $\omega \in \limsup_n A_n$ if and only if for all $n \geq 1$, there exists such a $k \geq n$ that $\omega \in A_k$. To put in words, the outcome ω is observed infinitely often, i.e. infinitely many A_n occur. Similarly, $\omega \in \liminf_n A_n$ if and only if there exists such an $n \geq 1$, that for all $k \geq n$, $\omega \in A_k$. That is, all A_n occur for n large enough.

Remark 96. Observe also the relation: $(\liminf_n A_n)^c = \limsup_n A_n^c$. This event means that infinitely many sets A_n do not occur.

Remark 97. Recall that if $(x_n)_{n=0}^\infty$ is a sequence of real numbers, then $\limsup_{n\to\infty} x_n = \inf_{n\geq 0}\left(\sup_{m\geq n} x_m\right)$ and $\liminf_{n\to\infty} x_n = \sup_{n\geq 0}\left(\inf_{m\geq n} x_m\right)$. The connection to sets is the following, which is left as an exercise to the reader: $\mathbb{1}_{\limsup_n A_n} = \limsup_{n\to\infty} \mathbb{1}_{A_n}$ and $\mathbb{1}_{\liminf_n A_n} = \liminf_{n\to\infty} \mathbb{1}_{A_n}$.

We fix the following convention: We write $\limsup_n A_n$ and $\{A_n \text{ occurs } i.o.\}$ interchangeably, and so do we with $\mathbb{P}(\limsup_n A_n) = 1$ and "A_n i.o.", where "i.o." abbreviates "infinitely often".

Example 98 (Outcomes in an infinite series of coin tosses)**.** Consider an infinite series $(X_1, X_2, \dots)$ of coin tosses, which are assumed to be independent of each other. Then $X_k \in \{0,1\}$, so that $\mathbb{P}(X_k = i) = \frac{1}{2}$ for $i = 0,1$. Let $A_n := \{X_n = 1\}$, i.e. the event that the nth coin flip results tails. Let us show that A_n i.o.

Indeed, for any $k \geq 1$, we have $\mathbb{P}(A_n^c \cap \dots \cap A_{n+k-1}^c) = \mathbb{P}(X_n = 0, \dots, X_{n+k-1} = 0) = \frac{1}{2^k}$, and therefore $\mathbb{P}\left(\bigcap_{i=n}^\infty A_i^c\right) = \lim_{k\to\infty} \frac{1}{2^k} = 0$. Hence,

$$\mathbb{P}\left(\bigcup_{n=1}^\infty \bigcap_{i=n}^\infty A_i^c\right) \leq \sum_{n=1}^\infty \mathbb{P}\left(\bigcap_{i=n}^\infty A_i^c\right) = 0,$$

yielding $\mathbb{P}(\limsup_n A_n) = \mathbb{P}\left(\bigcap_{n=1}^\infty \bigcup_{i=n}^\infty A_i\right) = 1$ as desired.

Theorem 99 (First Borel-Cantelli lemma)**.** *Suppose $A_1, A_2, \dots$ is a sequence of events such that $\sum_{n=1}^\infty \mathbb{P}(A_n) < \infty$. Then $\mathbb{P}(\limsup_n A_n) = 0$, i.e. only finitely many A_n occur a.s.*

Proof. For all $m = 1, 2, \dots$, we have

$$\mathbb{P}(\limsup_n A_n) = \mathbb{P}\left(\bigcap_{n=1}^\infty \bigcup_{k=n}^\infty A_k\right) \leq \mathbb{P}\left(\bigcup_{k=m}^\infty A_k\right) \leq \sum_{k=m}^\infty \mathbb{P}(A_k).$$

Since the series $\sum_{n=1}^\infty \mathbb{P}(A_n)$ converges, we have $\sum_{k=m}^\infty \mathbb{P}(A_k) \xrightarrow[m\to\infty]{} 0$. The claim follows. □

If the above series is not convergent, we need an independence assumption to deduce a "converse" result:

Theorem 100 (Second Borel-Cantelli lemma)**.** *Suppose $A_1, A_2, \dots$ is a sequence of independent events such that $\sum_{n=1}^\infty \mathbb{P}(A_n) = \infty$. Then A_n i.o. That is, $\mathbb{P}(\limsup_n A_n) = 1$.*

Proof. The proof relies on the inequality $1 + x \leq e^x$ for all $x \in \mathbb{R}$. There are many ways to prove it, one is that the line $1 + x$ is tangent to e^x at $x = 0$, and $x \mapsto e^x$ is convex. To see how this can be applied, observe that $(\limsup_n A_n)^c = \bigcup_{n=1}^\infty \bigcap_{k=n}^\infty A_k^c = \liminf_n A_n^c$, and we can thus equivalently show $\mathbb{P}(\liminf_n A_n^c) = 0$. Now fix $N > n \geq 1$. Then by independence of the sets $A_1, A_n \dots$ (and hence their complements),

$$\mathbb{P}\left(\bigcap_{k=n}^N A_k^c\right) = \prod_{k=n}^N \mathbb{P}(A_k^c) = \prod_{k=n}^N (1 - \mathbb{P}(A_k)) \leq \prod_{k=n}^N \exp\left(-\mathbb{P}(A_k)\right) = \exp\left(-\sum_{k=n}^N \mathbb{P}(A_k)\right)$$

where the right hand side tends to 0 as $N \to \infty$ by the assumption $\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty$. Hence, $\lim_{N\to\infty} \mathbb{P}\left(\bigcap_{k=n}^{N} A_k^c\right) = 0$. Note that the sets $\bigcap_{k=n}^{N} A_k^c$ are decreasing in N, and thus by Lemma 16, we deduce $\lim_{N\to\infty} \mathbb{P}\left(\bigcap_{k=n}^{N} A_k^c\right) = \mathbb{P}\left(\bigcap_{k=n}^{\infty} A_k^c\right)$. Hence, we have $\mathbb{P}(\liminf_n A_n^c) \leq \sum_{n=1}^{\infty} \mathbb{P}\left(\bigcap_{k=n}^{\infty} A_k^c\right) = 0$ as desired. □

Remark 101. If the events $A_1, A_2 \dots$ are independent, the two Borel-Cantelli lemmas combined form a $0-1$-law: then always $\mathbb{P}(\limsup_n A_n) \in \{0,1\}$. In fact, this is a special case of Kolmogorov's $0-1$-law, which will be presented in the following section.

Remark 102. The assumption that $A_1, A_2 \dots$ are independent is necessary in the 2nd Borel-Cantelli lemma. For example, consider $\Omega = [0,1]$ with the Borel σ-algebra and the uniform distribution $\mathbb{P}$. Let $A_n := \left[0, \frac{1}{n}\right]$ $(n = 1, 2, \dots)$. Then $\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \sum_{n=1}^{\infty} \frac{1}{n} = \infty$, but $\mathbb{P}(\limsup_n A_n) = \mathbb{P}(\{0\}) = 0$.

Example 103. Let $X_1, X_2, \dots$ be a sequence of independent exponentially distributed random variables with parameter $\lambda > 0$. Recall from Example 57 that the law of X_n is then given by $\mathbb{P}(X_n > t) = e^{-\lambda t}$ for all $t \geq 0$. We show first that $\limsup_{n\to\infty} X_n = \infty$ almost surely. To see this, fix $M > 0$ and define $A_n := \{X_n > M\}$. Then the sets A_n are measurable and independent by definition, and $\mathbb{P}(A_n) = e^{-\lambda M}$ for all $n = 1, 2, \dots$. In particular, $\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty$. Hence by the 2nd Borel-Cantelli lemma, $\mathbb{P}(\limsup_n A_n) = 1$, that is, $X_n > M$ i.o. which implies $\limsup_{n\to\infty} X_n \geq M$ a.s. The latter simply means $\mathbb{P}(\limsup_{n\to\infty} X_n \geq M) = 1$. Since the sets $\{X_n \geq M\}$ are decreasing in M, we finally deduce $\mathbb{P}(\limsup_{n\to\infty} X_n = \infty) = \mathbb{P}\left(\bigcap_{M=1}^{\infty} \{\limsup_{n\to\infty} X_n \geq M\}\right) = \lim_{M\to\infty} \mathbb{P}(\limsup_{n\to\infty} X_n \geq M) = 1$.

In fact, we can say more about the convergence. More precisely, $\limsup_{n\to\infty} \frac{X_n}{\lambda^{-1}\log(n)} = 1$ a.s. The claim follows if we can show the following two inequalities for all $\epsilon > 0$:

(i) $X_n \leq (1+\epsilon)\frac{\log(n)}{\lambda}$ for n large enough (a.s.),

(ii) $X_n \geq (1-\epsilon)\frac{\log(n)}{\lambda}$ i.o.

Then (i) and (ii) combined would yield $\mathbb{P}\left(1 - \frac{1}{k} \leq \limsup_{n\to\infty} \frac{X_n}{\lambda^{-1}\log(n)} \leq 1 + \frac{1}{k}\right) = 1$ for all $k \geq 1$, and thus

$$\mathbb{P}\left(\limsup_{n\to\infty} \frac{X_n}{\lambda^{-1}\log(n)} = 1\right) = \mathbb{P}\left(\bigcap_{k=1}^{\infty} \left\{1 - \frac{1}{k} \leq \limsup_{n\to\infty} \frac{X_n}{\lambda^{-1}\log(n)} \leq 1 + \frac{1}{k}\right\}\right) = 1.$$

Therefore, it is enough to show (i) and (ii). Here, the Borel-Cantelli lemmas will be applied. First, $\mathbb{P}\left(X_n > (1+\epsilon)\frac{\log(n)}{\lambda}\right) = \exp\left(-\lambda(1+\epsilon)\frac{\log(n)}{\lambda}\right) = n^{-(1+\epsilon)}$ which is summable since $\epsilon > 0$. Therefore by the 1st Borel-Cantelli, $X_n > (1+\epsilon)\frac{\log(n)}{\lambda}$ only a finite number of times, which means that eventually $X_n \leq (1+\epsilon)\frac{\log(n)}{\lambda}$, showing the claim (i). Then, $\mathbb{P}\left(X_n \geq (1-\epsilon)\frac{\log(n)}{\lambda}\right) = \exp\left(-\lambda(1-\epsilon)\frac{\log(n)}{\lambda}\right) = n^{-(1-\epsilon)}$ for $0 < \epsilon < 1$, which is not summable. Since the random variables $X_1, X_2, \dots$ are independent, so are the events $\left\{X_n \geq (1-\epsilon)\frac{\log(n)}{\lambda}\right\}$. Hence by the 2nd Borel-Cantelli, $X_n \geq (1-\epsilon)\frac{\log(n)}{\lambda}$ i.o., concluding (ii).

3.2 Kolmogorov's zero-one law

In this section, we generalize the *zero-one law* for independent events described above. It leads us to a remarkable result by Kolmogorov, which roughly tells that a sequence of independent random variables cannot have a nontrivial behavior at ∞. To understand the behavior of random variables for large indices, we first need to understand σ-algebras which contain information up to ∞, but in which information corresponding to a finite number of indices does not matter.

Definition 104. Let $\mathcal{F}_1, \mathcal{F}_2, \dots$ be independent σ-algebras. For all $n = 1, 2, \dots$, let $\mathcal{T}_n := \sigma(\mathcal{F}_n, \mathcal{F}_{n+1}, \dots)$. Then the *tail σ-algebra* is defined by $\mathcal{T} := \bigcap_{n=1}^{\infty} \mathcal{T}_n$.

Remark 105. The tail σ-algebra describes the events whose occurrences are not affected by knowledge of just a finite number of σ-algebras $\mathcal{F}_n$.

Example 106. Let $X_1, X_2 \dots$ be independent random variables, and define $\mathcal{F}_n := \sigma(X_n)$. Then for example $E_1 := \{\omega : \lim_{n\to\infty} X_n(\omega) \ \text{ exists}\}$ and $E_2 := \{\limsup_{n\to\infty} X_n = \infty\}$ belong to $\mathcal{T}$, since their occurrence clearly does not depend on a finite number of the indices n. Formally, let us show the claim for E_1. In that case, we note that $\lim_{n\to\infty} X_n$ exists if and only if $(X_n)_{n=0}^{\infty}$ is a Cauchy sequence; equivalently, for all $\epsilon > 0$ there exists an $N > 0$ such that for all integers $m, n > N$, $|X_n - X_m| < \epsilon$. Therefore, we find $E_1 = \bigcap_{k=1}^{\infty} \bigcup_{N=1}^{\infty} \bigcap_{m,n>N} \left\{ |X_n - X_m| < \frac{1}{k} \right\}$. Note that $\mathcal{T}_N \supset \mathcal{T}_{N+1} \supset \cdots \supset \bigcap_{i=1}^{\infty} \mathcal{T}_i = \mathcal{T}$, hence $\left\{ |X_n - X_m| < \frac{1}{k} \right\} \in \mathcal{T}_N$ if $n, m > N$, and so does its complement. Thus,

$$\bigcap_{N=1}^{\infty} \bigcup_{m,n>N} \left\{ |X_n - X_m| < \frac{1}{k} \right\}^c \in \mathcal{T},$$

so finally $E_1^c = \bigcup_{k=1}^{\infty} \bigcap_{N=1}^{\infty} \bigcup_{m,n>N} \left\{ |X_n - X_m| < \frac{1}{k} \right\}^c \in \mathcal{T}$, yielding $E_1 \in \mathcal{T}$.

Theorem 107 (Kolmogorov's $0-1$-law)**.** *Let $\mathcal{F}_1, \mathcal{F}_2, \dots$ be independent σ-algebras and $\mathcal{T}$ their tail σ-algebra. Then for any $A \in \mathcal{T}$, we have $\mathbb{P}(A) \in \{0, 1\}$. That is, the σ-algebra $\mathcal{T}$ consists of sets that either a.s. occur or a.s. do not occur.*

Proof. First, note that for all $n = 1, 2, \dots$, the σ-algebras $\sigma(\mathcal{F}_1, \dots, \mathcal{F}_n)$ and $\mathcal{T}_{n+1}$ are independent, which formally follows from a slight generalization of the regrouping lemma 85. Then, since $\mathcal{T} = \bigcap_{k\geq 1} \mathcal{T}_k \subset \mathcal{T}_{n+1}$, the σ-algebras $\sigma(\mathcal{F}_1, \dots, \mathcal{F}_n)$ and $\mathcal{T}$ are independent. Consider now $\mathcal{P} := \bigcup_{n=1}^{\infty} \sigma(\mathcal{F}_1, \dots, \mathcal{F}_n)$. If $A \in \mathcal{P}$ and $B \in \mathcal{P}$, then $A \in \sigma(\mathcal{F}_1, \dots, \mathcal{F}_k)$ and $B \in \sigma(\mathcal{F}_1, \dots, \mathcal{F}_m)$ for some $k, m \geq 1$, and then $A \cap B \in \sigma(\mathcal{F}_1, \dots, \mathcal{F}_{\min(k,m)})$. Therefore, $\mathcal{P}$ is a π-system, and so is also $\mathcal{T}$ since it is even a σ-algebra. By the previous independence considerations, if now $A \in \mathcal{P}$ and $B \in \mathcal{T}$, then $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. Now it follows from Proposition 73 (which followed from Dynkin's lemma) that $\sigma(\mathcal{P})$ and $\sigma(\mathcal{T})$ are independent, that is, $\sigma(\mathcal{F}_1, \mathcal{F}_2, \dots)$ and $\mathcal{T}$ are independent. But $\mathcal{T} \subset \sigma(\mathcal{F}_1, \mathcal{F}_2, \dots)$, so it is in fact independent of itself! Hence, for all $A \in \mathcal{T}$, we have $\mathbb{P}(A) = \mathbb{P}(A \cap A) = \mathbb{P}(A)^2$, which only holds if $\mathbb{P}(A) = 0$ or $\mathbb{P}(B) = 1$. □

Example 108. Let $\mathcal{F}_1, \mathcal{F}_2, \dots$ be independent σ-algebras and $A_i \in \mathcal{F}_i$ for all $i = 1, 2, \dots$. Let $A := \limsup_n A_n = \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k$. Then it is easy to see that $A \in \mathcal{T}$, so by Kolmogorov's $0-1$-law, $\mathbb{P}(A) \in \{0, 1\}$. This is the zero-one law that follows from the two Borel-Cantelli-lemmas.

3.3 Infinite product spaces and sequences of independent random variables

In this section, we discuss briefly the existence of an infinite sequence of independent random variables and infinite product measures. Although the material is classical, we present most of the results without proofs. These results will be crucial in the following chapter, where we apply the theory introduced so far to the percolation model.

Recall from Definitions 76 and 78 the finite product σ-algebras and product measures. The generalization of these concepts to countable infinite setting is based on *cylinder sets*.

Definition 109 (Cylinder σ-algebra)**.** Let

$$\Omega := \{\omega = (\omega_1, \omega_2, \dots) \; : \; \omega_i \in \mathbb{R} \quad \forall\, i = 1, 2, \dots\}.$$

Let $n \in \mathbb{N}$ and $B_1, \dots, B_n \in \mathcal{B}(\mathbb{R})$. A set of the form

$$\{\omega \in \Omega \; : \; \omega_i \in B_i \quad \forall\, i = 1, 2, \dots, n\} = B_1 \times \cdots \times B_n \times \prod_{i=n+1}^{\infty} \mathbb{R}$$

is called a *cylinder set*. The σ-algebra $\mathcal{F}$ on Ω generated by the cylinder sets is called the *cylinder σ-algebra*.

Let $\mu_1, \mu_2, \dots$ be probability measures on $\mathbb{R}$. We construct now the infinite product measure $\mu := \otimes_{i=1}^{\infty} \mu_i$. For $B_1, \dots, B_n \in \mathcal{B}$, we denote by $\mathrm{Cyl}(B_1, \dots, B_n)$ the cylinder set $B_1 \times \cdots \times B_n \times \prod_{i=n+1}^{\infty} \mathbb{R}$. Then, we set $\mu\left(\mathrm{Cyl}(B_1, \dots, B_n)\right) = \prod_{i=1}^{n} \mu_i(B_i)$. The extension of μ to the cylinder σ-algebra $\mathcal{F}$ follows from the following important theorem, which we present without proof. For a proof, see eg. [2].

Theorem 110 (Kolmogorov's extension theorem)**.** *Let $(\Omega, \mathcal{F})$ be as in the previous definition. For every $n = 1, 2, \dots$, let ν_n be a probability measure on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$. Furthermore, assume that the measures ν_n are consistent in the sense that $\nu_{n+1}(B_1 \times \cdots \times B_n \times \mathbb{R}) = \nu_n(B_1 \times \cdots \times B_n)$ for all $n = 1, 2, \dots$ and $B_1, \dots, B_n \in \mathcal{B}(\mathbb{R})$. Then there exists a unique probability measure ν on $(\Omega, \mathcal{F})$ such that $\nu\left(\mathrm{Cyl}(B_1, \dots, B_n)\right) = \nu_n(B_1 \times \cdots \times B_n)$ for all $B_1, \dots, B_n \in \mathcal{B}(\mathbb{R})$, $n \in \mathbb{N}$.*

We proceed now with the construction of the infinite product measure. With the notation of Kolmogorov's extension theorem, we set $\nu_n = \mu_1 \otimes \cdots \otimes \mu_n$ for each $n = 1, 2, \dots$. Now for all $B_1, \dots, B_n \in \mathcal{B}(\mathbb{R})$, we have $\nu_{n+1}(B_1 \times \cdots \times B_n \times \mathbb{R}) = \mu_1(B_1) \cdots \mu_n(B_n) \mu_{n+1}(\mathbb{R}) = \mu_1(B_1) \cdots \mu_n(B_n) = \nu_n(B_1 \times \cdots \times B_n)$. Thus, let us define $\mathbb{P} = \mu$ as the unique probability measure on $(\Omega, \mathcal{F})$ satisfying $\mu\left(\mathrm{Cyl}(B_1, \dots, B_n)\right) = \nu_n(B_1 \times \cdots \times B_n) = \mu_1(B_1) \cdots \mu_n(B_n)$ for all $B_1, \dots, B_n \in \mathcal{B}(\mathbb{R})$, which is guaranteed by Kolmogorov's extension theorem.

The existence of an infinite sequence of independent random variables follows now by defining the projection mappings $X_i(\omega) := \omega_i$ for all $i = 1, 2, \dots$ and $\omega = (\omega_1, \omega_2, \dots) \in \Omega$. Now for any $B \in \mathcal{B}(\mathbb{R})$, we have $X_i^{-1}(B) = \{\omega \in \Omega \; : \; \omega_i \in B\}$, which is a cylinder set and hence belongs to $\mathcal{F}$. Thus, X_i is a random variable. Let now $n \in \{1, 2, \dots\}$ and $i_1, \dots, i_n$ be distinct indices, which can be assumed to be increasing without loss of generality. Set

$B_k = \mathbb{R}$ for $k \notin \{i_1, \dots, i_n\}$ and let $B_{i_1}, \dots, B_{i_n} \in \mathcal{B}(\mathbb{R})$ be arbitrary. Then

$$\begin{aligned}\mathbb{P}(X_{i_1} \in B_{i_1}, \dots, X_{i_n} \in B_{i_n}) = \mu\left(\mathrm{Cyl}(B_1, \dots, B_{i_n})\right) &= \prod_{k=1}^{i_n} \mu_k(B_k) \\ &= \prod_{k=1}^{n} \mu_{i_k}(B_{i_k}) = \prod_{k=1}^{n} \mathbb{P}(X_{i_k} \in B_{i_k}).\end{aligned}$$

By Definition 65, the random variables X_i, $i = 1, 2, \dots,$ are hence mutually independent.

Remark 111. In fact, Kolmogorov's extension theorem holds in a slightly more general setting. In particular, the index set $\mathbb{N}$ in the product measure can be replaced by an arbitrary index set (with a small extra assumption), and $\mathbb{R}$ can be replaced by a general Polish space (i.e. a complete and separable metric space).

Remark 112. As an exercise, the reader is encouraged to try an alternative construction of a sequence of independent random variables using digits of the binary expansion of a uniformly distributed random variable on $(0, 1)$. This construction is as follows: First, write a uniform random variable U as

$$U = \sum_{i=1}^{\infty} \frac{X_i}{2^i} \text{ and } \forall i \geq 1, X_i \in \{0, 1\},$$

and then show that the random variables $X_1, X_2, \dots$ are independent and identically distributed (i.i.d.) uniform on $\{0, 1\}$. After this, consider a bijection between $\mathbb{N}$ and $\mathbb{N} \times \mathbb{N}$, showing then that it is possible to construct a bi-indexed sequence of i.i.d. random variables $(X_{n,i})_{n\geq1,i\geq1}$ with the same law as above. Deduce from this that the σ-algebras $\sigma(X_{n,1}, X_{n,2}, \dots)$ are mutually independent, and that there exists a sequence $(U_n)_{n\geq1}$ of independent uniform random variables on $(0, 1)$. Finally, if μ_n $(n = 1, 2, \dots)$ are any given laws, apply this to show that there is an infinite sequence of independent random variables Y_n with laws μ_n, respectively.

Chapter 4

Introduction to percolation

In this chapter, we apply some of the above developed theory to a classical model originating from polymer chemistry, which has then drawn much attention of the statistical physics community over decades. This model, called (bond) *percolation*, turns out to be surprisingly simple to define mathematically, yet analyzing it in detail requires deep mathematical tools from measure-theoretic probability. Many major breakthroughs have been achieved, and percolation has also been a focal point of research interests of three fields medalists (Wendelin Werner, Stanislav Smirnov and Hugo Duminil-Copin) as well as numerous other prominent mathematicians. Many generalizations have been investigated since then, and the research around percolation continues strong for the time being; see eg. the seminar *Percolation today*.

4.1 Definition of the bond percolation model

We start by introducing the model in its simplest and most classical form on a d-dimensional square lattice, i.e. we consider the Bernoulli bond percolation on $\mathbb{Z}^d$. This model was introduced by Broadbent and Hammersley in 1957. We then proceed to some of the elementary results concerning a *phase transition*. Denote by $\mathbb{L}^d$ the pair $(\mathbb{Z}^d, E^d)$, where $\mathbb{Z}^d = \{(z_1, \ldots, z_d) \ : \ z_i \in \mathbb{Z}, \ i = 1, \ldots, d\}$ and $E^d = \{(x, y) \ : \ x, y \in \mathbb{Z}^d, \ |x - y|_1 = 1\}$, where $|x - y|_1 = |x_1 - y_1| + \cdots + |x_d - y_d|$. In other words, E^d is the set of nearest neighbor pairs of elements in $\mathbb{Z}^d$, which we call *edges*; we also denote $x \sim y$ if $(x, y) \in E^d$. We call a $z \in \mathbb{Z}^d$ a *vertex*, and thus view the pair $(\mathbb{Z}^d, E^d)$ as an infinite graph (so that an interested reader could easily proceed to study the model on other graphs).

To construct the probability space, we decide for each edge whether it is *open* or *closed*, which is formally done as follows. Let $\Omega := \{0, 1\}^{E^d}$, that is, we assign to each edge either value 0 (closed) or 1 (open). Intuitively, if we model eg. porous medium as in the original motivation of percolation, an edge is open if it lets liquid pass through. Let $\mathcal{F}$ be the σ-algebra generated by the cylinder sets of the form $A_{e_1} \times \cdots \times A_{e_n} \times \{0, 1\} \times \cdots$, where $n \in \mathbb{N}$ and $e_1, \ldots, e_n$ are some edges on $\mathbb{L}^d$. Fix $p \in [0, 1]$, representing the probability of a single edge being open. Formally, for each $e \in E^d$, let ω_e be a Bernoulli(p)-distributed random variable, i.e. with law given by

$$\begin{cases} \mathbb{P}_p(\omega_e = 1) = p \\ \mathbb{P}_p(\omega_e = 0) = 1 - p \end{cases} \tag{4.1}$$

independently for all $e \in E^d$. Then (by a slight abuse of notation), define a probability measure $\mathbb{P}_p$ on Ω as the product measure of Bernoulli distributions over all edges $e \in E^d$.

We call a subgraph of $\mathbb{L}^d$ induced by open edges a *configuration.* A basic question to ask about a configuration is whether there exists an infinite connected component of open edges. We call a component in a configuration a *cluster*, so our question is to determine whether there exists an infinite cluster or not. In the intuitive picture from porous medium, an infinite cluster would mean that liquid could run through the whole system.

Quite remarkably, it turns out that Kolmogorov's zero-one law gives us a solution key to this problem. More precisely, for a fixed enumeration of the edges $e_1, e_2, \dots$ and each $n = 1, 2, \dots,$ define $\mathcal{F}_n := \{\emptyset, \Omega, \{\omega_{e_n} = 1\}, \{\omega_{e_n} = 0\}\}$. Now obviously each $\mathcal{F}_n$ is a σ-algebra, and they are mutually independent by construction. Hence, we may define the tail σ-algebra $\mathcal{T} = \bigcap_{n=1}^{\infty} \sigma(\mathcal{F}_n, \mathcal{F}_{n+1}, \dots)$. Now by Kolmogorov's $0-1$-law, each event $A \in \mathcal{T}$ satisfies $\mathbb{P}_p(A) \in \{0, 1\}$. It is also easy to see that $\{\exists \text{ infinite cluster}\} \in \mathcal{T}$, since this event does not depend on removing a finite number of edges. Hence, we have deduced $\mathbb{P}_p(\exists \text{ infinite cluster}) \in \{0, 1\}$. It thus remains a question how this probability depends on the parameter p.

4.2 Percolation probabilities and critical point

Let us now fix some notation which is common in percolation theory. First, let $x, y \in \mathbb{Z}^d$, and let us denote $x \leftrightarrow y$ if there exists a path of open edges linking x and y. Furthermore, let us denote $x \leftrightarrow \infty$ if x belongs to an infinite cluster (i.e. there is a path of infinite length starting from x). Then $\{x \leftrightarrow y\} \in \mathcal{F}$ and $\{x \leftrightarrow \infty\} \in \mathcal{F}$. The former holds since it only depends on a finite number of edges, and the latter is a countable intersection of events stating that the cluster of x contains a vertex of distance n from x. Next, we introduce two functions which play a central role in the analysis of the p-dependency of the infinite cluster probability.

Definition 113. The *connectivity function* is the function $\theta : [0, 1] \to [0, 1]$ given by $\theta(p) = \mathbb{P}_p(0 \leftrightarrow \infty)$. The existence of an infinite cluster is described by the function $\psi : [0, 1] \to [0, 1]$, $\psi(p) = \mathbb{P}_p(\exists \text{ infinite cluster}) = \mathbb{P}_p\left(\bigcup_{x \in \mathbb{Z}^d} \{x \leftrightarrow \infty\}\right)$.

The first big question in percolation theory is whether there is a *critical value* of p, denoted by p_c, such that $\psi(p) = 0$ if $p \in [0, p_c)$ and $\psi(p) = 1$ if $p \in (p_c, 1]$. Moreover, perhaps even more important subsequent question is to determine what happens at $p = p_c$, (which still contains a myriad of open problems in most dimensions). During the rest of the chapter, we just scratch the surface of this big program, providing some ideas to an extensive further study.

Proposition 114. *The functions θ and ψ are increasing.*

Proof. Let $(U_e)_{e \in E^d}$ be independent uniform random variables on $[0, 1]$. Note that the existence of this collection is guaranteed by the previous chapter. Then, let $p \in [0, 1]$ and let us define a percolation configuration by setting for each edge $e \in E^d$

$$\omega_e = \begin{cases} 1 & \text{if } U_e \leq p \\ 0 & \text{if } U_e > p. \end{cases}$$

Let us show that $\mathbb{P}_p$ is the law of this configuration ω. Indeed, if $\mathbb{P}$ is now the product measure induced by the uniform measures, we have

$$\begin{cases} \mathbb{P}(\omega_e = 1) = p \\ \mathbb{P}(\omega_e = 0) = 1 - p \end{cases}$$

for all $e \in E^d$. By definition, the random variables ω_e are measurable functions of independent random variables U_e, and hence by (a slight generalization of) the regrouping lemma, they are themselves mutually independent. Thus by the uniqueness of the product measure provided by Kolmogorov's extension theorem, $\mathbb{P}_p$ is the law of ω.

Next, let $q \in [p, 1]$, and let us define

$$\omega'_e = \begin{cases} 1 & \text{if } U_e \leq q \\ 0 & \text{if } U_e > q. \end{cases}$$

Now similarly as above, ω' has law $\mathbb{P}_q$, and furthermore by its definition, $\omega_e \leq \omega'_e$ for all $e \in E^d$. This implies that if $\omega \in \{0 \leftrightarrow \infty\}$, then also $\omega' \in \{0 \leftrightarrow \infty\}$. Hence, $\theta(p) = \mathbb{P}_p(0 \leftrightarrow \infty) = \mathbb{P}(\{\omega : 0 \leftrightarrow \infty\}) \leq \mathbb{P}(\{\omega' : 0 \leftrightarrow \infty\}) = \mathbb{P}_q(0 \leftrightarrow \infty) = \theta(q)$. The claim for ψ is proven similarly. □

Remark 115. In the above proof, the measure $\mathbb{P}$ is a *coupling* measure between $\mathbb{P}_p$ and $\mathbb{P}_q$, $0 \leq p \leq q \leq 1$. Coupling is a general technique in probability theory, which is used eg. when one wants to construct a common probability space in order to compare different probability measures (as above).

Definition 116 (Critical point)**.** The *critical probability* for percolation on $\mathbb{L}^d$ is the number $p_c := p_c(\mathbb{Z}^d) := \inf\{p \in [0,1] : \theta(p) > 0\}$.

Next, we analyze the *phase transition* around $p = p_c$ a bit further.

Proposition 117. *The percolation on $\mathbb{L}^d$ has the following phases.*

- *If $0 \leq p < p_c$, then $\theta(p) = \psi(p) = 0$.*
- *If $p > p_c$, then $\theta(p) > 0$ and $\psi(p) = 1$.*

Proof. The claim for θ follows directly from the definition of p_c. For the rest, recall that $\psi(p) = \mathbb{P}_p(\exists \text{ infinite cluster}) = \mathbb{P}_p\left(\bigcup_{x \in \mathbb{Z}^d}\{x \leftrightarrow \infty\}\right) \in \{0, 1\}$ by Kolmogorov's $0-1$-law, since $\bigcup_{x \in \mathbb{Z}^d}\{x \leftrightarrow \infty\} \in \mathcal{T}$. Let $p < p_c$. Then $\mathbb{P}_p(0 \leftrightarrow \infty) = \theta(p) = 0$, and obviously $\mathbb{P}_p(x \leftrightarrow \infty) = \mathbb{P}_p(0 \leftrightarrow \infty) = 0$ for all $x \in \mathbb{Z}^d$ (by the translational symmetry of $\mathbb{Z}^2$). Using the subadditivity of $\mathbb{P}$, we then deduce $\psi(p) = \mathbb{P}_p\left(\bigcup_{x \in \mathbb{Z}^d}\{x \leftrightarrow \infty\}\right) \leq \sum_{x \in \mathbb{Z}^d} \mathbb{P}_p(x \leftrightarrow \infty) = 0$. Then, let $p > p_c$. We simply bound ψ from below, $\psi(p) \geq \mathbb{P}_p(0 \leftrightarrow \infty) = \theta(p) > 0$. Hence, $\psi(p) = 1$. □

Theorem 118. *The phase transition of percolation on $\mathbb{L}^d$ is*

- *trivial if $d = 1$: $p_c(\mathbb{Z}) = 1$;*
- *non-trivial if $d \geq 2$: $0 < p_c(\mathbb{Z}^d) < 1$.*

Proof. We only show the claim for $d = 1$. For the non-triviality claim, see [1]. That said, assume $p < 1$, and note that $\theta(1) = 1$ by definition. Now for all $e \in E^1$, the events $\{\omega_e = 0\}$ are independent with probability $\mathbb{P}(\omega_e = 0) = 1 - p$. Thus by the 2nd Borel-Cantelli lemma, $\omega_e = 0$ infinitely often. We apply the lemma to both sides of the origin, showing that there are infinitely many closed edges on both sides, implying that a.s. an open infinite cluster containing the origin cannot occur. Hence, $\theta(p) = \mathbb{P}_p(0 \leftrightarrow \infty) = 0$. □

Finally, we collect some remarkable results about the existence of the infinite cluster. For proofs, see [1] and the references therein.

- *Uniqueness of the infinite cluster*: Let $p > p_c$. Then $\mathbb{P}_p(\exists! \text{ infinite cluster}) = 1$. (Aizenman - Kesten - Newman, 1987)
- When $d = 2$, $p_c = p_c(\mathbb{Z}^2) = \frac{1}{2}$ and there is no infinite cluster at $p = p_c$ (and in particular, $\theta(p_c) = 0$). (Kesten, 1980)
- In fact, $p_c \geq \frac{1}{2}$ for all d.

It is conjectured that $\theta(p_c) = 0$ for all $d \geq 2$. However, no exact formula for p_c is found or expected to be found in $d \geq 3$. This is one reason why most of the developments about the *critical* percolation have been found in $d = 2$. Another reason is that in $d = 2$, the *interfaces* (i.e. separating domain walls) between open and closed components are random curves rather than random (hyper)surfaces, so their analysis is way simpler. In fact, even though there is no infinite cluster at $p = p_c$, the interfaces between finite clusters are nowadays well understood, thanks to their conformally invariant scaling properties (reflecting the fact that the $p = p_c$ case corresponds to a 2-dimensional conformal field theory in statistical physics). We do not discuss these properties further, but an interested reader could see a glimpse in [1].

Chapter 5

Limit theorems

In this chapter, we move on from sequences of random variables to their limits. First, we introduce some important notions of convergence in probability theory and study their relations with one another. Then, we proceed to infinite sums of random variables and investigate their asymptotic behavior at infinity. This leads us to the *laws of large numbers.* Finally, we investigate *convergence in distribution*, which is the weakest yet perhaps most useful form of convergence in probability theory. This form of convergence is generally suitable for the *central limit theorems.*

5.1 Convergence in probability and almost sure convergence

Definition 119 (Convergence in probability)**.** Let $X_1, X_2, \dots$ and X be real-valued random variables on the same probability space. Then $X_n \xrightarrow[n\to\infty]{} X$ *in probability* if

$$\mathbb{P}(|X_n - X| > \epsilon) \xrightarrow[n\to\infty]{} 0$$

for all $\epsilon > 0$. In this case, we denote $X_n \xrightarrow[n\to\infty]{\mathbb{P}} X$.

Remark 120. The above definition means that for all $\epsilon > 0$ and $\delta > 0$, there exist an index $N \in \mathbb{N}$ such that $\mathbb{P}(|X_n - X| \leq \epsilon) > 1 - \delta$ whenever $n > N$. That is, with arbitrarily high probability, X_n is eventually close to X.

Definition 121 (Convergence a.s.)**.** Let us consider the setting of Definition 119. Then $X_n \xrightarrow[n\to\infty]{} X$ *almost surely* if there exists an event E with $\mathbb{P}(E) = 1$ such that for all $\omega \in E$, $X_n(\omega) \xrightarrow[n\to\infty]{} X(\omega)$. In this case, we denote $X_n \xrightarrow[n\to\infty]{a.s.} X$.

Remark 122 (Choice of a measurable E)**.** Let $(X_i)_{i\geq 1}$ be a sequence of random variables, and consider $E = \{\omega : \lim_{n\to\infty} X_n(\omega) \text{ exists}\} = \{\omega : \limsup_{n\to\infty} X_n(\omega) = \liminf_{n\to\infty} X_n(\omega)\}$. Now the set E is measurable, since the random variables $\limsup_{n\to\infty} X_n$ and $\liminf_{n\to\infty} X_n$ are. Namely,

$$\left\{\limsup_{n\to\infty} X_n \geq x\right\} = \bigcap_{k=1}^{\infty} \left\{\left\{X_n \geq x - \frac{1}{k}\right\} \text{ occurs i.o.}\right\} = \bigcap_{k=1}^{\infty} \left\{\limsup_{n} \left\{X_n \geq x - \frac{1}{k}\right\}\right\},$$

where we recall that the lim sup of any sequence of measurable sets is measurable. The measurability of $\liminf_{n\to\infty} X_n$ follows by noting that $\liminf_{n\to\infty} X_n = -\limsup_{n\to\infty}(-X_n)$.

Next, we show that a.s. convergence is at least as strong as convergence in probability, i.e. a.s. convergence always implies convergence in probability. However, a converse statement would only hold when one passes to subsequences.

Proposition 123. *(i) Assume $X_n \xrightarrow[n\to\infty]{a.s.} X$. Then $X_n \xrightarrow[n\to\infty]{\mathbb{P}} X$.*

(ii) Assume $X_n \xrightarrow[n\to\infty]{\mathbb{P}} X$. Then there exists a subsequence $(n_k)_{k\geq 1}$ such that $X_{n_k} \xrightarrow[k\to\infty]{a.s.} X$.

Proof. (i) Assume $X_n \xrightarrow[n\to\infty]{a.s.} X$, and let $\epsilon > 0$. Define events $E_m := \{|X_m - X| \leq \epsilon\}$. Then $\{X_n \xrightarrow[n\to\infty]{} X\} \subset \bigcup_{n=1}^{\infty}\bigcap_{m=n}^{\infty} E_m$, so $\mathbb{P}\left(\bigcup_{n=1}^{\infty}\bigcap_{m=n}^{\infty} E_m\right) = 1$. Moreover, we have

$$\mathbb{P}(|X_n - X| > \epsilon) = 1 - \mathbb{P}(E_n) \leq 1 - \mathbb{P}\left(\bigcap_{m=n}^{\infty} E_m\right).$$

Now by Lemma 15, the right hand side converges to zero as $n \to \infty$, showing that $X_n \xrightarrow[n\to\infty]{\mathbb{P}} X$.

(ii) Assume $X_n \xrightarrow[n\to\infty]{\mathbb{P}} X$, and let $\epsilon_k := \frac{1}{k}$ for all $k = 1, 2, \ldots$. By the assumption, $\mathbb{P}(|X_n - X| > \epsilon_1) \xrightarrow[n\to\infty]{} 0$, and thus there exist a subsequence $(n_k^{(1)})_{k\geq 1}$ such that $\sum_{k=1}^{\infty}\mathbb{P}\left(\left|X_{n_k^{(1)}} - X\right| > \epsilon_1\right) < \infty$. Now by the 1st Borel-Cantelli lemma, $\left|X_{n_k^{(1)}} - X\right| > \epsilon_1$ can only occur finitely many times a.s., implying that $\left|X_{n_k^{(1)}} - X\right| \leq \epsilon_1$ almost surely for k large enough. We can further iterate this idea by choosing a subsequence $(n_k^{(m+1)})_{k\geq 1}$ of $(n_k^{(m)})_{k\geq 1}$ such that $\left|X_{n_k^{(m+1)}} - X\right| \leq \epsilon_{m+1}$ almost surely for k large enough and $m = 1, 2, \ldots$. We may then choose the diagonal sequence $(n_k^{(k)})_{k\geq 1}$, which satisfies for all $n = 1, 2, \ldots$ $\left|X_{n_k^{(k)}} - X\right| \leq \frac{1}{n}$ a.s. provided k is large enough. Thit implies that almost surely, $X_{n_k^{(k)}} \to X$. □

The above proposition in fact provides a characterization for convergence in probability:

Corollary 124. $X_n \xrightarrow[n\to\infty]{\mathbb{P}} X$ *if and only if for all subsequences $(n_k)_{k\geq 1}$ there exist a subsequence $(n_{k_m})_{m\geq 0}$ such that $X_{n_{k_m}} \xrightarrow[m\to\infty]{a.s.} X$.*

Proof. It is enough to show the converse implication by assuming that $X_n \xrightarrow[n\to\infty]{\mathbb{P}} X$ does not hold. That is, there exists $\epsilon > 0$ and $\delta > 0$ and a subsequence $(n_k)_{k\geq 1}$ such that $\mathbb{P}(|X_{n_k} - X| > \epsilon) \geq \delta$ for all $k = 1, 2, \ldots$. But since by assumption there exists a subsequence $(n_{k_m})_{m\geq 1}$ of $(n_k)_{k\geq 1}$ such that $X_{n_{k_m}} \xrightarrow[m\to\infty]{a.s.} X$, we also have $X_{n_{k_m}} \xrightarrow[m\to\infty]{\mathbb{P}} X$. That is, $\mathbb{P}(\left|X_{n_{k_m}} - X\right| > \epsilon) \xrightarrow[m\to\infty]{} 0$, and hence the subsequence $(n_{k_m})_{m\geq 0}$ contradicts the assumption. □

Remark 125 (Convergence in probability is weaker than a.s. convergence)**.** Let $X_1, X_2, \dots$ be independent random variables with law given by $\mathbb{P}(X_n = 0) = 1 - \frac{1}{n}$ and $\mathbb{P}(X_n = 1) = \frac{1}{n}$. Then for all $\epsilon \in (0,1)$, $\mathbb{P}(X_n > \epsilon) = \frac{1}{n} \xrightarrow[n\to\infty]{} 0$, and hence $X_n \xrightarrow[n\to\infty]{\mathbb{P}} 0$. However, since the events $\{X_n = 1\}$ are independent and $\sum_{n=1}^{\infty} \mathbb{P}(X_n = 1) = \infty$, the second Borel-Cantelli lemma implies $X_n = 1$ i.o. This implies that X_n does not converge to 0 a.s.

5.2 Laws of large numbers

In this section, we study under which assumptions and in which sense does a rescaled sum of independent and identically distributed (i.i.d.) random variables converge towards their mean. The setting is as follows: Let $X_1, X_2, \dots$ be i.i.d. and $S_n := \sum_{i=1}^n X_i$. Then the question is, whether $\frac{S_n}{n}$ converges towards $m := \mathbb{E}(X_1)$ in some sense. To guarantee finite expectation, we assume at least that $\mathbb{E}(|X_1|) < \infty$. It will turn out that no other assumptions are needed; however, extra assumptions give better control (or *large deviation* bounds).

5.2.1 Weak law of large numbers

We proceed on studying sums of independent and identically distributed (i.i.d.) random variables. We prove first a weak form of *law of large numbers*, where the word *weak* refers to the convergence in probability.

Proposition 126 (Weak law of large numbers)**.** *Let $X_1, X_2, \dots$ be i.i.d. with finite second moment, i.e. $\mathbb{E}(X_1^2) < \infty$. Denote $S_n := \sum_{i=1}^n X_i$. Then $\frac{S_n}{n} \xrightarrow[n\to\infty]{\mathbb{P}} m$, where $m := \mathbb{E}(X_1)$.*

Proof. Let $\epsilon > 0$. Observe that $\mathrm{Var}(S_n) = \sum_{i=1}^n \mathrm{Var}(X_i) = n\,\mathrm{Var}(X_1)$ since the random variables X_i are i.i.d. By the finite second moment assumption, we may apply Chebyshev's inequality together with the assumption of identical distributions to deduce

$$\begin{aligned}\mathbb{P}\left(\left|\frac{S_n}{n} - m\right| > \epsilon\right) &= \mathbb{P}(|S_n - nm| > n\epsilon) = \mathbb{P}\left(\left|S_n - \sum_{i=1}^n \mathbb{E}(X_i)\right| > n\epsilon\right) \\ &= \mathbb{P}(|S_n - \mathbb{E}(S_n)| > n\epsilon) \le \frac{\mathrm{Var}(S_n)}{n^2\epsilon^2} = \frac{n\,\mathrm{Var}(X_1)}{n^2\epsilon^2} = \frac{\mathrm{Var}(X_1)}{\epsilon^2}\frac{1}{n} \xrightarrow[n\to\infty]{} 0.\end{aligned}$$

□

5.2.2 Strong law of large numbers

In this section, we study under which assumptions, and amendments of the proof of the weak law of large numbers, does the law of large numbers hold almost surely. We start by generalizing the proof under extra assumptions on the moments of the random variables, and then improve the assumptions yielding more general results. First, we apply the idea of the proof of the weak law of large numbers. However, since we want to show a stronger (a.s.) convergence property, we also need a stronger moment condition in order to directly apply a similar proof idea.

Proposition 127. *Let $X_1, X_2, \dots$ be i.i.d. with finite fourth moment, i.e. $\mathbb{E}(X_1^4) < \infty$. Denote $S_n := \sum_{i=1}^n X_i$. Then $\frac{S_n}{n} \xrightarrow[n\to\infty]{a.s.} m$, where $m := \mathbb{E}(X_1)$.*

Proof. Assume, without loss of generality, that $m = 0$ (otherwise one might consider $Y_n = X_n - m$). Now

$$\mathbb{E}(S_n^4) = \mathbb{E}\left(\left(\sum_{i=1}^{n} X_i\right)^4\right) = \mathbb{E}\left(\sum_{i,j,k,l=1}^{n} X_i X_j X_k X_l\right) = \sum_{i,j,k,l=1}^{n} \mathbb{E}(X_i X_j X_k X_l).$$

For $\mathbb{E}(X_i X_j X_k X_l)$, there are three possible cases. The first one is that some index, say i, is distinct from all the other indices. In that case, $\mathbb{E}(X_i X_j X_k X_l) = \mathbb{E}(X_i)\mathbb{E}(X_j X_k X_l) = 0$ by independence and the assumption $m = 0$. The second case is that two indices, say i and j, coincide and there is no single distinct index. Then $\mathbb{E}(X_i X_j X_k X_l) = \mathbb{E}(X_i X_j)\mathbb{E}(X_k X_l) = \mathbb{E}(X_i^2)\mathbb{E}(X_k^2) = \mathbb{E}(X_1^2)^2$. The final one is that all of them coincide, so $\mathbb{E}(X_i X_j X_k X_l) = \mathbb{E}(X_1^4)$. Therefore,

$$\mathbb{E}(S_n^4) = \sum_{1 \le i,j \le n, i \ne j} \mathbb{E}(X_1^2)^2 + \sum_{i=1}^{n} \mathbb{E}(X_1^4) \le C \cdot n^2$$

for some constant $C > 0$. Hence by Markov's inequality,

$$\mathbb{P}\left(\left|\frac{S_n}{n}\right| > \epsilon\right) = \mathbb{P}(S_n^4 > (n\epsilon)^4) \le \frac{\mathbb{E}(S_n^4)}{n^4 \epsilon^4} \le \frac{C}{\epsilon^4 n^2}.$$

Now $\sum_{n=1}^{\infty} \mathbb{P}\left(\left|\frac{S_n}{n}\right| > \epsilon\right) < \infty$, hence the first Borel-Cantelli lemma gives $\mathbb{P}(E_\epsilon) = 1$, where $E_\epsilon := \left\{\left|\frac{S_n}{n}\right| \le \epsilon \text{ eventually}\right\}$. Therefore, if we take $\epsilon = \frac{1}{k}$ and define $E := \bigcap_{k=1}^{\infty} E_{1/k}$, we have $\mathbb{P}(E) = \mathbb{P}\left(\bigcap_{k=1}^{\infty} E_{1/k}\right) = \lim_{k\to\infty} \mathbb{P}(E_{1/k}) = 1$. Since $\frac{S_n(\omega)}{n} \xrightarrow[n\to\infty]{} 0$ for all $\omega \in E$, the claim follows.

□

Remark 128. In the previous two proofs, we used *large deviations* estimates based on moment bounds of S_n. In the proof of the weak law of large numbers, we noticed (assuming $m = 0$) that $\mathbb{P}(|S_n| > n\epsilon) \le \frac{\mathrm{Var}(X_1)}{\epsilon^2}\frac{1}{n}$, which converges to zero but is not summable. To prove a strong law of large numbers, we needed summability in order to apply the 1st Borel-Cantelli lemma: $\mathbb{P}(S_n^4 > (n\epsilon)^4) \le \frac{C}{\epsilon^4 n^2}$. Here, we only considered even moments, since odd powers of S_n would result parity issues which would not make this approach work (the reader may consider S_n^3 to see where the problems arise).

More generally, if $\mathbb{E}(X_1^{2k}) < \infty$, one can show that $\mathbb{P}(|S_n| > n\epsilon) \le \frac{C}{n^k}$ for $k = 1, 2, \dots$ and some constant C depending on k and ϵ. Thus, assumptions on higher (even) moments give faster decaying large deviation bounds.

Proposition 129. *Let $X_1, X_2, \dots$ be i.i.d. with finite variance, i.e. $\mathbb{E}(X_1^2) < \infty$. Denote $S_n := \sum_{i=1}^{n} X_i$. Then $\frac{S_n}{n} \xrightarrow[n\to\infty]{a.s.} m$, where $m := \mathbb{E}(X_1)$.*

Proof. In the proof of the weak law of large numbers, we showed that $\mathbb{P}\left(\left|\frac{S_n}{n} - m\right| > \epsilon\right) \le \frac{\mathrm{Var}(X_1)}{\epsilon^2}\frac{1}{n}$, which is not summable. However, if we consider the subsequence $(n_k)_{k\ge 0}$ defined by $n_k := k^2$, we have $\mathbb{P}\left(\left|\frac{S_{n_k}}{n_k} - m\right| > \epsilon\right) \le \frac{\mathrm{Var}(X_1)}{\epsilon^2}\frac{1}{k^2}$ which is summable. Applying then the 1st Borel-Cantelli lemma as in the previous proof, we find $\frac{S_{n_k}}{n_k} \xrightarrow[n\to\infty]{a.s.} m$. The rest of the proof is to show that we can in fact pass from this subsequence to the whole sequence $n = 1, 2, \dots$.

First, assume $X_i \geq 0$ for all $i = 1, 2, \dots$, and fix a large enough n. Then there exists a k such that $n_k \leq n \leq n_{k+1}$, and moreover by the non-negativity of $X_1, \dots, X_n$, we have $S_{n_k} \leq S_n \leq S_{n_{k+1}}$. Furthermore, $\frac{S_{n_k}}{n_{k+1}} \leq \frac{S_n}{n} \leq \frac{S_{n_{k+1}}}{n_k}$, and since $\frac{S_{n_k}}{n_{k+1}} = \frac{S_{n_k}}{n_k}\frac{n_k}{n_{k+1}} = \frac{S_{n_k}}{n_k}\frac{k^2}{(k+1)^2}$, we see that $\frac{S_{n_k}}{n_{k+1}} \xrightarrow[k\to\infty]{a.s.} m$ (and similarly, $\frac{S_{n_{k+1}}}{n_k} \xrightarrow[k\to\infty]{a.s.} m$). That is, there exists an event E with $\mathbb{P}(E) = 1$ such that for all $\omega \in E$, we have $\frac{S_{n_k}(\omega)}{n_{k+1}} \xrightarrow[k\to\infty]{} m$ and $\frac{S_{n_{k+1}}(\omega)}{n_k} \xrightarrow[k\to\infty]{} m$. Since $\frac{S_n(\omega)}{n} \in \left[\frac{S_{n_k}(\omega)}{n_{k+1}}, \frac{S_{n_{k+1}}(\omega)}{n_k}\right]$, this shows that also $\frac{S_n(\omega)}{n} \xrightarrow[n\to\infty]{} m$ for all $\omega \in E$. That yields $\frac{S_n(\omega)}{n} \xrightarrow[n\to\infty]{a.s.} m$.

In the general case, we write $X_i = X_i^+ - X_i^-$, where the random variables $X_i^+ \geq 0$ and $X_i^- \geq 0$ are defined in the way familiar from measure theory:

$$X_i^+ = \begin{cases} X_i & \text{if } X_i \geq 0 \\ 0 & \text{else} \end{cases}, \qquad X_i^- = \begin{cases} 0 & \text{if } X_i \geq 0 \\ -X_i & \text{else.} \end{cases}$$

By the second moment assumption $\mathbb{E}(X_i^2) = \mathbb{E}(X_1^2) < \infty$, we have $\mathbb{E}((X_i^+)^2) < \infty$ and $\mathbb{E}((X_i^-)^2) < \infty$. Moreover, $(X_i^+)_{i\geq 1}$ (resp. $(X_i^-)_{i\geq 1}$) is a sequence of images of i.i.d. random variables $(X_i)_{i\geq 1}$ in a measurable mapping, hence an i.i.d. sequence of random variables. Therefore by the previous consideration for non-negative random variables, we have $\frac{\sum_{i=1}^n X_i^+}{n} \xrightarrow[n\to\infty]{a.s.} \mathbb{E}(X_1^+)$ and $\frac{\sum_{i=1}^n X_i^-}{n} \xrightarrow[n\to\infty]{a.s.} \mathbb{E}(X_1^-)$. Hence by linearity,

$$\frac{S_n}{n} = \frac{\sum_{i=1}^n X_i^+ - \sum_{i=1}^n X_i^-}{n} \xrightarrow[n\to\infty]{a.s.} \mathbb{E}(X_1^+) - \mathbb{E}(X_1^-) = \mathbb{E}(X_1) = m.$$

□

We are now ready to give the most general statement for the strong law of large numbers:

Theorem 130 (Strong law of large numbers)**.** *Let $X_1, X_2, \dots$ be i.i.d. with $\mathbb{E}(|X_1|) < \infty$. Denote $S_n := \sum_{i=1}^n X_i$. Then $\frac{S_n}{n} \xrightarrow[n\to\infty]{a.s.} m$, where $m := \mathbb{E}(X_1)$.*

Proof. As in the previous proof, we may assume $X_i \geq 0$ for all $i = 1, 2, \dots$ (otherwise we decompose them in the positive and negative parts). This time, in order to mitigate the lack of the second moment bound on X_i, we introduce the truncated random variables

$$Y_n := X_n \mathbb{1}_{X_n \leq n} = \begin{cases} X_n & \text{if } X_n \leq n \\ 0 & \text{else.} \end{cases}$$

Let also $T_n := \sum_{i=1}^n Y_i$.

Step 1. Let us show that $\frac{T_n}{n} \xrightarrow[n\to\infty]{a.s.} m$ implies $\frac{S_n}{n} \xrightarrow[n\to\infty]{a.s.} m$. First, we note that the definition of Y_n gives $\mathbb{P}(X_n \neq Y_n) = \mathbb{P}(X_n > n) = \mathbb{P}(X_1 > n)$, and thus $\sum_{n=1}^\infty \mathbb{P}(X_n \neq Y_n) = \sum_{n=1}^\infty \mathbb{P}(X_1 > n) \leq \int_0^\infty \mathbb{P}(X_1 > t)dt = \mathbb{E}(X_1) < \infty$. The inequality followed from the fact that $t \mapsto \mathbb{P}(X_1 > t)$ is a decreasing function of t, and thus $\mathbb{P}(X_1 > n+1) \leq \int_n^{n+1} \mathbb{P}(X_1 > t)dt$

for all $n \in \mathbb{N}$ (this is a Riemann lower sum approximation of the integral). The last equality is left as an exercise to the reader, following from the definition of the expectation and Fubini's theorem.

By the 1st Borel-Cantelli lemma, we then obtain that a.s. $X_n = Y_n$ eventually. That is, if we denote the event $E := \{X_n = Y_n \ \text{ eventually}\}$, then $\mathbb{P}(E) = 1$. If $\omega \in E$, then there exists such an $N = N(\omega) > 0$ that $X_n(\omega) = Y_n(\omega)$ for all $n > N$. Hence for all $\omega \in E$,

$$\frac{|T_k(\omega) - S_k(\omega)|}{k} \leq \frac{\sum_{n=1}^{k} |Y_n(\omega) - X_n(\omega)|}{k} \leq \frac{\sum_{n=1}^{N} |X_n(\omega)|}{k} \xrightarrow[k\to\infty]{} 0.$$

Therefore $\frac{T_n(\omega)}{n} \xrightarrow[n\to\infty]{} m$ implies $\frac{S_n(\omega)}{n} \xrightarrow[n\to\infty]{} m$ for all $\omega \in E$, giving the desired a.s. convergence for $\frac{S_n}{n}$.

Step 2. Let $\alpha > 1$ and consider the subsequence $n_k = \lfloor \alpha^k \rfloor$. We show in this step that $\frac{T_{n_k}}{n_k} \xrightarrow[k\to\infty]{a.s.} m$ implies $\frac{T_n}{n} \xrightarrow[n\to\infty]{a.s.} m$ (which combined with step 1 implies $\frac{S_n}{n} \xrightarrow[n\to\infty]{a.s.} m$). In this step, we generalize the idea in the proof of Proposition 129 as follows. First, fix a large enough n, and note that then there exists a k such that $n_k \leq n \leq n_{k+1}$. By the non-negativity of $Y_1, \dots, Y_n$, we then have $T_{n_k} \leq T_n \leq T_{n_{k+1}}$, and again, $\frac{T_{n_k}}{n_{k+1}} \leq \frac{T_n}{n} \leq \frac{T_{n_{k+1}}}{n_k}$. Since $\frac{T_{n_k}}{n_{k+1}} = \frac{T_{n_k}}{n_k}\frac{n_k}{n_{k+1}} = \frac{T_{n_k}}{n_k}\frac{\lfloor \alpha^k \rfloor}{\lfloor \alpha^{k+1} \rfloor}$, we see that $\frac{T_{n_k}}{n_{k+1}} \xrightarrow[k\to\infty]{a.s.} \frac{m}{\alpha}$ (and similarly, $\frac{T_{n_{k+1}}}{n_k} \xrightarrow[k\to\infty]{a.s.} \alpha m$). Therefore, $\frac{m}{\alpha} \leq \liminf_{n\to\infty} \frac{T_n}{n} \leq \limsup_{n\to\infty} \frac{T_n}{n} \leq \alpha m$ a.s. for all $\alpha > 1$. Letting then $\alpha \searrow 1$ gives $\liminf_{n\to\infty} \frac{T_n}{n} = \limsup_{n\to\infty} \frac{T_n}{n} = m$ a.s., which finally yields $\frac{T_n}{n} \xrightarrow[n\to\infty]{a.s.} m$.

Step 3. In this technical step, we finally show $\frac{T_{n_k}}{n_k} \xrightarrow[k\to\infty]{a.s.} m$. The idea is, as in the previous proofs, to apply a version of Markov's inequality. Observe now that the random variables Y_i are independent with $\mathbb{E}(Y_i^2) \leq i^2 < \infty$ for all $i = 1, 2, \dots$, hence we can even apply Chebyshev's inequality as follows:

$$\begin{aligned}\sum_{k=1}^{\infty} \mathbb{P}\left(\left|\frac{T_{n_k}}{n_k} - \frac{\mathbb{E}(T_{n_k})}{n_k}\right| > \epsilon\right) &\leq \sum_{k=1}^{\infty} \frac{\mathrm{Var}(T_{n_k})}{\epsilon^2 n_k^2} = \sum_{k=1}^{\infty} \frac{1}{\epsilon^2 n_k^2} \sum_{j=1}^{n_k} \mathrm{Var}(Y_j) \\ &= \frac{1}{\epsilon^2} \sum_{k=1}^{\infty} \frac{1}{n_k^2} \sum_{j=1}^{\infty} \mathrm{Var}(Y_j) \mathbb{1}_{j \leq n_k} = \frac{1}{\epsilon^2} \sum_{j=1}^{\infty} \mathrm{Var}(Y_j) \sum_{k\geq 1 : n_k \geq j} \frac{1}{n_k^2}.\end{aligned}$$

In the last equality, we applied Fubini's theorem to the sums of positive functions.

For the last sum, recall $n_k = \lfloor \alpha^k \rfloor$, where $\alpha > 1$, and denote $k_0 := \inf\{k \geq 1 : \alpha^k \geq j\}$. Observe that for k large enough, $n_k \geq \frac{1}{2}\alpha^k$. Thus,

$$\sum_{k\geq 1 : n_k \geq j} \frac{1}{n_k^2} \leq 4 \sum_{k\geq 1 : n_k \geq j} \frac{1}{\alpha^{2k}} \leq 4 \sum_{k=k_0}^{\infty} \frac{1}{\alpha^{2k}} \leq \frac{4}{\alpha^{2k_0}} \sum_{k=0}^{\infty} \frac{1}{\alpha^{2k}} = \frac{4}{\alpha^{2k_0}} \frac{1}{1-\alpha^{-2}} \leq \frac{C}{j^2}$$

where $C := \frac{4}{1-\alpha^{-2}}$. On the other hand,

$$\begin{aligned}\mathrm{Var}(Y_j) = \mathbb{E}(Y_j^2) - \mathbb{E}(Y_j)^2 \leq \mathbb{E}(Y_j^2) &= \int_0^{\infty} 2t\mathbb{P}(Y_j > t)dt = \int_0^{j} 2t\mathbb{P}(Y_j > t)dt \\ &\leq \int_0^{j} 2t\mathbb{P}(X_j > t)dt = \int_0^{j} 2t\mathbb{P}(X_1 > t)dt\end{aligned}$$

which follows from the general identity $\mathbb{E}(Z^k) = \int_0^\infty kt^{k-1}\mathbb{P}(Z > t)dt$, $k = 1, 2, \dots$, for a non-negative random variable Z (exercise). Putting these together, we obtain

$$\sum_{j=1}^{\infty} \operatorname{Var}(Y_j) \sum_{k \geq 1: n_k \geq j} \frac{1}{n_k^2} \leq C \sum_{j=1}^{\infty} \frac{1}{j^2} \int_0^\infty 2t\mathbb{P}(X_1 > t)\mathbb{1}_{t<j} dt = C \int_0^\infty 2t\mathbb{P}(X_1 > t) \sum_{j>t} \frac{1}{j^2} dt$$
$$\leq C \int_0^\infty 2t\mathbb{P}(X_1 > t) \int_t^\infty \frac{1}{x^2} dx dt = 2C \int_0^\infty \mathbb{P}(X_1 > t)dt = 2C\mathbb{E}(X_1) < \infty.$$

Therefore,

$$\sum_{k=1}^{\infty} \mathbb{P}\left(\left| \frac{T_{n_k}}{n_k} - \frac{\mathbb{E}(T_{n_k})}{n_k} \right| > \epsilon \right) \leq \frac{2C\mathbb{E}(X_1)}{\epsilon^2} < \infty,$$

so the 1st Borel-Cantelli lemma implies that for all $\epsilon > 0$, a.s. $\left| \frac{T_{n_k}}{n_k} - \frac{\mathbb{E}(T_{n_k})}{n_k} \right| \leq \epsilon$ for large enough k. As in the previous proofs, it then implies $\frac{T_{n_k}}{n_k} - \frac{\mathbb{E}(T_{n_k})}{n_k} \xrightarrow[k\to\infty]{a.s.} 0$.

It remains to show $\frac{\mathbb{E}(T_{n_k})}{n_k} \xrightarrow[k\to\infty]{} m$. In fact, this is a straightforward consequence of the dominated convergence theorem:

$$\mathbb{E}(T_{n_k}) = \mathbb{E}\left(\sum_{j=1}^{n_k} Y_j \right) = \mathbb{E}\left(\sum_{j=1}^{n_k} X_j \mathbb{1}_{X_j \leq j} \right) = \mathbb{E}(S_{n_k}) - \sum_{j=1}^{n_k} \mathbb{E}(X_j \mathbb{1}_{X_j > j})$$
$$= mn_k - \sum_{j=1}^{n_k} \mathbb{E}(X_1 \mathbb{1}_{X_1 > j}).$$

Observe that since $\mathbb{E}(X_1 \mathbb{1}_{X_1>j}) \leq \mathbb{E}(X_1) < \infty$, the dominated convergence theorem gives $\lim_{j\to\infty} \mathbb{E}(X_1 \mathbb{1}_{X_1>j}) = \mathbb{E}(X_1 \lim_{j\to\infty} \mathbb{1}_{X_1>j}) = 0$. Hence the Cesàro mean converges to zero, $\frac{1}{n_k} \sum_{j=1}^{n_k} \mathbb{E}(X_1 \mathbb{1}_{X_1>j}) \xrightarrow[k\to\infty]{} 0$. (This follows from a general theorem, but can also be seen directly by the following argument: Denoting $n_k := l$ and $\mathbb{E}(X_1 \mathbb{1}_{X_1>j}) := a_j$, we have for all $\epsilon > 0$ that there exists $N > 0$ large enough such that $a_j < \epsilon$ if $j > N$. Now $\frac{1}{l} \sum_{j=1}^{l} a_j = \frac{1}{l} \sum_{j=1}^{N} a_j + \frac{1}{l} \sum_{j=N+1}^{l} a_j$, where $\frac{1}{l} \sum_{j=1}^{N} a_j \xrightarrow[l\to\infty]{} 0$ and $\frac{1}{l} \sum_{j=N+1}^{l} a_j \leq \frac{l-N}{l}\epsilon$, which can be made arbitrarily small.) Combining this with the previous a.s. convergence, we conclude $\frac{T_{n_k}}{n_k} \xrightarrow[k\to\infty]{a.s.} \mathbb{E}(X_1) = m$.

□

5.3 Convergence in distribution

Recall that convergence in probability and a.s. convergence required the random variables to be defined on the same probability space. This is a rather strong restriction. For example, a typical situation arises when one wants to approximate distributions arising from a large discrete random system by a continuous limit distribution. In order to have a well-defined limit distribution, one needs to consider a form of convergence which takes into account the lack of the common probability space. A classical example would be the *normal approximation* given by the *central limit theorem*, which we will consider later in Chapter 7.

A solution would be given by the *convergence in distribution*, which is also known as the *weak convergence* (for reasons we will explain later in this section). Roughly speaking, this form of convergence would allow us to compare statistical properties of random variables, which are in some sense independent of the probability space itself, but rather only depend on the distribution. Next, we will define this for distributions of real-valued random variables, and comment later how to generalize it to metric spaces.

Definition 131 (Convergence in distribution)**.** Let $F_1, F_2, \dots$ and F be distribution functions (as in Definition 33 and Theorem 36). Then F_n converges to F *weakly* if $F_n(x) \xrightarrow[n\to\infty]{} F(x)$ for every $x \in \mathbb{R}$ such that F is continuous at x. We say that random variables $X_1, X_2, \dots$ converge *in distribution* to X if $F_{X_n} \xrightarrow[n\to\infty]{} F_X$ weakly, where F_Y denotes the distribution function of a random variable Y. In this case, we denote $X_n \xrightarrow[n\to\infty]{d} X$.

Remark 132. Let us stress that the random variables X_n in the definition of the convergence in distribution do not need to be defined on the same probability space, but rather their statistical properties are assumed to be close to each other.

Example 133. The requirement to consider only the continuity points of F is needed for a meaningful definition as the following simple example shows us: Let $a \in \mathbb{R}$, and let X_n, $n = 1, 2, \dots,$ be a random variable with distribution function

$$F_{X_n}(x) = \mathbb{P}(X_n \le x) = \begin{cases} 0 & \text{if } x < a - \frac{1}{n} \\ \frac{1}{2} & \text{if } a - \frac{1}{n} \le x < a + \frac{1}{n} \\ 1 & \text{if } x \ge a + \frac{1}{n}. \end{cases}$$

Now we see that

$$F_{X_n}(x) \xrightarrow[n\to\infty]{} \begin{cases} 0 & \text{if } x < a \\ \frac{1}{2} & \text{if } x = a \\ 1 & \text{if } x > a. \end{cases}$$

But since $X_n = a \pm \frac{1}{n}$ a.s., we would like to have $X_n \xrightarrow[n\to\infty]{d} a$, where $X \equiv a$ has the distribution function

$$F_X(x) = \begin{cases} 0 & \text{if } x < a \\ 1 & \text{if } x \ge a. \end{cases}$$

The convergence of $F_{X_n}(x)$ towards $F_X(x)$ indeed holds if and only if x is a continuity point of F_X.

Example 134. The requirement of F being a distribution function is also meaningful. Consider eg. the random variables $X_n = n$ a.s. Now for any $x \in \mathbb{R}$, we easily see that $\mathbb{P}(X_n \le x) \xrightarrow[n\to\infty]{} 0$, where $F \equiv 0$ is everywhere continuous but not a distribution function. This is natural, since $X_n \xrightarrow[n\to\infty]{a.s.} \infty$, so we also expect this to hold in the weak sense.

Indeed, convergence in distribution is weaker than convergence in probability (and thus also weaker than a.s. convergence):

Proposition 135. *Let $X_1, X_2, \dots$ and X be real-valued random variables on the same probability space. Then $X_n \xrightarrow[n\to\infty]{\mathbb{P}} X$ implies $X_n \xrightarrow[n\to\infty]{d} X$.*

Proof. Let $\epsilon > 0$ and $x \in \mathbb{R}$. We note that

$$\begin{aligned} F_{X_n}(x) = \mathbb{P}(X_n \le x) &= \mathbb{P}(X_n \le x, |X_n - X| \le \epsilon) + \mathbb{P}(X_n \le x, |X_n - X| > \epsilon) \\ &\le \mathbb{P}(X_n \le x, X_n - \epsilon \le X \le X_n + \epsilon) + \mathbb{P}(|X_n - X| > \epsilon) \\ &\le \mathbb{P}(X \le x + \epsilon) + \mathbb{P}(|X_n - X| > \epsilon) \end{aligned}$$

and similarly

$$\begin{aligned} \mathbb{P}(X \le x - \epsilon) &\le \mathbb{P}(X \le x - \epsilon, |X_n - X| \le \epsilon) + \mathbb{P}(|X_n - X| > \epsilon) \\ &\le \mathbb{P}(X \le x - \epsilon, X_n \le X + \epsilon) + \mathbb{P}(|X_n - X| > \epsilon) \\ &\le \mathbb{P}(X_n \le x) + \mathbb{P}(|X_n - X| > \epsilon), \end{aligned}$$

giving

$$F_{X_n}(x) = \mathbb{P}(X_n \le x) \ge \mathbb{P}(X \le x - \epsilon) - \mathbb{P}(|X_n - X| > \epsilon).$$

Since $X_n \xrightarrow[n\to\infty]{\mathbb{P}} X$, we have $\mathbb{P}(|X_n - X| > \epsilon) \xrightarrow[n\to\infty]{} 0$. Therefore,

$$F_X(x-\epsilon) = \mathbb{P}(X \le x - \epsilon) \le \liminf_{n\to\infty} F_{X_n}(x) \le \limsup_{n\to\infty} F_{X_n}(x) \le \mathbb{P}(X \le x + \epsilon) = F_X(x+\epsilon).$$

Letting $\epsilon \to 0$ gives then $F_X(x-) \le \liminf_{n\to\infty} F_{X_n}(x) \le \limsup_{n\to\infty} F_{X_n}(x) \le F_X(x)$, where we used the fact that F_X is càdlàg. Finally, assuming x is a continuity point of F_X, we obtain $F(x-) = F(x)$, giving $\lim_{n\to\infty} F_{X_n}(x) = F(x)$, and finally yielding $X_n \xrightarrow[n\to\infty]{d} X$. □

In fact, the implication of the previous proposition can be reverted in the special case of random variables with a constant limit. The following result is a special case of so-called Slutsky's theorem. The proof is left as an exercise to the reader.

Proposition 136. *(i) Let $X_1, X_2, \dots$ be real-valued random variables and $a \in \mathbb{R}$ a constant such that $X_n \xrightarrow[n\to\infty]{d} a$. Then for all $\epsilon > 0$, $\mathbb{P}(|X_n - a| > \epsilon) \xrightarrow[n\to\infty]{} 0$, where the probability measure $\mathbb{P} := \mathbb{P}^{(n)}$ may depend on n. In particular, if $X_1, X_2, \dots$ are defined on the same probability space, we have $X_n \xrightarrow[n\to\infty]{\mathbb{P}} a$.*

(ii) Let $X_1, X_2, \dots$ and X, as well as $Y_1, Y_2, \dots$, be real-valued random variables such that $X_n \xrightarrow[n\to\infty]{d} X$ and $Y_n \xrightarrow[n\to\infty]{d} 0$. Then $X_n + Y_n \xrightarrow[n\to\infty]{d} X$.

Next, our aim is to generalize the convergence in distribution to random variables which are not necessarily real-valued, in which case they do not have a distribution function. We also wish to obtain various equivalent characterizations for convergence in distribution. For that purpose, we define what is meant by laws of random variables to converge in a weak sense.

Definition 137 (Weak convergence of measures)**.** Let $(\mathcal{S}, d)$ be a metric space equipped with the Borel σ-algebra $\mathcal{B}(\mathcal{S})$. Let $\mu_1, \mu_2, \dots$ and μ be probability measures on $(\mathcal{S}, \mathcal{B}(\mathcal{S}))$. Then $\mu_n \xrightarrow[n\to\infty]{} \mu$ weakly if for all bounded and continuous functions $f : \mathcal{S} \to \mathbb{R}$,

$$\int_{\mathcal{S}} f d\mu_n \xrightarrow[n\to\infty]{} \int_{\mathcal{S}} f d\mu.$$

Example 138. Let $\mathcal{S} = \mathbb{R}$ (with its usual euclidean metric), and let $x_n \in \mathbb{R}$ and $x \in \mathbb{R}$ be points such that $x_n \xrightarrow[n\to\infty]{} x$. Consider the Dirac measures δ_{x_n} (recall Example 10). Let $f : \mathbb{R} \to \mathbb{R}$ be bounded and continuous. Then

$$\int_{\mathbb{R}} f d\delta_{x_n} = f(x_n) \xrightarrow[n\to\infty]{} f(x) = \int_{\mathbb{R}} f d\delta_x$$

by the continuity of f. This shows by definition that $\delta_{x_n} \xrightarrow[n\to\infty]{} \delta_x$ weakly.

Example 139. Consider the same setting as in the previous example, and for $n = 1, 2, \dots$, let $\mu_n := \frac{1}{n}\sum_{i=0}^{n-1} \delta_{\frac{i}{n}}$. That is, μ_n is the uniform measure on $\left\{0, \frac{1}{n}, \dots, \frac{n-1}{n}\right\}$. Let $f : \mathbb{R} \to \mathbb{R}$ be bounded and continuous. Then using linearity,

$$\int_{\mathbb{R}} f d\mu_n = \frac{1}{n}\sum_{i=0}^{n-1} \int_{\mathbb{R}} f d\delta_{\frac{i}{n}} = \frac{1}{n}\sum_{i=0}^{n-1} f\left(\frac{i}{n}\right) \xrightarrow[n\to\infty]{} \int_0^1 f(x)dx.$$

Therefore, $\mu_n \xrightarrow[n\to\infty]{} \mu$ weakly, where μ is the Lebesgue measure on $[0,1]$, i.e. the uniform distribution. Note that there exists an event A such that $\mu_n(A)$ does not converge towards $\mu(A)$. For example, take $A = [0,1] \cap \mathbb{Q}$, in which case $\mu_n(A) = 1$ for all n but $\mu(A) = 0$.

Let us observe easily that in the case of $\mathcal{S} = \mathbb{R}$, Definition 137 implies Definition 131.

Proposition 140. *Let $\mu_1, \mu_2, \dots$ and μ be probability measures on $\mathbb{R}$. Define functions F_n and F by setting $F_n(x) := \mu_n((-\infty, x])$ and $F(x) := \mu((-\infty, x])$. Then $\mu_n \xrightarrow[n\to\infty]{} \mu$ weakly implies $F_n \xrightarrow[n\to\infty]{} F$ weakly.*

Proof. Assume $\mu_n \xrightarrow[n\to\infty]{} \mu$ weakly. Recall that by the proof of Proposition 34, the functions F_n and F are càdlàg. Let $x < y$ and $f : \mathbb{R} \to \mathbb{R}$ the piecewise affine function

$$f(t) = \begin{cases} 1 & \text{if } t \le x \\ \frac{t-y}{x-y} & \text{if } x < t \le y \\ 0 & \text{if } t > y. \end{cases}$$

Note that f is bounded and continuous, and $\mathbb{1}_{(-\infty,x]} \le f \le \mathbb{1}_{(-\infty,y]}$. Thus we have

$$\begin{aligned} F_n(x) = \mu_n((-\infty, x]) = \int_{\mathbb{R}} \mathbb{1}_{(-\infty,x]} d\mu_n(t) \le \int_{\mathbb{R}} f(t) d\mu_n(t) \\ \xrightarrow[n\to\infty]{} \int_{\mathbb{R}} f(t) d\mu(t) \le \int_{\mathbb{R}} \mathbb{1}_{(-\infty,y]} d\mu(t) = F(y). \end{aligned}$$

Therefore, $\limsup_{n\to\infty} F_n(x) \le F(y)$ for all $y > x$, and letting $y \searrow x$ yields

$$\limsup_{n\to\infty} F_n(x) \le F(x)$$

by the right-continuity of F.

With a similar argument (by changing the roles of x and y), we can also show that $\liminf_{n\to\infty} F_n(y) \ge F(x)$ for all $y > x$, giving $\liminf_{n\to\infty} F_n(y) \ge F(y-)$ as $x \nearrow y$. If y is a continuity point of F, we conclude $\liminf_{n\to\infty} F_n(y) \ge F(y)$.

Putting the two together, for all x where F is continuous, we have shown

$$\limsup_{n\to\infty} F_n(x) \leq F(x) \leq \liminf_{n\to\infty} F_n(x),$$

yielding that $F_n \xrightarrow[n\to\infty]{} F$ weakly. □

In fact, changing our viewpoint from random variables to measures reveals quite a surprisingly close connection between convergence in distribution and almost sure convergence. We state and prove the following result for probability measures on $\mathbb{R}$, but it also holds for metric spaces (with a small extra assumption that the support of the limit measure is separable).

Theorem 141 (Skorokhod's representation)**.** *Let μ_n, $n = 1, 2, \ldots$ and μ be probability measures on $\mathbb{R}$ with associated distribution functions F_n and F, respectively. Assume $F_n \xrightarrow[n\to\infty]{} F$ weakly. Then there exists a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and random variables X_n, $n = 1, 2, \ldots$ and X on $(\Omega, \mathcal{F}, \mathbb{P})$ such that X_n has law μ_n, X has law μ and $X_n \xrightarrow[n\to\infty]{a.s.} X$.*

Proof. We build on the idea of Theorem 36. Let U be a uniform random variable on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where $\Omega = [0, 1]$, $\mathcal{F} = \mathcal{B}$ and $\mathbb{P}$ is the Lebesgue measure. Let F_n and F be the distribution functions defined as in the previous proposition. Then recall from the proof of Theorem 36 that $X_n := F_n^{-1}(U)$ and $X := F^{-1}(U)$ are random variables with laws μ_n and μ, respectively, where $F^{-1}(u) := \sup\{y \in \mathbb{R} : F(y) < u\}$ for all $u \in [0, 1]$ is the "càdlàg inverse" of F. Recall that it satisfies $u \leq F(x)$ if an only if $F^{-1}(u) \leq x$ for all $u \in [0, 1]$. Now it is enough to show $F_n^{-1}(U) \xrightarrow[n\to\infty]{a.s.} F^{-1}(U)$.

By the assumption, $F_n(x) \xrightarrow[n\to\infty]{} F(x)$ for all x such that F is continuous at x. Moreover, since F is increasing, its set of discontinuity points is countable by a general theorem from measure theory. Based on these observation, the rest of the proof consists of rather simple limit arguments as follows. Let $u \in (0, 1)$ arbitrary, and let $\epsilon > 0$. Let x be a continuity point of F such that $F^{-1}(u) - \epsilon < x < F^{-1}(u)$. Now $F(x) < u$, and since $F_n(x) \xrightarrow[n\to\infty]{} F(x)$, also $F_n(x) < u$ for n large enough. Hence, $F^{-1}(u) - \epsilon < x < F_n^{-1}(u)$ for n large, showing that $\liminf_{n\to\infty} F_n^{-1}(u) \geq F^{-1}(u)$.

For the other direction, observe that F^{-1} also has a countable set of discontinuity points (since it is again monotone). Now, suppose u is a continuity point of F^{-1}, let $u' > u$ and $y \in \mathbb{R}$ a continuity point of F such that $F^{-1}(u') < y < F^{-1}(u') + \epsilon$. Then $u < u' \leq F(F^{-1}(u')) \leq F(y)$, since F is increasing. Again, since $F_n(y) \xrightarrow[n\to\infty]{} F(y)$, there exists n large enough such that $u < F_n(y)$, and hence $F_n^{-1}(u) \leq y < F^{-1}(u') + \epsilon$. It implies that $\limsup_{n\to\infty} F_n^{-1}(u) \leq F^{-1}(u')$ for all $u' > u$. Finally, since u is a continuity point of F^{-1}, we can pass $u' \searrow u$ to obtain $\limsup_{n\to\infty} F_n^{-1}(u) \leq F^{-1}(u)$. Thus, it follows $\lim_{n\to\infty} F_n^{-1}(u) = F^{-1}(u)$ for all continuity points $u \in (0, 1)$. By the fact that there are only countably many discontinuity points, the set of continuity points has measure one. Hence, $F_n^{-1}(U) \xrightarrow[n\to\infty]{a.s.} F^{-1}(U)$.

□

Skorokhod's representation theorem is very powerful for applications in probability theory, since it often tells us that weak convergence is enough for a number of properties related to distributions. For example, it often allows us to exchange limits and integrals in the case when random variables only converge in distribution, instead of the required almost sure convergence. The following corollary serves as an example.

Corollary 142 (Fatou's lemma for convergence in distribution)**.** *Let X_n, $n = 1, 2, \dots$ and X be non-negative random variables such that $X_n \xrightarrow[n\to\infty]{d} X$. Then $\mathbb{E}(X) \leq \liminf_{n\to\infty} \mathbb{E}(X_n)$.*

Proof. By Skorokhod's representation theorem, there exist random variables $\tilde{X}_n$ and $\tilde{X}$ on a common probability space and having the same law as X_n and X, respectively, such that $\tilde{X}_n \xrightarrow[n\to\infty]{a.s.} \tilde{X}$. The claim follows then from Fatou's lemma, since $\mathbb{E}(X_n) = \mathbb{E}(\tilde{X_n})$ and $\mathbb{E}(X) = \mathbb{E}(\tilde{X})$. □

We are now ready to show the equivalence of the two definitions of weak convergence.

Theorem 143. *Let $\mu_1, \mu_2, \dots$ and μ be probability measures on $\mathbb{R}$. Define functions F_n and F by setting $F_n(x) := \mu_n((-\infty, x])$ and $F(x) := \mu((-\infty, x])$. Then $\mu_n \xrightarrow[n\to\infty]{} \mu$ weakly if and only if $F_n \xrightarrow[n\to\infty]{} F$ weakly.*

Proof. The first implication in this theorem is shown in Proposition 140. Thus, let us assume $F_n \xrightarrow[n\to\infty]{} F$ weakly. Then by Skorokhod's representation theorem, there exists random variables X_n $(n = 1, 2, \dots)$ and X with laws μ_n and μ, respectively, such that $X_n \xrightarrow[n\to\infty]{a.s.} X$. Let $f : \mathbb{R} \to \mathbb{R}$ be a bounded and continuous function. By continuity, we then also have $f(X_n) \xrightarrow[n\to\infty]{a.s.} f(X)$. Then by the dominated convergence theorem,

$$\int_{\mathbb{R}} f(t)d\mu_n(t) = \mathbb{E}(f(X_n)) \xrightarrow[n\to\infty]{} \mathbb{E}(f(X)) = \int_{\mathbb{R}} f(t)d\mu(t),$$

which yields $\mu_n \xrightarrow[n\to\infty]{} \mu$ weakly by definition. □

As a corollary, we find an equivalent definition of convergence in distribution for a sequence of random variables.

Corollary 144 (Convergence in distribution, 2nd definition)**.** *Random variables $X_1, X_2, \dots$ converge* in distribution *to a random variable X if $\mathbb{E}(f(X_n)) \xrightarrow[n\to\infty]{} \mathbb{E}(f(X))$ for all bounded and continuous functions $f : \mathbb{R} \to \mathbb{R}$.*

5.4 Fundamental theorems of weak convergence

We start this subsection by stating and proving several equivalent characterizations of weak convergence. If A is a set of a metric space, we use the notation A° for the interior of A, $\bar{A}$ for the closure and $\partial A = \bar{A} \setminus A^o$ for the boundary of A. Recall that always $A^\circ \subset A \subset \bar{A} = A^\circ \cup \partial A$, where A° is open and $\bar{A}$ and ∂A are closed.

Theorem 145 (Portmanteau theorem)**.** *Let $\mu_1, \mu_2, \dots$ be probability measures on $\mathbb{R}$. Then the following conditions are equivalent.*

(i) $\mu_n \xrightarrow[n\to\infty]{} \mu$ *weakly.*

(ii) *For all* $G \subset \mathbb{R}$ *open,* $\liminf_{n\to\infty} \mu_n(G) \geq \mu(G)$.

(iii) *For all* $F \subset \mathbb{R}$ *closed,* $\limsup_{n\to\infty} \mu_n(F) \leq \mu(F)$.

(iv) *For all Borel sets* $B \in \mathcal{B}(\mathbb{R})$*, if* $\mu(\partial B) = 0$*, then* $\mu_n(B) \xrightarrow[n\to\infty]{} \mu(B)$.

Proof. Let us assume first that (i) holds. Let $G \subsetneq \mathbb{R}$ be an open set (the case $G = \mathbb{R}$ yields trivially the equality $1 = 1$). Recall from topology that the function $x \mapsto d(x, G^c)$ is continuous (in fact, even Lipschitz-continuous). Define $f_M(x) := \min\{Md(x, G^c), 1\}$, which is then continuous and bounded, and satisfies $f_M \leq \mathbb{1}_G$ and $f_M \nearrow \mathbb{1}_G$ as $M \to \infty$. First, by (i), we deduce

$$\int_{\mathbb{R}} f_M d\mu_n \xrightarrow[n\to\infty]{} \int_{\mathbb{R}} f_M d\mu.$$

On the other hand, $\int_{\mathbb{R}} f_M d\mu_n \leq \int_{\mathbb{R}} \mathbb{1}_G d\mu_n = \mu_n(G)$, so $\liminf_{n\to\infty} \mu_n(G) \geq \int_{\mathbb{R}} f_M d\mu$. Finally, by the monotone convergence theorem,

$$\int_{\mathbb{R}} f_M d\mu \xrightarrow[M\to\infty]{} \int_{\mathbb{R}} \mathbb{1}_G d\mu = \mu(G),$$

which yields $\liminf_{n\to\infty} \mu_n(G) \geq \mu(G)$.

Let us then show that (ii) and (iii) are equivalent. If we assume (ii) and let $F \subset \mathbb{R}$ be closed, then $G := F^c$ is open, and the assumption (ii) yields $\liminf_{n\to\infty} \mu_n(F^c) \geq \mu(F^c)$. This is equivalent to $\liminf_{n\to\infty}(1 - \mu_n(F)) \geq 1 - \mu(F)$, which in turn is equivalent to $1 - \limsup_{n\to\infty} \mu_n(F) \geq 1 - \mu(F)$. Hence the claim (iii) follows. The reverse implication $(iii) \longrightarrow (ii)$ is proven similarly.

Now, assume (iii), and therefore also (ii), to hold. Let $B \in \mathcal{B}(\mathbb{R})$ such that $\mu(\partial B) = 0$. By the inclusions $B^\circ \subset B \subset \bar{B} = B^\circ \cup \partial B$, we then see that $\mu(B^\circ) = \mu(B) = \mu(\bar{B})$. Using this together with the inclusions, we obtain

$$\liminf_{n\to\infty} \mu_n(B) \geq \liminf_{n\to\infty} \mu_n(B^\circ) \geq \mu(B^\circ) = \mu(B) = \mu(\bar{B}) \geq \limsup_{n\to\infty} \mu_n(\bar{B}) \geq \limsup_{n\to\infty} \mu_n(B).$$

Hence, $\mu(B) = \limsup_{n\to\infty} \mu_n(B) = \liminf_{n\to\infty} \mu_n(B)$.

Finally, let us assume (iv). Observe that for all $x \in \mathbb{R}$, we have $\partial(-\infty, x] = \{x\}$. Therefore, if x is a continuity point of $F(x) = \mu((-\infty, x])$, we have $F_n(x) = \mu_n((-\infty, x]) \xrightarrow[n\to\infty]{} \mu((-\infty, x]) = F(x)$, since in this case $\mu(\{x\}) = 0$. Then (i) follows from Theorem 143. □

Remark 146. The Portmanteau theorem also works for $\mu_1, \mu_2, \ldots$ defined on a metric space. In this case, the proof of the final implication $(iv) \implies (i)$ is more cumbersome, whereas all the other implications are proven similarly.

Remark 147. Intuitively, the Portmanteau theorem has the following interpretation. By (ii), open sets can only lose probability mass in the limit, whereas by (iii) tells us that closed sets can only gain probability mass in the limit. The interpretation of (iv) is that a set can only gain or lose mass trough its boundary, which cannot happen if the boundary is trivial in the sense that it has measure zero.

Example 148. Recall from Example 139 that for $n = 1, 2, \ldots$, the discrete uniform measures $\mu_n := \frac{1}{n}\sum_{i=0}^{n-1} \delta_{\frac{i}{n}}$ converge weakly towards the Lebesgue measure μ on $[0, 1]$. Recall that for the set $A := \mathbb{Q} \cap [0, 1]$, we have $\mu_n(A) = 1$ for all $n = 1, 2, \ldots$, but $\mu(A) = 0$. Indeed, in this case $\partial A = [0,1] \setminus \mathbb{Q}$, so $\mu(\partial(A)) = 1 \neq 0$. Hence, we cannot deduce convergence of $\mu_n(A)$ even though μ_n converges weakly. Observe also that A is neither an open nor a closed set.

We would like to understand better what it means to lose probability mass at the boundary of an event. In particular, one may raise the question whether a sequence of measures has a weakly converging subsequence. The general answer to this question turns out to be negative, but a slightly weaker statement holds.

Theorem 149 (Helly's selection principle)**.** *Let F_n, $n = 1, 2, \dots$ be distribution functions. Then there exists a subsequence $(n_k)_{k=1}^{\infty}$ and an increasing càdlàg function $F : \mathbb{R} \to [0,1]$ such that $F_{n_k}(x) \xrightarrow[k\to\infty]{} F(x)$ in every $x \in \mathbb{R}$ such that F is continuous at x.*

Proof. The first part of the proof is a diagonal argument, based on the countability of $\mathbb{Q}$ and on the fact that $[0,1]$ is a compact set. We write $\mathbb{Q} = \{q_n : n = 1, 2, \dots\}$. First, by compactness, there exists a subsequence $(n_k^{(1)})$ such that $F_{n_k^{(1)}}(q_1) \in [0,1]$ converges as $k \to \infty$. Denote the limit by $f(q_1)$. Iterating this, there exists a subsequence $(n_k^{(2)})$ of $(n_k^{(1)})$ and $f(q_2) \in [0,1]$ such that $\lim_{k\to\infty} F_{n_k^{(2)}}(q_2) = f(q_2)$. More generally, for any $j = 1, 2, \dots$, there exists a subsequence $(n_k^{(j+1)})$ of $(n_k^{(j)})$ such that $F_{n_k^{(j+1)}}(q_{j+1}) \xrightarrow[k\to\infty]{} f(q_{j+1})$. We choose $n_k := n_k^{(k)}$ for all $k = 1, 2, \dots$, in which case we have $F_{n_k}(q) \xrightarrow[k\to\infty]{} f(q)$ for all $q \in \mathbb{Q}$. This constructs a function $f : \mathbb{Q} \to [0,1]$, which is increasing since the distribution functions F_{n_k} are increasing.

Let us then extend f to whole real line by setting $F(x) := \inf\{f(q) : q \in \mathbb{Q}, q > x\}$. Then F is also increasing, since making x larger would decrease the size of the set on which the infimum is taken. For the claim on the limit function, it is now enough to show that F is right-continuous, since every right-continuous increasing function must also have left limits. So, let $x \in \mathbb{R}$ and $\epsilon > 0$. Then by the definition of F and the infimum, there exists a $q > x$ such that $F(x) \leq f(q) < F(x) + \epsilon$. Therefore for all $y \in (x, q)$, by monotonicity $F(x) - \epsilon < F(x) \leq F(y) \leq f(q) < F(x) + \epsilon$, which implies $|F(y) - F(x)| < \epsilon$. Therefore, $\lim_{y \searrow x} F(y) = F(x)$ by the definition of right-continuity.

It remains to show that $F_{n_k}(x) \xrightarrow[k\to\infty]{} F(x)$ in all the continuity points x of F. For that purpose, fix x such that F is continuous at x, and let $\epsilon > 0$. Then for any $y < x$ close enough to x, we have $F(x) - \epsilon < F(y) < F(x) + \epsilon$. By definition of F, we then find $q, r \in \mathbb{Q}$ such that $y < q < x < r$ and $f(r) < F(x) + \epsilon$. Then since $F_{n_k}(q) \xrightarrow[k\to\infty]{} f(q) \in (F(x) - \epsilon, F(x)]$ and $F_{n_k}(r) \xrightarrow[k\to\infty]{} f(r) \in [F(x), F(x) + \epsilon)$, we have $F(x) - \epsilon \leq F_{n_k}(q) \leq F_{n_k}(x) \leq F_{n_k}(r) \leq F(x) + \epsilon$ for all k large enough, which yield $|F(x) - F_{n_k}(x)| \leq \epsilon$. Hence, $F_{n_k}(x) \xrightarrow[k\to\infty]{} F(x)$. □

Example 150. Consider $F_n := \mathbb{1}_{[n,\infty)}$ for all $n = 1, 2, \dots$. It is easy to see that F_n are distribution functions with $F_n(x) \xrightarrow[n\to\infty]{} 0$ for all $x \in \mathbb{R}$. That is, F_n converges to a càdlàg function 0, which is not a distribution function. This is caused by the following phenomenon. If X_n is a random variable distributed as F_n, then $\mathbb{P}(X_n \geq n) = 1$ for all $n \geq 1$, so X_n gets arbitrary high values with high probability. Intuitively, this could be rephrased as "probability mass escaping in the infinity".

Remark 151. Helly's selection principle is a more general theorem in convex geometry, for which the above one for distribution functions is a special case. It was originally proven by Eduard Helly, who used to work at the University of Vienna.

5.4.1 Tightness

In this subsection, we aim at understanding systematically the "escaping mass" phenomenon of Example 150, as well as finding a criterion for the existence of weakly converging subsequences of probability measures. Observe that the proof of Helly's selection theorem applied compactness on the euclidean topology to find a converging subsequence of distribution functions. In order to extend this to the weak convergence of laws, we would need a compactness assumption in the space of measures (in a suitable sense). The concept of *tightness* addresses these problems.

Definition 152 (Tightness)**.** A sequence of probability measures $(\mu_n)_{n=1}^{\infty}$ defined on a metric space $\mathcal{S}$ is *tight* if for every $\epsilon > 0$, there exists a compact set $K \subset \mathcal{S}$ such that $\mu_n(K) \geq 1-\epsilon$ for all $n = 1, 2, \ldots$. When $\mathcal{S} = \mathbb{R}$ with the euclidean metric, equivalently $(\mu_n)_{n=1}^{\infty}$ is tight if for all $\epsilon > 0$, there exists $a, b \in \mathbb{R}$ such that $\mu_n([a, b]) \geq 1 - \epsilon$ for all $n = 1, 2, \ldots$.

Remark 153. We easily notice that tightness is, in fact, a property of unordered collections of measures and generalises as follows. A collection of probability measures $\{\mu_i : i \in I\}$, where I is an arbitrary index set, is tight if for every $\epsilon > 0$, there exists a compact set $K \subset \mathcal{S}$ such that $\mu_i(K) \geq 1 - \epsilon$ for all $i \in I$. We will use this general definition in Proposition 192 later.

Example 154. Consider the distribution functions $F_n := \mathbb{1}_{[n,\infty)}$ for all $n = 1, 2, \ldots$ from Example 150. Let μ_n denote the associated laws, and let X_n be a random variable distributed as μ_n. Then for all $a, b \in \mathbb{R}$, we have $\mu_n([a, b]) = \mathbb{P}(X_n \in [a, b]) \leq \mathbb{P}(X_n \leq b)$. We see that if $n > b$, $\mathbb{P}(X_n \leq b) = 0$. Hence, the sequence $(\mu_n)_{n=1}^{\infty}$ cannot be tight.

Theorem 155 (Prokhorov's theorem)**.** *Let $(\mu_n)_{n=1}^{\infty}$ be a sequence of probability measures on $(\mathbb{R}, \mathcal{B})$. Then the following are equivalent.*

(i) *$(\mu_n)_{n=1}^{\infty}$ is tight.*

(ii) *For all subsequences $(n_k)_{k=1}^{\infty}$, there exists a subsequence $(n_{k_l})_{l=1}^{\infty}$ and a probability measure μ on $(\mathbb{R}, \mathcal{B})$ such that $\mu_{n_{k_l}} \xrightarrow[l\to\infty]{} \mu$ weakly.*

Proof. We show first the implication $(i) \implies (ii)$. Thus, let $(\mu_n)_{n=1}^{\infty}$ be tight, and let (n_k) be an arbitrary subsequence. Then by Helly's selection principle, there exists a subsequence (n_{k_l}) and an increasing càdlàg function $F : \mathbb{R} \to [0, 1]$ such that $F_{n_{k_l}}(x) \xrightarrow[k\to\infty]{} F(x)$ in every continuity point $x \in \mathbb{R}$ of F, where $F_{n_{k_l}}$ are the associated distribution functions of $\mu_{n_{k_l}}$. It is enough to show that F is a distribution function, which then defines the desired measure μ by Theorem 36. It remains to show that $\lim_{x\to\infty} F(x) = 1$ and $\lim_{x\to-\infty} F(x) = 0$.

Therefore, let $\epsilon > 0$ and $a, b \in \mathbb{R}$ such that $\mu_n((a, b)) \geq 1 - \epsilon$, where the latter condition follows from tightness (by eg. considering an open enlargement of a closed interval provided by the definition). Without loss of generality, we may choose a, b such that they are continuity points of F. Then we have $F_{n_{k_l}}(a) = \mu_{n_{k_l}}((-\infty, a]) \leq \epsilon$ for all $l = 1, 2, \ldots$, so $F(a) = \lim_{l\to\infty} F_{n_{k_l}}(a) \leq \epsilon$. Since $\epsilon > 0$ was arbitrary and F is increasing, the first claim follows. Similarly, we deduce $F_{n_{k_l}}(b) = \mu_{n_{k_l}}((-\infty,b]) \geq \mu_{n_{k_l}}((a,b)) \geq 1-\epsilon$, so $F(b) = \lim_{l\to\infty} F_{n_{k_l}}(b) \geq 1-\epsilon$, giving the second limit.

Conversely, let us now prove the implication $(ii) \implies (i)$. Let us assume that (ii) holds and the sequence $(\mu_n)_{n=1}^\infty$ is not tight. Then there exists an $\epsilon > 0$ such that for all $a,b \in \mathbb{R}$, we have $\mu_n([a,b]) < 1-\epsilon$ for some n. Thus, let us choose n_k, $k = 1,2,\ldots,$ such that $\mu_{n_k}([-k,k]) < 1-\epsilon$. By assumption (ii), we then find a subsequence $(n_{k_l})_{l=1}^\infty$ of $(n_k)_{k=1}^\infty$ such that $\mu_{n_{k_l}} \xrightarrow[l\to\infty]{} \mu$ weakly, where μ is some probability measure on $\mathbb{R}$. If F is now the distribution function associated with μ, let a and b be continuity points of F. Then $\mu(\{a\}) = 0 = \mu(\{b\})$, and moreover we can choose a,b such that $\mu([a,b]) > 1-\epsilon$. Now by the Portmanteau theorem, $\mu_{n_{k_l}}([a,b]) \xrightarrow[l\to\infty]{} \mu([a,b]) > 1-\epsilon$. This implies that for l large enough, $1-\epsilon > \mu_{n_{k_l}}([-k_l,k_l]) \geq \mu_{n_{k_l}}([a,b]) > 1-\epsilon$, which is a contradiction. Hence, $(\mu_n)_{n=1}^\infty$ must be tight. □

Remark 156. The implication $(i) \implies (ii)$ of Prokhorov's theorem is typically more useful in practice. It also works on arbitrary metric spaces as such, while the other implication needs an assumption that the metric space is complete and separable.

As a corollary, we get the following very useful sufficient condition for weak convergence. It tells that under tightness, weak convergence along subsequences towards a unique limit implies the weak convergence of the full sequence of probability measures. In practice, one then often shows weak convergence by showing tightness of the sequence of measures separately with the uniqueness of any subsequential weak limits.

Corollary 157. *Let $(\mu_n)_{n=1}^\infty$ be a tight sequence of probability measures and μ a probability measure on $\mathbb{R}$. Assume that all weakly converging subsequences $(n_k)_{k=1}^\infty$ have the same limit measure μ. Then $\mu_n \xrightarrow[n\to\infty]{} \mu$ weakly.*

Proof. Assume that $\mu_n \xrightarrow[n\to\infty]{} \mu$ does not converge weakly. Then by (the proof of) the Portmanteau theorem, there exists an $x \in \mathbb{R}$ such that $\mu(\{x\}) = 0$ but $\mu_n((-\infty,x])$ does not converge to $\mu((-\infty,x])$. Thus, there exists $\epsilon > 0$ and a subsequence $(n_k)_{k=1}^\infty$ such that $|\mu_{n_k}((-\infty,x]) - \mu((-\infty,x])| \geq \epsilon$ for all $k = 1,2,\ldots$. But since $(\mu_n)_{n=1}^\infty$ is tight, there exists a subsequence $(n_{k_l})_{l=1}^\infty$ of $(n_k)_{k=1}^\infty$ such that $\mu_{n_{k_l}} \xrightarrow[l\to\infty]{} \mu$ weakly, which cannot happen. This contradiction shows that $\mu_n \xrightarrow[n\to\infty]{} \mu$ weakly. □

Chapter 6

Characteristic function

In this chapter, we introduce and study the *Fourier transform* of a random variable, which encodes various informations about the distribution and the convergence properties of a random variable. In probability theory, the Fourier transform is called the *characteristic function*. Besides characterizing probability distributions and being useful in proving convergence of probability distributions, it also allows us to compute moments of random variables. What distinguishes it from eg. the moment generating function is the fact that it is always well-defined, at the cost that we must allow it to obtain complex values.

Definition 158 (Characteristic function)**.** Let $X : \Omega \to \mathbb{R}$ be a random variable with law μ. The *characteristic function* of X is the function $\varphi_X : \mathbb{R} \to \mathbb{C}$ defined by

$$\varphi_X(t) := \mathbb{E}(e^{itX}) = \int_{\mathbb{R}} e^{itx} d\mu(x).$$

Remark 159. Recall from complex numbers that $e^{ix} = \cos(x) + i\sin(x)$. In particular, for any $x \in \mathbb{R}$, we have $|e^{ix}| = \sqrt{\cos^2(x) + \sin^2(x)} = 1$ (recall that $i^2 = -1$ and that $|x+iy|^2 = (x+iy)(\overline{x+iy}) = (x+iy)(x-iy) = x^2 + y^2$ for all $x, y \in \mathbb{R}$). Therefore, e^{itX} is always integrable, where we recall that the complex integral is defined as $\int_{\mathbb{R}} e^{itx} d\mu(x) = \int_{\mathbb{R}} \cos(tx) d\mu(x) + i \int_{\mathbb{R}} \sin(tx) d\mu(x)$. Hence, the characteristic function is well-defined for any real random variable.

Remark 160. Observe that the characteristic function φ_X only depends on the law of X (via Proposition 45), but not on the function X itself.

Example 161. Let $X \sim \mathrm{Exp}(\lambda)$ with $\lambda > 0$, that is, X has the density $f(x) = \lambda e^{-\lambda x} \mathbb{1}_{x>0}$. Then

$$\varphi_X(t) = \int_{\mathbb{R}} e^{itx} f(x) dx = \lambda \int_0^\infty e^{(it-\lambda)x} dx = \frac{\lambda}{it-\lambda} \Big(\lim_{x\to\infty} (e^{itx} e^{-\lambda x}) - e^{(it-\lambda)\cdot 0} \Big) = \frac{\lambda}{\lambda - it}.$$

Above, $\lim_{x\to\infty}(e^{itx}e^{-\lambda x}) = 0$ since $\left|e^{itx}e^{-\lambda x}\right| = e^{-\lambda x} \xrightarrow[x\to\infty]{} 0$. Note that the integral is in fact a complex integral over the half-infinite line segment $y = (-\lambda + it)x$ parametrized by $x \in (0, \infty)$. The reader may rest assured that the usual integration techniques on the real line remain valid in this case. The integration could always be made rigorous using complex contour integration techniques, which hold for more complicated complex integrals as well (and are in any case useful in order to estimate more complicated characteristic functions). The reader is encouraged to work on these examples using the definition of complex contour integration.

Example 162. Let $X \sim \text{Unif}([-1,1])$. Then U has the density $f(x) = \frac{1}{2}\mathbb{1}_{-1 \le x \le 1}$. Thus,

$$\varphi_X(t) = \int_{-1}^{1} \frac{1}{2} e^{itx} dx = \frac{1}{2it}\left(e^{it} - e^{-it}\right) = \frac{\sin t}{t}.$$

Observe that $\varphi_X(t) \in \mathbb{R}$. In fact, this holds for all *symmetric* distributions, i.e. for those for which X and $-X$ have the same distribution. This is proven as a part of the following proposition.

We readily obtain some general basic properties of the characteristic function.

Proposition 163 (Basic properties of the characteristic function)**.** *Let $\varphi \equiv \varphi_X$ be the characteristic function of some random variable X. Then the following properties hold.*

(i) $\varphi(0) = 1$ *and* $|\varphi(t)| \le 1$.

(ii) $\varphi(-t) = \overline{\varphi(t)}$. *In particular, if X and $-X$ have the same distribution, then $\varphi(t) \in \mathbb{R}$ for all $t \in \mathbb{R}$.*

(iii) *φ is uniformly continuous. That is,* $\sup_{t\in\mathbb{R}} |\varphi(t+\epsilon) - \varphi(t)| \longrightarrow 0$ *as* $\epsilon \to 0$.

(iv) *If $X_1, \ldots, X_n$ are independent random variables, then*

$$\varphi_{X_1+\cdots+X_n}(t) = \prod_{k=1}^{n} \varphi_{X_k}(t).$$

In particular, if $X_1, \ldots, X_n$ are i.i.d., then $\varphi_{X_1+\cdots+X_n}(t) = \varphi_{X_1}(t)^n$.

Proof. (i) Directly from the definition, we find $\varphi(0) = \mathbb{E}(e^{i\cdot 0}) = \mathbb{E}(1) = 1$ and $|\varphi(t)| \le \mathbb{E}(\left|e^{itX}\right|) = \mathbb{E}(1) = 1$.

(ii) We observe that for any complex random variable $Z = V + iW$, where V and W are real random variables, we have $\mathbb{E}(V+iW) = \mathbb{E}(V) + i\mathbb{E}(W)$ by linearity of the complex integral. Therefore the complex conjugate satisfies $\mathbb{E}(\overline{Z}) = \mathbb{E}(\overline{V+iW}) = \mathbb{E}(V - iW) = \mathbb{E}(V) - i\mathbb{E}(W) = \overline{\mathbb{E}(V) + i\mathbb{E}(W)} = \overline{\mathbb{E}(V+iW)} = \overline{\mathbb{E}(Z)}$. Hence,

$$\begin{aligned}\varphi(-t) &= \mathbb{E}(e^{-itX}) = \mathbb{E}(\cos(tX) - i\sin(tX)) = \mathbb{E}(\overline{\cos(tX) + i\sin(tX)}) \\ &= \overline{\mathbb{E}(\cos(tX) + i\sin(tX))} = \overline{\mathbb{E}(e^{itX})} = \overline{\varphi(t)}.\end{aligned}$$

If X and $-X$ have the same distribution, then $\varphi_X(t) = \mathbb{E}(e^{itX}) = \mathbb{E}(e^{it(-X)}) = \mathbb{E}(e^{i(-t)X}) = \varphi_X(-t) = \overline{\varphi_X(t)}$. That is, $\varphi(t) \in \mathbb{R}$.

(iii) If $\epsilon \in \mathbb{R}$ and $t \in \mathbb{R}$, we have $|\varphi(t+\epsilon) - \varphi(t)| \le \mathbb{E}(\left|e^{itX}\right|\left|e^{i\epsilon X} - 1\right|) = \mathbb{E}(\left|e^{i\epsilon X} - 1\right|)$, where $\left|e^{i\epsilon X} - 1\right| \le 2$. Hence by dominated convergence,

$$\sup_{t\in\mathbb{R}} |\varphi(t+\epsilon) - \varphi(t)| \le \mathbb{E}(\left|e^{i\epsilon X} - 1\right|) \longrightarrow 0$$

as $\epsilon \to 0$.

(iv) If $Y = R + iS$ and $Z = V + iW$ are two complex random variables such that (R,S) and (V,W) are independent, then using the linearity of expectation we find $\mathbb{E}(YZ) = \mathbb{E}((R+iS)(V+iW)) = \mathbb{E}(R+iS)\mathbb{E}(V+iW) = \mathbb{E}(Y)\mathbb{E}(Z)$. Applying this to $Y = e^{itX_1}$ and $Z = e^{itX_2}$ gives $\varphi_{X_1+X_2}(t) = \mathbb{E}(e^{itX_1}e^{itX_2}) = \mathbb{E}(e^{itX_1})\mathbb{E}(e^{itX_2}) = \varphi_{X_1}(t)\varphi_{X_2}(t)$. The general claims follows then by induction.

□

Above we already got some idea how the characteristic function carries information on the distribution of a random variable. We will next find much more refined results on this. One important question is to compute or estimate moments of random variables. It turns out that the characteristic function works very well in this. A more basic object to study would be the *moment generating function* a.k.a. the *Laplace transform* $t \mapsto \mathbb{E}(e^{tX})$, which is just a real integral of the random variable e^{tX}. If the moment generating function is well-defined, we would find $\mathbb{E}(X^n) = \frac{d^n}{dt^n}\mathbb{E}(e^{tX})_{|t=0}$ in an efficient way, which is also often easier to analyse computationally. However, the MGF is not always well-defined, which poses a great restriction to its use. We will see in the next theorem that the MGF can always be replaced by the characteristic function, with the cost that its derivatives will be complex valued.

Proposition 164. *Let X be a real random variable and assume $\mathbb{E}(|X|^n) < \infty$ for some $n \in \mathbb{N}$. Then the characteristic function $\varphi \equiv \varphi_X$ is n times differentiable and $\varphi^{(n)}(t) = \mathbb{E}\left(\frac{d^n}{dt^n} e^{itX}\right) = \mathbb{E}\left((iX)^n e^{itX}\right)$. In particular, $\varphi^{(n)}(0) = i^n \mathbb{E}(X^n)$. Moreover, if all the moments are finite (i.e. $\mathbb{E}(|X|^n) < \infty$ for all $n \in \mathbb{N}$) and $t \in \mathbb{R}$ is such that $\frac{t^n \mathbb{E}(|X|^n)}{n!} \xrightarrow[n\to\infty]{} 0$, then*

$$\varphi(t) = \sum_{n=0}^{\infty} \frac{(it)^n \mathbb{E}(X^n)}{n!}.$$

For the proof, we apply the following bound of the tail sum of the complex exponential, which is proven in [2, Lemma 3.3.19].

Lemma 165. *For all $x \in \mathbb{R}$ and $n \in \mathbb{N}$,*

$$\left| e^{ix} - \sum_{m=0}^{n} \frac{(ix)^m}{m!} \right| \le \min\left\{ \frac{|x|^{n+1}}{(n+1)!}, \frac{2|x|^n}{n!} \right\}.$$

Proof of Proposition 164. The case $n = 0$ is just the definition of the characteristic function. Thus, we assume $\mathbb{E}(|X|) < \infty$ $(n = 1)$, and conclude the general case by induction. Using the linearity of expectation, the difference quotient of φ satisfies $\frac{\varphi(t+\epsilon)-\varphi(t)}{\epsilon} = \mathbb{E}\left(\frac{e^{i(t+\epsilon)X} - e^{itX}}{\epsilon}\right) = \mathbb{E}\left(e^{itX} \frac{e^{i\epsilon X}-1}{\epsilon}\right)$. Therefore by the triangle inequality of expectation and Lemma 165,

$$\begin{aligned}\left| \frac{\varphi(t+\epsilon)-\varphi(t)}{\epsilon} - \mathbb{E}\left(iXe^{itX}\right) \right| = \left| \mathbb{E}\left(e^{itX} \frac{e^{i\epsilon X} - 1 - i\epsilon X}{\epsilon} \right) \right| &\le \mathbb{E}\left(\left| \frac{e^{i\epsilon X} - 1 - i\epsilon X}{\epsilon} \right| \right) \\ &\le \mathbb{E}\left(\min\left\{ \frac{\epsilon X^2}{2}, 2|X| \right\} \right).\end{aligned}$$

We note that $\mathbb{E}(|X|) < \infty$ by assumption, and $\frac{\epsilon X^2}{2} \longrightarrow 0$ as $\epsilon \to 0$. Hence the dominated convergence theorem implies $\left| \frac{\varphi(t+\epsilon)-\varphi(t)}{\epsilon} - \mathbb{E}\left(iXe^{itX}\right) \right| \longrightarrow 0$ as $\epsilon \to 0$.

Assume now that $n \in \mathbb{N}$ is such that $\mathbb{E}(|X|^{n+1}) < \infty$. Then repeating the above argument replacing φ by $\varphi^{(n)}$ and $\mathbb{E}\left(iXe^{itX}\right)$ by $\mathbb{E}\left((iX)^{n+1}e^{itX}\right)$ yields the claim. Finally, assume that all the moments of X exist and $t \in \mathbb{R}$ is such that $\frac{t^n \mathbb{E}(|X|^n)}{n!} \xrightarrow[n\to\infty]{} 0$. We may assume $t \ge 0$, otherwise we consider the complex conjugates via the identities $\varphi(-t) = \overline{\varphi(t)}$ and $-it = \overline{it}$. Then

$$\begin{aligned}\left|\varphi(t) - \sum_{k=0}^{n} \frac{(it)^k \mathbb{E}(X^k)}{k!}\right| = \left|\mathbb{E}\left(e^{itX} - \sum_{k=0}^{n} \frac{(it)^k X^k}{k!}\right)\right| &\leq \mathbb{E}\left(\left|e^{itX} - \sum_{k=0}^{n} \frac{(it)^k X^k}{k!}\right|\right) \\ &\leq \mathbb{E}\left(\min\left\{\frac{t^{n+1}|X|^{n+1}}{(n+1)!}, \frac{2t^n |X|^n}{n!}\right\}\right).\end{aligned}$$

By assumption, the quantity on the right hand side converges to zero as $n \to \infty$, yielding $\varphi(t) = \sum_{n=0}^{\infty} \frac{(it)^n \mathbb{E}(X^n)}{n!}$. □

The following theorem is an important result from Fourier analysis, which holds in the special case of continuous probability distributions, i.e. if a density function exists. In this case, the tail of the characteristic function converges to zero.

Theorem 166 (Riemann-Lebesgue lemma)**.** *Let μ be a probability measure on $\mathbb{R}$ with density f, and φ the associated characteristic function. Then $\varphi(t) \xrightarrow[t\to\infty]{} 0$.*

Proof. By definition, $\varphi(t) = \int_{\mathbb{R}} e^{itx} d\mu(x) = \int_{\mathbb{R}} e^{itx} f(x) dx$. Assume first $f = \mathbb{1}_{(a,b)}$ for some $a < b$. Then $\varphi(t) = \int_a^b e^{itx} dx = \frac{1}{it}(e^{itb} - e^{ita}) \xrightarrow[t\to\infty]{} 0$ since $\left|e^{itb} - e^{ita}\right| \leq 2$. Then, if f is a step function, it is a linear combination of indicator functions of intervals, and the claim follows from the linearity of expectation. Finally, assume $f \geq 0$ is any density function, and let $\epsilon > 0$. By the fact that step functions are dense in the L^1-space of integrable functions, there exists a step function g such that $\int_{\mathbb{R}} |f(x) - g(x)| dx \leq \frac{\epsilon}{2}$, which also satisfies $\left|\int_{\mathbb{R}} e^{itx} g(x) dx\right| \leq \frac{\epsilon}{2}$ for t large enough by the previous observations. Then we have

$$\left|\int_{\mathbb{R}} e^{itx} f(x) dx - \int_{\mathbb{R}} e^{itx} g(x)\right| \leq \int_{\mathbb{R}} \left|e^{itx}\right| |f(x) - g(x)| dx = \int_{\mathbb{R}} |f(x) - g(x)| dx \leq \frac{\epsilon}{2}.$$

Hence,

$$|\varphi(t)| = \left|\int_{\mathbb{R}} e^{itx} f(x) dx\right| \leq \int_{\mathbb{R}} |f(x) - g(x)| dx + \left|\int_{\mathbb{R}} e^{itx} g(x) dx\right| \leq \epsilon$$

for t large enough, showing the claim. □

6.1 Fourier inversion

In this section, our goal is to explain why the characteristic function indeed is called *characteristic*. We will see that given a probability measure μ on the real line, its characteristic function in fact gives an explicit expression for the measure μ, showing the uniqueness of the law assigned to the characteristic function. This result is a special case of the *Fourier inversion* in Fourier analysis.

Theorem 167 (Fourier inversion theorem)**.** *Let μ be a probability measure on $\mathbb{R}$, and let $\varphi(t) = \int_{\mathbb{R}} e^{itx} d\mu(x)$ be its characteristic function. Let $a < b$. Then*

$$\lim_{T\to\infty} \frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it} \varphi(t) dt = \mu((a,b)) + \frac{1}{2}(\mu(a) + \mu(b)).$$

In particular, if μ_1 and μ_2 have the same characteristic function, then $\mu_1 = \mu_2$.

Proof. From the definition of the characteristic function φ, we have

$$I_T := \frac{1}{2\pi}\int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it}\varphi(t)dt = \frac{1}{2\pi}\int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it}\int_{\mathbb{R}} e^{itx}d\mu(x)dt.$$

We can rewrite the integrand of this double integral as $\frac{e^{-ita}-e^{-itb}}{it}e^{itx} = \int_a^b e^{-it(y-x)}dy$, which gives $\left|\frac{e^{-ita}-e^{-itb}}{it}e^{itx}\right| \leq \int_a^b \left|e^{it(x-y)}\right| dy = b - a$. This constant upper bound is integrable over the product measure $d\mu \otimes \mathbb{1}_{[-T,T]}dt$, and hence Fubini's theorem may be applied. Thus,

$$I_T = \frac{1}{2\pi}\int_{\mathbb{R}}\int_{-T}^{T} \frac{e^{it(x-a)} - e^{it(x-b)}}{it}dtd\mu(x).$$

Here, the inner integral can be simplified as

$$\begin{aligned}\int_{-T}^{T} \frac{e^{it(x-a)} - e^{it(x-b)}}{it}dt &= \int_{0}^{T} \frac{e^{it(x-a)} - e^{it(x-b)}}{it}dt + \int_{0}^{T} \frac{e^{-it(x-a)} - e^{-it(x-b)}}{-it}dt \\ &= 2\int_0^T \frac{\sin(t(x-a)) - \sin(t(x-b))}{t}dt\end{aligned}$$

where we used the identities $\sin(t(x-a)) = \frac{1}{2i}(e^{it(x-a)} - e^{-it(x-a)})$ and $\sin(t(x-b)) = \frac{1}{2i}(e^{it(x-b)} - e^{-it(x-b)})$. Making the change of variable $u = t(x-a)$ and $v = t(x-b)$ allows us to rewrite the above integral as

$$\begin{aligned}\int_0^T \frac{\sin(t(x-a)) - \sin(t(x-b))}{t}dt &= \operatorname{sgn}(x-a)\int_0^{T|x-a|}\frac{\sin(u)}{u}du - \operatorname{sgn}(x-b)\int_0^{T|x-b|}\frac{\sin(v)}{v}dv \\ &=: \Psi_T(x).\end{aligned}$$

where $\operatorname{sgn}(y) = \begin{cases} -1 & \text{if } y < 0, \\ 0 & \text{if } y = 0, \\ 1 & \text{if } y > 0 \end{cases}$ is the sign function. When taking $T \to \infty$, we detect the remarkable (and quite miraculous) identity $\int_0^\infty \frac{\sin(u)}{u}du = \frac{\pi}{2}$. We leave its proof as an exercise to the reader (one way is to use and show [2, Exercise 1.7.5]). To conclude, it thus remains to take into account the signs in the expression of $\Psi_T(x)$ and carry out the limit $T \to \infty$. We find

$$\Psi(x) := \lim_{T\to\infty}\Psi_T(x) = \begin{cases} 0 & (x < a) \\ \frac{\pi}{2} & (x = a) \\ \pi & (a < x < b) \\ \frac{\pi}{2} & (x = b) \\ 0 & (x > b). \end{cases}$$

Since $|\Psi(x)| \leq \pi$, we have $|\Psi_T(x)| \leq 2\pi < \infty$ for all T large enough. Therefore by the dominated convergence theorem, we conclude

$$\lim_{T\to\infty} I_T = \lim_{T\to\infty}\frac{1}{\pi}\int_{\mathbb{R}}\Psi_T(x)d\mu(x) = \frac{1}{\pi}\int_{\mathbb{R}}\Psi(x)d\mu(x) = \mu((a,b)) + \frac{1}{2}(\mu(a) + \mu(b)).$$

□

Remark 168. The integral $\int_{-\infty}^{\infty} \frac{e^{-ita}-e^{-itb}}{it}\varphi(t)dt$ does not necessarily exist, even though $\lim_{T\to\infty} \int_{-T}^{T} \frac{e^{-ita}-e^{-itb}}{it}\varphi(t)dt$ does. For example, we may take $\mu = \delta_0$ (i.e. $X \equiv 0$ a.s.) in which case $\varphi(t) = \mathbb{E}(e^{it\cdot 0}) = 1$ for all $t \in \mathbb{R}$. The integral is then not well-defined. Observe that μ does not have a density, since otherwise by Riemann-Lebesgue lemma, we would have $\varphi(t) \xrightarrow[t\to\infty]{} 0$. In fact, integrable characteristic functions always give rise to a density, as the following corollary points out.

Corollary 169. *Let X is a random variable with characteristic function φ. Assume that the characteristic function is integrable, i.e. $\int_{\mathbb{R}} |\varphi(t)|dt < \infty$. Then X has a density f given by $f(x) = \frac{1}{2\pi}\int_{\mathbb{R}} e^{-itx}\varphi(t)dt$.*

Proof. Left as an exercise to the reader. □

Corollary 170. *Let X be a real random variable. Then X and $-X$ have the same law if and only if $\varphi_X(t) \in \mathbb{R}$ for all $t \in \mathbb{R}$.*

Proof. The implication from the first claim to the second claim is proven in Proposition 163 (ii). Thus, let us just show the converse implication. Assume $\varphi_X(t) \in \mathbb{R}$ for all $t \in \mathbb{R}$. Then by using again 163 (ii), we find $\varphi_X(t) = \overline{\varphi_X(t)} = \varphi_X(-t) = \varphi_{-X}(t)$. That is, the characteristic functions of X and $-X$ are the same, and hence by Fourier inversion theorem, X and $-X$ have the same law. □

6.2 Weak convergence and characteristic functions

This section provides a very important characterization of weak convergence via pointwise convergence of characteristic functions. This is, in fact, one of the most practical ways to show weak convergence of probability measures.

Theorem 171 (Lévy's continuity theorem)**.** *Let μ_n, $n = 1, 2, \dots,$ and μ be probability measures with characteristic functions φ_n and φ, respectively. Then the following are equivalent.*

(i) $\mu_n \xrightarrow[n\to\infty]{} \mu$ weakly.

(ii) $\varphi_n(t) \xrightarrow[n\to\infty]{} \varphi(t)$ for all $t \in \mathbb{R}$ (pointwise).

Proof. We assume first (i), i.e. $\mu_n \xrightarrow[n\to\infty]{} \mu$ weakly. Note that the function $t \mapsto e^{itx} = \cos(tx) + i\sin(tx)$ is bounded and continuous, and so are its real and imaginary parts. Thus,

$$\varphi_n(t) = \int_{\mathbb{R}} e^{itx} d\mu_n(x) = \int_{\mathbb{R}} \cos(tx)d\mu_n(x) + i\int_{\mathbb{R}} \sin(tx)d\mu_n(x) \xrightarrow[n\to\infty]{} \int_{\mathbb{R}} e^{itx}d\mu(x) = \varphi(t)$$

for all $t \in \mathbb{R}$ by the definition of the weak convergence.

Assume now that (ii) holds. We show first that the sequence $(\mu_n)_{n=1}^{\infty}$ is tight (so that we could apply Prokhorov's theorem). In order to find a good estimate of μ_n for tightness, we fix $u > 0$ and consider the integral

$$I_n(u) := \frac{1}{u}\int_{-u}^{u} (1 - \varphi_n(t))dt = \frac{1}{u}\int_{-u}^{u} \left(1 - \int_{\mathbb{R}} e^{itx}d\mu_n(x)\right) dt = \frac{1}{u}\int_{-u}^{u}\int_{\mathbb{R}} (1 - e^{itx})d\mu_n(x)dt.$$

We notice that the function $t \mapsto 1 - e^{itx}$ is bounded and the measure $d\mu_n(x) \otimes \mathbb{1}_{[-u,u]}dt$ is a finite measure, hence we may apply Fubini's theorem to write

$$\begin{aligned} I_n(u) = \int_{\mathbb{R}} \frac{1}{u} \int_{-u}^{u} (1 - e^{itx}) dt d\mu_n(x) &= \int_{\mathbb{R}} \frac{1}{u} \left(2u - \frac{1}{ix}(e^{ixu} - e^{-ixu}) \right) d\mu_n(x) \\ &= 2 \int_{\mathbb{R}} \left(1 - \frac{\sin(xu)}{xu} \right) d\mu_n(x). \end{aligned}$$

Since $\left| \frac{\sin(xu)}{xu} \right| \leq 1$, the integrand is positive, and thus we may estimate

$$I_n(u) \geq 2 \int_{|x| \geq \frac{2}{u}} \left(1 - \frac{\sin(xu)}{xu} \right) d\mu_n(x) \geq \int_{|x| \geq \frac{2}{u}} 2 \left(1 - \frac{1}{2} \right) d\mu_n(x) = \mu_n \left(\left\{ x : |x| \geq \frac{2}{u} \right\} \right).$$

Consider now the compact interval $K_u = \left[-\frac{2}{u}, \frac{2}{u} \right]$. We have shown that $\mu_n(K_u^c) \leq I_n(u)$, and thus in light of Definition 152, we should show that $I_n(u)$ can be made arbitrarily small. Thus, let $\epsilon > 0$, and recall that φ is continuous with $\varphi_n(t) \xrightarrow[n\to\infty]{} \varphi(t)$ for all $t \in \mathbb{R}$ and $\varphi(0) = 1$. Let us choose $u > 0$ such that $|1 - \varphi(t)| \leq \frac{\epsilon}{2}$ for all $t \in [-u, u]$. Then

$$\frac{1}{u} \left| \int_{-u}^{u} (1 - \varphi(t)) dt \right| \leq \frac{1}{u} \int_{-u}^{u} |1 - \varphi(t)| \, dt \leq \epsilon$$

and by dominated convergence, $I_n(u) \xrightarrow[n\to\infty]{} \frac{1}{u} \int_{-u}^{u} (1 - \varphi(t)) dt$. Therefore, there exists an $n_0 \in \mathbb{N}$ such that $I_n(u) < 2\epsilon$ for all $n \geq n_0$. Hence for $K = K_u$ we have $\mu_n(K^c) < 2\epsilon$, and the tightness follows by definition.

By Prokhorov's theorem, there exists a subsequence $(n_k)_{k=1}^{\infty}$ and a probability measure ν on $(\mathbb{R}, \mathcal{B})$ such that $\mu_{n_k} \xrightarrow[k\to\infty]{} \nu$ weakly. By the implication $(i) \implies (ii)$, we then have $\varphi_{n_k}(t) \xrightarrow[k\to\infty]{} \psi(t)$ for all $t \in \mathbb{R}$, where ψ is the characteristic function associated to ν. But by assumption, we also have $\varphi_{n_k}(t) \xrightarrow[k\to\infty]{} \varphi(t)$ for all $t \in \mathbb{R}$, showing that $\varphi = \psi$ by uniqueness of the limit. Hence, by Fourier inversion, $\nu = \mu$, which holds for all converging subsequences. The final claim follows from Corollary 157. □

Lévy's continuity theorem has important corollaries, which follow from the following observation: In the proof, we did not use the full continuity of the limiting characteristic function φ, but only its continuity at $t = 0$ (together with the property $\varphi(0) = 1$). We present two common reformulations.

Corollary 172. *Let μ_n, $n = 1, 2, \ldots$, be probability measures with characteristic functions φ_n, and assume $\varphi_n(t) \xrightarrow[n\to\infty]{} g(t)$ for all $t \in \mathbb{R}$, where $g : \mathbb{R} \to \mathbb{C}$ is a function continuous at $t = 0$. Then g is a characteristic function of a probability measure μ and $\mu_n \xrightarrow[n\to\infty]{} \mu$ weakly.*

Proof. Since g is continuous at $t = 0$ and $g(0) = \lim_{n\to\infty} \varphi_n(0) = 1$, we can replace φ in the proof of Lévy's continuity theorem by g, showing that the sequence $(\mu_n)_{n=1}^{\infty}$ is tight. By Prokhorov's theorem, there exists then a subsequence $(n_k)_{k=1}^{\infty}$ and a probability measure μ such that $\mu_{n_k} \xrightarrow[k\to\infty]{} \mu$ weakly. But by assumption, we also have $\varphi_{n_k}(t) \xrightarrow[k\to\infty]{} g(t)$ for all $t \in \mathbb{R}$, showing that $\varphi = g$. Therefore g is a characteristic function of μ, and the rest of the claim follows from Lévy's continuity theorem. □

Corollary 173. *Let μ_n, $n = 1, 2, \dots$, be probability measures with characteristic functions φ_n. Assume that $(\mu_n)_{n=1}^\infty$ is tight and the limit $\lim_{n\to\infty} \varphi_n(t) = g(t)$ exists for all $t \in \mathbb{R}$. Then g is a characteristic function of a probability measure μ and $\mu_n \xrightarrow[n\to\infty]{} \mu$ weakly.*

Proof. Now the tightness is given, so the same uniqueness of the limit argument as in the previous proof shows the desired claims for g and μ. □

Chapter 7

Central limit theorem

We recall the (strong) law of large numbers: if $X_1, X_2, \ldots$ are i.i.d. random variables with $\mathbb{E}(|X_1|) < \infty$ and we denote $S_n := \sum_{i=1}^{n} X_i$, then $\frac{S_n}{n} \xrightarrow[n\to\infty]{a.s.} m$, where $m := \mathbb{E}(X_1)$. That means, S_n is approximately nm when n is large. Informally, the central limit theorem concerns approximation of fluctuations of S_n around nm, which turn out to be normally distributed. To obtain this, we need to understand well both normal distribution and weak convergence of renormalized and rescaled sums of i.i.d. random variables. Our most important tool is the newly introduced characteristic function.

7.1 Normal distribution

Definition 174 (Gaussian random variable)**.** Let X be a real random variable. Then X has the *standard normal* a.k.a. *standard Gaussian* distribution if X has the density $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$. In this case, we denote $X \sim N(0,1)$ and call it a *standard Gaussian* random variable.

More generally, if $m \in \mathbb{R}$ and $\sigma^2 > 0$, then X has the *normal* a.k.a. *Gaussian* distribution with parameters m and σ^2 if X has the density $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-m)^2}{2\sigma^2}}$. In this case, we denote $X \sim N(m, \sigma^2)$ and call it a Gaussian random variable with parameters m and σ^2.

Remark 175. Obviously $f \geq 0$ in the previous definition. In addition, a standard exercise in multivariate calculus shows that $\int_{\mathbb{R}} e^{-\frac{x^2}{2}} dx = \sqrt{2\pi}$ (consider eg. the integral $\int_{\mathbb{R}^2} e^{-\frac{x^2+y^2}{2}} dxdy$ and polar coordinates). This shows directly that f is indeed a density for standard normal distribution, and the general case follows from a change of variable.

Next, we compute the characteristic function of a Gaussian random variable, which allows us to find the moments by differentiating it; recall Proposition 164 for a converse result, which generally holds for all characteristic functions. It turns out that the characteristic function of the normal distribution is infinitely many times differentiable, thus allowing us to deduce the existence of all the moments.

Proposition 176. *Let $X \sim N(m, \sigma^2)$. Then $\varphi_X(t) = \exp\left(itm - \frac{t^2\sigma^2}{2}\right)$. Furthermore, the moments $\mathbb{E}(X^n) < \infty$ exists for all $n \in \mathbb{N}$, $\mathbb{E}(X) = m$ and* $\mathrm{Var}(X) = \sigma^2$*.*

Proof. Let us first compute the moment generating function (the Laplace transform) of X, i.e. the function $z \mapsto \mathbb{E}(e^{zX})$ for $z \in \mathbb{R}$ (whenever it exists). This is a direct computation using the density of the Gaussian distribution and a change of variable $y = x - m$:

$$\begin{aligned}\mathbb{E}(e^{zX}) &= \int_{\mathbb{R}} e^{zx} f(x)dx = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{\mathbb{R}} e^{zx} e^{-\frac{(x-m)^2}{2\sigma^2}} dx = \frac{e^{zm}}{\sqrt{2\pi\sigma^2}} \int_{\mathbb{R}} e^{zy - \frac{1}{2\sigma^2}y^2} dy \\ &= \frac{e^{zm}}{\sqrt{2\pi\sigma^2}} \int_{\mathbb{R}} e^{-\frac{1}{2\sigma^2}\left(y - z\sigma^2\right)^2} e^{\frac{z^2\sigma^2}{2}} dy = e^{zm + \frac{z^2\sigma^2}{2}} \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\left(y-z\sigma^2\right)^2}{2\sigma^2}} dy \\ &= e^{zm + \frac{z^2\sigma^2}{2}}\end{aligned}$$

since $y \mapsto \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\left(y-z\sigma^2\right)^2}{2\sigma^2}}$ is a density of $N(z\sigma^2, \sigma^2)$. This computation shows in particular that $\mathbb{E}(e^{zX})$ is finite for all $z \in \mathbb{R}$, and the right-hand side expression $e^{zm+\frac{z^2\sigma^2}{2}}$ can be analytically continued to the whole complex plane $\mathbb{C}$ (by analyticity of the complex exponential). Therefore, $\mathbb{E}(e^{zX}) = e^{zm+\frac{z^2\sigma^2}{2}}$ for all $z \in \mathbb{C}$, and in particular, $\varphi_X(t) = \mathbb{E}(e^{itX}) = \exp\left(itm - \frac{t^2\sigma^2}{2}\right)$.

From its expression as a (complex) exponential, we see that $\varphi_X(t)$ can be differentiated infinitely many times, and by dominated convergence, the derivatives and integrals can be exchanged. Hence we see that $\frac{d^n}{dt^n}\varphi_X(t) = \mathbb{E}\left(\frac{d^n}{dt^n} e^{itX}\right) = \mathbb{E}((iX)^n e^{itX})$ for all $n \in \mathbb{N}$. By setting $t = 0$, we find $i^n \mathbb{E}(X^n) = \varphi_X^{(n)}(0) < \infty$ for all $n \in \mathbb{N}$. In particular, $\mathbb{E}(X) = -i\varphi'_X(0) = -i \cdot im = m$ and $\operatorname{Var}(X) = \mathbb{E}(X^2) - m^2 = -\varphi''_X(0) - m^2 = \sigma^2 + m^2 - m^2 = \sigma^2$. □

Remark 177. Note that by the Taylor series of the exponential function, we can write

$$\varphi_X(t) = \sum_{n=0}^{\infty} \frac{\varphi_X^{(n)}(0)}{n!} t^n = \sum_{n=0}^{\infty} \frac{\mathbb{E}(X^n)(it)^n}{n!}.$$

This may be compared with the expression of Proposition 164. Therefore, we have shown that the existence of all the moments and the existence of the convergent Taylor series expansion are equivalent in the case of the normal distribution. This result can be generalized to some other distributions as well.

Proposition 178. *Let $X_1, \ldots, X_n$ be independent random variables such that $X_j \sim N(m_j, \sigma_j^2)$ for all $j = 1, \ldots, n$, where $m_j \in \mathbb{R}$ and $\sigma_j^2 > 0$. Then $\sum_{j=1}^n X_j \sim N\left(\sum_{j=1}^n m_j, \sum_{j=1}^n \sigma_j^2\right)$. In particular, if $X_1, \ldots, X_n$ are i.i.d. with $X_1 \sim N(0,1)$, then $\frac{1}{\sqrt{n}} \sum_{j=1}^n X_j \sim N(0,1)$.*

Proof. Let $t \in \mathbb{R}$, and consider first the case $n = 2$. Then by independence of e^{itX_1} and e^{itX_2} and Proposition 176, we find

$$\begin{aligned}\varphi_{X_1+X_2}(t) &= \mathbb{E}\left(e^{it(X_1+X_2)}\right) = \mathbb{E}(e^{itX_1}e^{itX_2}) = \mathbb{E}(e^{itX_1})\mathbb{E}(e^{itX_2}) \\ &= e^{itm_1 - \frac{t^2\sigma_1^2}{2}} e^{itm_2 - \frac{t^2\sigma_2^2}{2}} = e^{it(m_1+m_2) - \frac{t^2(\sigma_1^2+\sigma_2^2)}{2}}.\end{aligned}$$

Thus, $\varphi_{X_1+X_2}$ is the characteristic function of a random variable $Y \sim N(m_1 + m_2, \sigma_1^2 + \sigma_2^2)$. Therefore by the Fourier inversion theorem, $X_1 + X_2 \sim N(m_1 + m_2, \sigma_1^2 + \sigma_2^2)$. The claim for general n follows then by induction.

Let us now assume $X_1, X_2, \dots$ are i.i.d. with $X_1 \sim N(0,1)$. Then by the general claim which we just showed, $S_n := X_1 + \cdots + X_n \sim N(0,n)$ and thus $\varphi_{S_n}(t) = e^{-\frac{t^2 n}{2}} = e^{-\frac{(t\sqrt{n})^2}{2}}$. It thus follows that $\varphi_{\frac{S_n}{\sqrt{n}}}(t) = \mathbb{E}\left(e^{it\frac{S_n}{\sqrt{n}}}\right) = \mathbb{E}\left(e^{i\frac{t}{\sqrt{n}}S_n}\right) = \varphi_{S_n}\left(\frac{t}{\sqrt{n}}\right) = e^{-\frac{t^2}{2}}$. This is the characteristic function of the standard normal distribution, and thus by Fourier inversion, $\frac{S_n}{\sqrt{n}} \sim N(0,1)$. □

If the i.i.d. sequence of random variables is no more (standard) Gaussian but has finite second moment, a suitably renormalized and rescaled sum of the random variables still turns out to be approximately a standard Gaussian as long as the sum is large enough. Since we are concerned about the distributions of these random sums, the correct notion is the convergence in distribution, and it is fairly easily obtained using characteristic functions. This leads us to the central limit theorem (CLT).

Theorem 179 (Central limit theorem). *Let $(X_i)_{i=1}^\infty$ be a sequence of i.i.d. random variables with $\mathbb{E}(X_1^2) < \infty$. Denote $S_n := X_1 + \cdots + X_n$, $m := \mathbb{E}(X_1)$ and $\sigma = \sqrt{\mathrm{Var}(X_1)}$. Then*

$$\frac{S_n - nm}{\sigma\sqrt{n}} \xrightarrow[n\to\infty]{d} Y$$

where $Y \sim N(0,1)$.

Proof. By Lévy's continuity theorem, it is sufficient to show that $\varphi_{\frac{S_n-nm}{\sigma\sqrt{n}}}(t) \xrightarrow[n\to\infty]{} \varphi_Y(t)$ for all $t \in \mathbb{R}$, where $Y \sim N(0,1)$. Define the renormalized random variables $Y_j := \frac{X_j - m}{\sigma}$ for all $j = 1, 2, \dots$. We note that Y_j are i.i.d. with $\mathbb{E}(Y_1) = 0$ and $\mathrm{Var}(Y_1) = 1 < \infty$. Now a direct computation shows

$$\begin{aligned}\varphi_{\frac{S_n-nm}{\sigma\sqrt{n}}}(t) &= \mathbb{E}\left(\exp\left(it\frac{S_n - nm}{\sigma\sqrt{n}}\right)\right) = \mathbb{E}\left(\exp\left(it\frac{1}{\sqrt{n}}\sum_{j=1}^n Y_j\right)\right) = \mathbb{E}\left(\prod_{j=1}^n \exp\left(it\frac{1}{\sqrt{n}}Y_j\right)\right)\\ &= \prod_{j=1}^n \mathbb{E}\left(\exp\left(it\frac{1}{\sqrt{n}}Y_j\right)\right) = \prod_{j=1}^n \varphi_{Y_j}\left(\frac{t}{\sqrt{n}}\right) = \varphi_{Y_1}\left(\frac{t}{\sqrt{n}}\right)^n.\end{aligned}$$

Recall that $\varphi_Y(t) = e^{-\frac{t^2}{2}} = \lim_{n\to\infty}\left(1 - \frac{t^2}{2n}\right)^n$. Therefore, it is enough to show $\varphi_{Y_1}\left(\frac{t}{\sqrt{n}}\right)^n \sim \left(1 - \frac{t^2}{2n}\right)^n$, and the rest of the claim follows then by elementary use of the triangle inequality. To be more precise, we apply the identity $a^n - b^n = (a-b)\sum_{j=0}^{n-1} a^j b^{n-j-1}$ $(a, b \in \mathbb{R})$ to estimate

$$\left|\varphi_{Y_1}\left(\frac{t}{\sqrt{n}}\right)^n - \left(1 - \frac{t^2}{2n}\right)^n\right| \leq \left|\varphi_{Y_1}\left(\frac{t}{\sqrt{n}}\right) - \left(1 - \frac{t^2}{2n}\right)\right| \sum_{j=0}^{n-1}\left|\varphi_{Y_1}\left(\frac{t}{\sqrt{n}}\right)\right|^j \left|1 - \frac{t^2}{2n}\right|^{n-j-1}.$$

We note that $\left|\varphi_{Y_1}\left(\frac{t}{\sqrt{n}}\right)\right| \leq 1$ and $\left|1 - \frac{t^2}{2n}\right| \leq 1$ if n is large enough, in which case we find

$$\begin{aligned}\left|\varphi_{Y_1}\left(\frac{t}{\sqrt{n}}\right)^n - \left(1 - \frac{t^2}{2n}\right)^n\right| &\leq n\left|\varphi_{Y_1}\left(\frac{t}{\sqrt{n}}\right) - \left(1 - \frac{t^2}{2n}\right)\right|\\ &= n\left|\mathbb{E}\left(e^{i\frac{t}{\sqrt{n}}Y_1}\right) - \left(1 + i\frac{t}{\sqrt{n}}\mathbb{E}(Y_1) - \frac{t^2}{2n}\mathbb{E}(Y_1^2)\right)\right|\\ &\leq n\mathbb{E}\left(\left|e^{i\frac{t}{\sqrt{n}}Y_1} - \left(1 + i\frac{t}{\sqrt{n}}Y_1 - \frac{t^2}{2n}Y_1^2\right)\right|\right).\end{aligned}$$

By Lemma 165, we have

$$\left|e^{i\frac{t}{\sqrt{n}}Y_1} - \left(1 + i\frac{t}{\sqrt{n}}Y_1 - \frac{t^2}{2n}Y_1^2\right)\right| \leq \min\left\{\frac{|tY_1|^3}{6n^{\frac{3}{2}}}, \frac{t^2|Y_1|^2}{n}\right\}.$$

Hence

$$\left|\varphi_{Y_1}\left(\frac{t}{\sqrt{n}}\right)^n - \left(1 - \frac{t^2}{2n}\right)^n\right| \leq \mathbb{E}\left(\min\left\{\frac{|tY_1|^3}{6n^{\frac{1}{2}}}, t^2|Y_1|^2\right\}\right) \xrightarrow[n\to\infty]{} 0$$

by DCT, since the right hand side is bounded by $t^2\mathbb{E}(Y_1^2) < \infty$. Finally,

$$\left|\varphi_{Y_1}\left(\frac{t}{\sqrt{n}}\right)^n - e^{-\frac{t^2}{2}}\right| \leq \left|\varphi_{Y_1}\left(\frac{t}{\sqrt{n}}\right)^n - \left(1 - \frac{t^2}{2n}\right)^n\right| + \left|\left(1 - \frac{t^2}{2n}\right)^n - e^{-\frac{t^2}{2}}\right| \xrightarrow[n\to\infty]{} 0.$$

□

Remark 180. The previous theorem shows by the definition of the convergence in distribution that

$$\mathbb{P}\left(\frac{S_n - nm}{\sigma\sqrt{n}} \leq x\right) \xrightarrow[n\to\infty]{} \mathbb{P}(Y \leq x)$$

for all $x \in \mathbb{R}$ since $N(0,1)$ is a continuous distribution. This allows us to approximate large sums of i.i.d. random variables by the Gaussian distribution. For example,

$$\mathbb{P}(S_n \geq 0) = \mathbb{P}\left(\frac{S_n - nm}{\sigma\sqrt{n}} \geq -\frac{nm}{\sigma\sqrt{n}}\right) \approx 1 - \mathbb{P}\left(Y \leq -\frac{nm}{\sigma\sqrt{n}}\right)$$

when n is large enough. Or if, for example, $X_i \sim \text{Bernoulli}(p)$ with $p > 0$, we have

$$\mathbb{P}\left(\left|\frac{S_n}{n} - p\right| \geq \epsilon\right) = \mathbb{P}\left(\left|\frac{S_n - np}{p(1-p)\sqrt{n}}\right| \geq \frac{\epsilon\sqrt{n}}{p(1-p)}\right) \approx \mathbb{P}\left(|Y| \geq \frac{\epsilon\sqrt{n}}{p(1-p)}\right).$$

Error bounds can be derived eg. by using the Berry-Esseen theorem under the assumption $\mathbb{E}(|X_1|^3) < \infty$, which gives

$$\left|\mathbb{P}\left(\frac{S_n - nm}{\sigma\sqrt{n}} \geq x\right) - \mathbb{P}(Y \geq x)\right| \leq \frac{C}{\sqrt{n}}$$

for some constant $C > 0$. We do not pursue this further in this course.

Chapter 8

Uniform integrability

In this chapter, we generalize the notion of integrabililty of a random variable to arbitrary collections of random variables. Recall that a random variable X is *integrable* if $\mathbb{E}(|X|) < \infty$. Recall also that if $\mathbb{E}(|X|^p) < \infty$ for some $p > 1$, then $\mathbb{E}(|X|) < \infty$ (Corollary 42). In the case $\mathbb{E}(|X|^p) < \infty$ for some $p \geq 1$, we denote $X \in L^p$.

Definition 181. Let $X_1, X_2, \ldots$ and X be random variables such that $X_n \in L^p$ for all $n = 1, 2, \ldots$ and $X \in L^p$, where $p \in [1, \infty)$. Then we say that X_n converge to X in L^p (or more precisely, in the L^p-norm) if $\mathbb{E}(|X_n - X|^p) \xrightarrow[n\to\infty]{} 0$. In this case, we denote $X_n \xrightarrow[n\to\infty]{L^p} X$.

We notice the following elementary relationship between L^p convergence and convergence in probability.

Proposition 182. *Let $X_1, X_2, \ldots$ and X be random variables such that $X_n \xrightarrow[n\to\infty]{L^p} X$. Then $X_n \xrightarrow[n\to\infty]{\mathbb{P}} X$.*

Proof. This is a simple application of Markov's inequality, which is left as an exercise to the reader. □

One might wonder if the converse holds. It turns out that this is not the case even if $p = 1$, which is the weakest form of L^p-convergence. In order to find a useful converse result, we introduce the notion of *uniform integrability.*

Definition 183 (Uniform integrability)**.** Let $\{X_i\}_{i\in I}$ be an arbitrary collection of random variables indexed by a set I. Then the collection $\{X_i\}_{i\in I}$ is *uniformly integrable* (UI) if

$$\sup_{i\in I} \mathbb{E}\left(|X_i| \, \mathbb{1}_{|X_i|\geq k}\right) \xrightarrow[k\to\infty]{} 0.$$

In other words, $\{X_i\}_{i\in I}$ is UI if for all $\epsilon > 0$ there exist $k_0 \in \mathbb{N}$ such that for all $k \geq k_0$, $\mathbb{E}(|X_i| \, \mathbb{1}_{|X_i|\geq k}) \leq \epsilon$ for all $i \in I$.

Remark 184. Intuitively, uniform integrability means that large values of X_i do not significantly contribute to $\mathbb{E}(X_i)$ uniformly over $i \in I$.

Remark 185. If X is integrable, then $\{X\}$ is uniformly integrable. To see this, we simply note that $\mathbb{E}(|X|\,\mathbb{1}_{|X|\geq k}) \xrightarrow[k\to\infty]{} 0$ by the dominated convergence theorem since $\mathbb{E}(|X|) < \infty$. The converse also holds as a special case of Proposition 187 below. It is also easy to see that a finite collection of integrable random variables $\{X_1, X_2, \dots, X_n\}$ is uniformly integrable.

Example 186. Fix $\alpha > 0$ and consider the random variables $X_n \in \{0, n^\alpha\}$, $n = 1, 2, \dots,$ whose law is given by $\mathbb{P}(X_n = 0) = 1 - \frac{1}{n}$ and $\mathbb{P}(X_n = n^\alpha) = \frac{1}{n}$. We claim that $\{X_n\}_{n=1}^\infty$ is UI if and only if $\alpha < 1$. To see this, we observe $\mathbb{E}(|X_n|\,\mathbb{1}_{|X_n|\geq k}) = n^\alpha \mathbb{1}_{n^\alpha \geq k} \cdot \frac{1}{n} = n^{\alpha-1}\mathbb{1}_{n^\alpha\geq k}$. Then

$$I(k) := \sup_{n\geq 1} \mathbb{E}(|X_n|\,\mathbb{1}_{|X_n|\geq k}) = \sup_{n\geq 1} n^{\alpha-1}\mathbb{1}_{n^\alpha \geq k} = \begin{cases} \infty & \text{if } \alpha > 1 \\ 1 & \text{if } \alpha = 1. \end{cases}$$

This shows that $\{X_n\}_{n=1}^\infty$ is not UI if $\alpha \geq 1$. If $0 < \alpha < 1$, then we have $I(k) \xrightarrow[k\to\infty]{} 0$ since $n^{\alpha-1} \xrightarrow[n\to\infty]{} 0$ and $\mathbb{1}_{n^\alpha \geq k} \xrightarrow[n\to\infty]{} 1$. Thus, in this case $\{X_n\}_{n=1}^\infty$ is UI.

Uniform integrability indeed gives good control on the random variables, and it is in fact stronger than boundedness in the L^1 norm.

Proposition 187. *If $\{X_i\}_{i\in I}$ is UI then $\{X_i\}_{i\in I}$ is bounded in L^1. That is, there exists a constant $C \in (0, \infty)$ such that $\mathbb{E}(|X_i|) \leq C$ for all $i \in I$.*

Proof. For any $k \geq 0$ fixed, we have

$$\mathbb{E}(|X_i|) = \mathbb{E}(|X_i|\,\mathbb{1}_{|X_i|<k}) + \mathbb{E}(|X_i|\,\mathbb{1}_{|X_i|\geq k}) \leq k + \sup_{j\geq 1} \mathbb{E}(|X_j|\,\mathbb{1}_{|X_j|\geq k}) < \infty.$$

□

The converse result does not hold. As a counterexample, consider the case $\alpha = 1$ in Example 186, where $\mathbb{E}(|X_n|) = 1$ for all $n \geq 1$ but uniform integrability does not hold. This shows that uniform integrability is indeed stronger that boundedness in L^1. However, boundedness in L^p for $p > 1$ turns out to be sufficiently strong to guarantee uniform integrability.

Proposition 188. *Let $\{X_i\}_{i\in I}$ be a family of random variables. Assume that there exist a constant $C < \infty$ and $p > 1$ such that $\mathbb{E}(|X_i|^p) \leq C$ for all $i \in I$. Then $\{X_i\}_{i\in I}$ is uniformly integrable.*

Proof. Let q be the Hölder conjugate of p, i.e. $q > 1$ such that $\frac{1}{p} + \frac{1}{q} = 1$. Then by Hölder's inequality, we find

$$\mathbb{E}(|X_i|\,\mathbb{1}_{|X_i|\geq k}) \leq \mathbb{E}(|X_i|^p)^{\frac{1}{p}} \mathbb{E}\left(\mathbb{1}_{|X_i|\geq k}\right)^{\frac{1}{q}} \leq C^{\frac{1}{p}} \mathbb{P}(|X_i| \geq k)^{\frac{1}{q}} = C^{\frac{1}{p}} \mathbb{P}(|X_i|^p \geq k^p)^{\frac{1}{q}}.$$

Now by Markov's inequality,

$$\mathbb{P}(|X_i|^p \geq k^p) \leq \frac{\mathbb{E}(|X_i|^p)}{k^p} \leq \frac{C}{k^p},$$

and hence $\sup_{i\in I} \mathbb{E}\left(|X_i|\,\mathbb{1}_{|X_i|\geq k}\right) \leq \frac{C}{k^{\frac{p}{q}}} \xrightarrow[k\to\infty]{} 0$ showing the claim. □

A simpler sufficient condition for uniform integrability is found when the family of random variables in question is bounded by an integrable random variable.

Proposition 189. *Let X_i, $i \in I$ and Z be random variables such that $|X_i| \leq Z$ for all $i \in I$ and $\mathbb{E}(Z) < \infty$. Then $\{X_i\}_{i\in I}$ is UI.*

Proof. We readily obtain

$$\sup_{i\in I} \mathbb{E}(|X_i| \mathbb{1}_{|X_i|\geq k}) \leq \mathbb{E}(Z\mathbb{1}_{Z\geq k}) \xrightarrow[k\to\infty]{} 0$$

by dominated convergence. □

We show next that the sum of two uniformly integrable families of random variables in again uniformly integrable.

Lemma 190. *Let $\{X_i\}_{i\in I}$ and $\{Y_i\}_{i\in I}$ be two uniformly integrable families of random variables (defined on the same index set I). Then $\{X_i + Y_i\}_{i\in I}$ is UI.*

Proof. We begin with an estimate

$$\begin{aligned} |X_i + Y_i| \mathbb{1}_{|X_i+Y_i|\geq 2k} &\leq (|X_i| + |Y_i|)\mathbb{1}_{\{|X_i|\geq k\}\cup\{|Y_i|\geq k\}} \leq (|X_i| + |Y_i|)(\mathbb{1}_{|X_i|\geq k} + \mathbb{1}_{|Y_i|\geq k}) \\ &= |X_i| \mathbb{1}_{|X_i|\geq k} + |Y_i| \mathbb{1}_{|Y_i|\geq k} + |X_i| \mathbb{1}_{|Y_i|\geq k} + |Y_i| \mathbb{1}_{|X_i|\geq k}. \end{aligned}$$

We further estimate the cross terms as

$$|X_i| \mathbb{1}_{|Y_i|\geq k} = |X_i| \mathbb{1}_{|Y_i|\geq k}\mathbb{1}_{|X_i|\geq |Y_i|} + |X_i| \mathbb{1}_{|Y_i|\geq k}\mathbb{1}_{|X_i|<|Y_i|} \leq |X_i| \mathbb{1}_{|X_i|\geq k} + |Y_i| \mathbb{1}_{|Y_i|\geq k},$$

and by symmetry, $|Y_i| \mathbb{1}_{|X_i|\geq k} \leq |X_i| \mathbb{1}_{|X_i|\geq k} + |Y_i| \mathbb{1}_{|Y_i|\geq k}$. Therefore,

$$|X_i + Y_i| \mathbb{1}_{|X_i+Y_i|\geq 2k} \leq 3(|X_i| \mathbb{1}_{|X_i|\geq k} + |Y_i| \mathbb{1}_{|Y_i|\geq k}).$$

Hence we obtain

$$\sup_{i\in I} \mathbb{E}(|X_i + Y_i| \mathbb{1}_{|X_i+Y_i|\geq 2k}) \leq 3 \left(\sup_{i\in I} \mathbb{E}(|X_i| \mathbb{1}_{|X_i|\geq k}) + \sup_{i\in I} \mathbb{E}(|Y_i| \mathbb{1}_{|Y_i|\geq k})\right)$$

and the claim follows by letting $k \to \infty$. □

We are now ready to show the main result of this chapter, which is a characterization of the convergence in L^1.

Theorem 191. *Let $(X_n)_{n=1}^{\infty}$ be a sequence of random variables and X a random variable. Then the following are equivalent.*

(i) $X_n \xrightarrow[n\to\infty]{\mathbb{P}} X$ *and* $\{X_n\}_{n=1}^{\infty}$ *is uniformly integrable.*

(ii) $X_n \xrightarrow[n\to\infty]{L^1} X$.

Proof. Let us first prove the implication $(ii) \implies (i)$. By Proposition 182, it is enough to show that $\{X_n\}_{n=1}^\infty$ is uniformly integrable. Observe that by the assumption (ii), X is integrable. Since $X_n = X_n - X + X$, it is enough to show that $\{X_n - X\}_{n=1}^\infty$ is uniformly integrable by the previous lemma and Remark 185. By assumption, $\mathbb{E}(|X_n - X|) \xrightarrow[n\to\infty]{} 0$, hence for all $\epsilon > 0$ there exists $n_0 \in \mathbb{N}$ such that $\mathbb{E}(|X_n - X|) < \epsilon$ if $n > n_0$. Since $\{X_1 - X, \ldots, X_{n_0} - X\}$ is a finite collection of integrable random variables, it is uniformly integrable. Thus, $\sup_{1\le n\le n_0} \mathbb{E}(|X_n - X| \mathbb{1}_{|X_n - X|\ge k}) < \epsilon$ for k large enough. We also have $\sup_{n>n_0} \mathbb{E}(|X_n - X| \mathbb{1}_{|X_n-X|\ge k}) \le \sup_{n>n_0} \mathbb{E}(|X_n - X|) \le \epsilon$. Hence,

$$\sup_{n\ge 1} \mathbb{E}(|X_n - X| \mathbb{1}_{|X_n-X|\ge k}) \le \epsilon$$

for k large enough, and the claim (i) follows.

Let us then show the implication $(i) \implies (ii)$. Assuming (i), then in particular $X_n \in L^1$ for all $n = 1, 2, \ldots$ by uniform integrability. On the other hand, since $X_n \xrightarrow[n\to\infty]{\mathbb{P}} X$, then there exists a subsequence $(n_k)_{k=1}^\infty$ such that $X_{n_k} \xrightarrow[k\to\infty]{a.s.} X$ by Proposition 125. Now by Fatou's lemma, $\mathbb{E}(|X|) \le \liminf_{k\to\infty} \mathbb{E}(|X_{n_k}|) \le C < \infty$, where C is some constant guaranteed by Proposition 187. Thus, $X \in L^1$.

Consider now $Y_n := X_n - X$. Then our claim is $\mathbb{E}(|Y_n|) \xrightarrow[n\to\infty]{} 0$. Now $\{Y_n\}_{n=1}^\infty$ is UI by the previous lemma and Remark 185. Let $\epsilon > 0$. Then by uniform integrability, $\mathbb{E}(|Y_n| \mathbb{1}_{|Y_n|\ge k}) \le \epsilon$ for k large enough. Hence,

$$\mathbb{E}(|Y_n|) = \mathbb{E}(|Y_n| \mathbb{1}_{|Y_n|\ge k}) + \mathbb{E}(|Y_n| \mathbb{1}_{|Y_n|\le \epsilon}) + \mathbb{E}(|Y_n| \mathbb{1}_{\epsilon<|Y_n|<k}) \le 2\epsilon + k\mathbb{P}(|Y_n| > \epsilon)$$

for k large enough. By the convergence in probability, we also have $\mathbb{P}(|Y_n| > \epsilon) \xrightarrow[n\to\infty]{} 0$. Thus, $\limsup_{n\to\infty} \mathbb{E}(|Y_n|) \le 2\epsilon$, which holds for all $\epsilon > 0$. Hence, $\lim_{n\to\infty} \mathbb{E}(|Y_n|) = 0$. □

Finally, let us explore the connection of uniform integrability to tightness by the following proposition.

Proposition 192. *Let X_i $(i \in I)$ be random variables with respective laws μ_i. If $\{X_i\}_{i\in I}$ is UI, then $\{\mu_i\}_{i\in I}$ is tight.*

Proof. Let $\epsilon > 0$. Then by uniform integrability, there exists $k \ge 1$ such that $\mathbb{E}(|X_i| \mathbb{1}_{|X_i|\ge k}) \le \epsilon$. Therefore, $k\mathbb{P}(|X_i| \ge k) \le \mathbb{E}(|X_i| \mathbb{1}_{|X_i|\ge k}) \le \epsilon$. This implies

$$\mu_i([-k, k]) = \mathbb{P}(-k \le X_i \le k) \ge \mathbb{P}(-k < X_i < k) \ge 1 - \frac{\epsilon}{k} \ge 1 - \epsilon.$$

Choosing $[a, b] = [-k, k]$ in the definition of tightness (as in Definition 152 and Remark 153) completes the proof. □

Chapter 9

Conditional expectation and martingales

The main motivation to consider conditional expectations is to understand random experiments with multiple layers of randomness. A simple example would be eg. to consider a collection of written texts, with X being the (random) number of typos in a page and N is the number of typos in a page which are corrected by a spell-checker. Let us assume that the spell-checker spots and corrects a typo independently with probability $p \in (0, 1)$ (where p should be close to 1 if the spell-checker does the work well). We would like to understand the average number of typos corrected on a given page, which should then be equal to pX. We denote this expected number of corrected typos by $\mathbb{E}(N|X)$ and call it the *conditional expectation* of N given the random variable X. In order to show that $\mathbb{E}(N|X) = pX$, we need some general theory, which turns out to generalize to way more complicated and interesting settings. This also includes *martingales*, which will be the main topic at the end of this course.

9.1 Conditional expectation

We begin with the construction of the conditional expectation in the case of discrete random variables, which is more intuitive. Then, we notice that it can be in fact easily generalized to very general settings.

9.1.1 Conditional expectation in the discrete case

Recall the definition of *conditional probability* (Definition 55) That is, if $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space and $B \in \mathcal{F}$ is an event such that $\mathbb{P}(B) > 0$, then $\mathbb{P}(A|B) = \frac{\mathbb{P}(A\cap B)}{\mathbb{P}(B)}$ for all $A \in \mathcal{F}$. This gives the probability of A given that we have observed $\omega \in B$. Assume now $0 < \mathbb{P}(B) < 1$. Since B and B^c partition Ω with $\mathbb{P}(B) > 0$ and $\mathbb{P}(B^c) > 0$, we may define a random variable Z on $(\Omega, \mathcal{F}, \mathbb{P})$ by setting

$$Z(\omega) := \begin{cases} \mathbb{P}(A|B) & \text{if } \omega \in B \\ \mathbb{P}(A|B^c) & \text{if } \omega \in B^c. \end{cases}$$

More generally:

Definition 193 (Conditional probability as a random variable)**.** If $B_1, B_2, \dots$ are events such that $B_i \cap B_j = \emptyset$ for $i \neq j$ and $\Omega = \bigcup_{i=1}^{\infty} B_i$, i.e. $\{B_1, B_2, \dots\}$ is a partition of Ω, then for an event A we define $Z(\omega) = \mathbb{P}(A|B_i)$ if $\omega \in B_i$ with the convention that $Z(\omega) = 0$ if $\mathbb{P}(B_i) = 0$.

Let $\mathcal{G} := \sigma(B_i : i = 1, 2, \dots)$. Then the the conditional probability of A given the σ-algebra $\mathcal{G}$ is the random variable $Z = \mathbb{P}(A|\mathcal{G}) = \sum_{n=1}^{\infty} \mathbb{P}(A|B_n)\mathbb{1}_{B_n}$.

We note that $\mathbb{P}(A|B) = \frac{\mathbb{E}(\mathbb{1}_A \mathbb{1}_B)}{\mathbb{P}(B)}$. The basic idea of defining the conditional expectation is then to replace $\mathbb{1}_A$ by a more general random variable X.

Definition 194 (Conditional expectation given an event)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $B \in \mathcal{F}$ such that $\mathbb{P}(B) > 0$. Let $X \geq 0$ be a random variable. Then the *conditional expectation* of X given B is defined by $\mathbb{E}(X|B) := \frac{\mathbb{E}(X\mathbb{1}_B)}{\mathbb{P}(B)}$.

We can now state an analogous definition as 193.

Definition 195 (Conditional expectation as a random variable)**.** Let $X \geq 0$ be a random variable and $\{B_1, B_2, \dots\}$ be a partition of Ω. Then the *conditional expectation* of X is defined as $\mathbb{E}(X|B_i)$ if $\omega \in B_i$ with the convention that $\mathbb{E}(X|B_i) = 0$ if $\mathbb{P}(B_i) = 0$.

Let $\mathcal{G} := \sigma(B_i : i = 1, 2, \dots)$. Then the conditional expectation of X given the σ-algebra $\mathcal{G}$ is the random variable[1] $Z := \mathbb{E}(X|\mathcal{G})$ given by

$$Z(\omega) = \sum_{n=1}^{\infty} \mathbb{E}(X|B_n)\mathbb{1}_{B_n}(\omega).$$

Remark 196. The conditional expectation in the case of a discrete partition of the space Ω can also be defined for $X \in L^1$. This is left as an exercise to the reader.

The following two examples are extracted from [4, Chapter 11].

Example 197. Consider a fair die, i.e. $\Omega = \{1, 2, 3, 4, 5, 6\}$ equipped with its power set as the σ-algebra and $\mathbb{P}$ the uniform measure on Ω. Consider the partition of Ω given by $B_1 = \{1, 3, 5\}$ and $B_2 = \{2, 4, 6\}$. Let $X(\omega) = \omega$ for $\omega \in \Omega$ represent a coin toss, and consider $\mathcal{G} = \sigma(B_1, B_2)$. Then by the previous definition,

$$\mathbb{E}(X|\mathcal{G}) = \mathbb{E}(X|B_1)\mathbb{1}_{B_1} + \mathbb{E}(X|B_2)\mathbb{1}_{B_2} = \frac{\mathbb{E}(X\mathbb{1}_{B_1})}{\mathbb{P}(B_1)} + \frac{\mathbb{E}(X\mathbb{1}_{B_2})}{\mathbb{P}(B_2)} = 3\mathbb{1}_{B_1} + 4\mathbb{1}_{B_2}.$$

That is, $\mathbb{E}(X|\mathcal{G})(\omega) = 3$ if $\omega \in B_1$ and $\mathbb{E}(X|\mathcal{G})(\omega) = 4$ if $\omega \in B_2$, matching well with our intuition.

Example 198. Consider now $\Omega = (0, 1]$, equipped with its Borel $\sigma-$algebra and the Lebesgue measure $\mathbb{P}$. Let $\mathcal{G}_n = \sigma\big\{ \big(\frac{i-1}{n}, \frac{i}{n}\big] : i \in \{1, \dots, n\}\big\}$ where $n \geq 1$ is fixed. Let $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$, that is, $\int_0^1 |X(\omega)|\, d\omega < \infty$. Then we find

$$\mathbb{E}(X|\mathcal{G}_n) = \sum_{i=1}^{n} \mathbb{E}\left(X \middle| \left(\frac{i-1}{n}, \frac{i}{n}\right]\right) \mathbb{1}_{\left(\frac{i-1}{n}, \frac{i}{n}\right]} = \sum_{i=1}^{n} \frac{\mathbb{E}\left(X\mathbb{1}_{\left(\frac{i-1}{n}, \frac{i}{n}\right]}\right)}{\mathbb{P}\left(\left(\frac{i-1}{n}, \frac{i}{n}\right]\right)} \mathbb{1}_{\left(\frac{i-1}{n}, \frac{i}{n}\right]} = \sum_{i=1}^{n} X_i \mathbb{1}_{\left(\frac{i-1}{n}, \frac{i}{n}\right]}$$

where $X_i = n\int_{\frac{i-1}{n}}^{\frac{i}{n}} X(\omega)d\omega$.

[1] Here we allow infinite values for a random variable with positive probability. That is, a random variable is a measurable mapping $\Omega \to [0, \infty]$ with this convention.

We observe the following key properties of the above defined conditional expectation.

Proposition 199. *Let $X \geq 0$ be a random variable on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with conditional expectation $\mathbb{E}(X|\mathcal{G})$ defined as in Definition 195. Then the following hold.*

(i) $\mathbb{E}(X|\mathcal{G})$ is measurable with respect to the sub-σ-algebra $\mathcal{G} \subset \mathcal{F}$.

(ii) $\mathbb{E}\left(\mathbb{E}(X|\mathcal{G})\mathbb{1}_B\right) = \mathbb{E}(X\mathbb{1}_B)$ for all $B \in \mathcal{G}$ (and in particular, $\mathbb{E}\left(\mathbb{E}(X|\mathcal{G})\right) = \mathbb{E}(X)$).

Proof. $\mathbb{E}(X|\mathcal{G})$ is by construction measurable with respect to $\mathcal{G}$. Let $B \in \mathcal{G}$. Without loss of generality, we may assume $B = B_n$ for some $n = 1, 2, \dots,$ where $\{B_1, B_2, \dots\}$ is the partition of Ω. Then

$$\mathbb{E}\left(\mathbb{E}(X|\mathcal{G})\mathbb{1}_{B_n}\right) = \mathbb{E}\left(\mathbb{E}(X|B_n)\mathbb{1}_{B_n}\right) = \mathbb{E}(X|B_n)\mathbb{E}\left(\mathbb{1}_{B_n}\right) = \frac{\mathbb{E}(X\mathbb{1}_{B_n})}{\mathbb{P}(B_n)}\mathbb{P}\left(B_n\right) = \mathbb{E}(X\mathbb{1}_{B_n}).$$

□

We use the properties of the above proposition as the general definition of the conditional expectation.

9.1.2 Conditional expectation in the general case

Definition 200 (Conditional expectation)**.** Let X be a random variable defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Assume furthermore that X is integrable. Let $\mathcal{G} \subset \mathcal{F}$ be a sub-σ-algebra. Then a random variable Z on $(\Omega, \mathcal{F}, \mathbb{P})$ is called a *conditional expectation* of X if the following holds:

(i) Z is measurable with respect to $\mathcal{G}$.

(ii) $\mathbb{E}\left(Z\mathbb{1}_B\right) = \mathbb{E}(X\mathbb{1}_B)$ for all $B \in \mathcal{G}$.

We denote $Z =: \mathbb{E}(X|\mathcal{G})$

For the existence and the uniqueness of the conditional expectation, we use the Radon-Nikodym theorem from measure theory. It is stated for σ-finite measures which satisfy an absolute continuity relation.

Definition 201 (Absolute continuity of measures)**.** A measure ν is absolutely continuous with respect to a measure μ if for all measurable sets A, $\mu(A) = 0$ implies $\nu(A) = 0$. In this case, we denote $\nu << \mu$.

Definition 202 (σ-finite measure)**.** A measure μ on Ω equipped with a σ-algebra $\mathcal{A}$ is σ-finite if there exist $B_1, B_2, \dots \in \mathcal{A}$ such that $\bigcup_{n=1}^{\infty} B_n = \Omega$ and $\mu(B_n) < \infty$ for all $n = 1, 2, \dots$.

Theorem 203 (Radon-Nikodym)**.** *Let μ and ν be σ-finite measures on a set Ω equipped with a σ-algebra $\mathcal{A}$. Then the following are equivalent.*

(i) $\nu << \mu$

(ii) There exists a function $f : \Omega \to \mathbb{R}_+$ which is measurable with respect to $\mathcal{A}$ and $\nu(A) = \int_\Omega f\mathbb{1}_A d\mu$ for all $A \in \mathcal{A}$.

Remark 204. If we allow a generalization of the notion of density function to arbitrary probability spaces (instead of just real numbers), then the function f in the Radon-Nikodym theorem is a density with respect to the measure μ. We call it the *Radon-Nikodym* derivative and denote $f = \frac{d\nu}{d\mu}$. The Radon-Nikodym derivative provides a useful change of measure also beyond the scope of this course.

Theorem 205 (Existence and uniqueness of the conditional expectation)**.** *Let X be an integrable random variable on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let $\mathcal{G} \subset \mathcal{F}$ be a sub-σ-algebra. Then the conditional expectation $\mathbb{E}(X|\mathcal{G})$ exists and is unique up to zero measurable sets. The latter means that if Z_1 and Z_2 are conditional expectations (i.e. satisfy Definition 200), then $Z_1 = Z_2$ almost surely.*

In addition, $\mathbb{E}(X|\mathcal{G})$ is integrable. Furthermore, if $X \geq 0$ a.s., then also $\mathbb{E}(X|\mathcal{G}) \geq 0$ a.s.

Proof. Assume first $X \geq 0$ a.s. Define a measure $\mathbb{Q}$ by $\mathbb{Q}(A) := \mathbb{E}(X\mathbb{1}_A)$ for all $A \in \mathcal{F}$ (where $\mathbb{E}$ is the expectation with respect to $\mathbb{P}$ as usual). The measure $\mathbb{Q}$ is σ-finite, since we may choose $B_n = \{X \leq n\}$ for every $n \in \mathbb{N}$. The measure $\mathbb{P}$ is trivially σ-finite. Moreover, if $\mathbb{P}(A) = 0$, then $\mathbb{Q}(A) = 0$, since $X\mathbb{1}_A = 0$ a.s. when A has zero measure. Therefore $\mathbb{Q} << \mathbb{P}$, and this relation passes to the restricted measures $\mathbb{Q}_{|\mathcal{G}}$ and $\mathbb{P}_{|\mathcal{G}}$. Hence by the Radon-Nikodym theorem, there exist a random variable $Z : \Omega \to \mathbb{R}_+$ which is measurable with respect to $\mathcal{G}$ and for which we have $\mathbb{E}(X\mathbb{1}_B) = \mathbb{Q}(B) = \int_\Omega Z\mathbb{1}_B d\mathbb{P} = \mathbb{E}(Z\mathbb{1}_B)$. By definition, $Z \geq 0$ is then a conditional expectation.

Assume then that X is a general integrable random variable. We write $X = X^+ - X^-$, where $X^+ = \max\{X, 0\}$ and $X^- = -\min\{X, 0\}$. Then X^+ and X^- are non-negative, so we could apply the above construction separately to them to construct $Z^\pm = \mathbb{E}(X^\pm|\mathcal{G})$. Define $Z := Z^+ - Z^-$, which is trivially measurable w.r.t. $\mathcal{G}$. Since $\mathbb{E}(Z^\pm) = \mathbb{E}(X^\pm) < \infty$, it follows that $\mathbb{E}(|Z|) < \infty$. By linearity of expectation and the definition of the conditional expectation, we then see that

$$\mathbb{E}(Z\mathbb{1}_B) = \mathbb{E}(Z^+\mathbb{1}_B) - \mathbb{E}(Z^-\mathbb{1}_B) = \mathbb{E}(X^+\mathbb{1}_B) - \mathbb{E}(X^-\mathbb{1}_B) = \mathbb{E}(X\mathbb{1}_B)$$

for all $B \in \mathcal{G}$. Thus, Z is a conditional expectation of X.

Finally, let us show the uniqueness. Let Z_1 and Z_2 be two conditional expectations of X. Then they are in particular $\mathcal{G}$-measurable, so $B := \{Z_1 < Z_2\} \in \mathcal{G}$. By the definition of the conditional expectation, we then have $\mathbb{E}(Z_1\mathbb{1}_B) = \mathbb{E}(X\mathbb{1}_B) = \mathbb{E}(Z_2\mathbb{1}_B)$, so by integrability of Z_1 and Z_2, we find $\mathbb{E}((Z_2 - Z_1)\mathbb{1}_B) = 0$. Since $(Z_2 - Z_1)\mathbb{1}_B \geq 0$ a.s., we must then have $(Z_2 - Z_1)\mathbb{1}_B = 0$ a.s. Thus, $\mathbb{P}(B) = 0$, which implies that $Z_1 \geq Z_2$ a.s. By symmetry, we also have $Z_2 \geq Z_1$ a.s., showing that $Z_1 = Z_2$ a.s. □

Remark 206 (Extension to general non-negative random variables)**.** The notion of conditional expectation can be extended to non-negative random variables which are not necessarily integrable. In this case, we have to accept a random variable Z to possibly attain value ∞, i.e. Z is a function $\Omega \to [0, \infty]$. This is principally because even if $X \geq 0$ is itself a.s. finite, its expectation may still be infinite, so the conditional expectation of X may be infinite with a positive probability.

With the above convention, we then define $\mathbb{E}(X|\mathcal{G}) := \lim_{n\to\infty} \mathbb{E}\left(\min\{X, n\}|\mathcal{G}\right)$ a.s.. If furthermore X is integrable, it can be shown that $\mathbb{E}(X|\mathcal{G})$ is the same unique limit as in Definition 200.

The definition of the conditional expectation is rather indirect, but there are several cases in which the conditional expectation can in fact be computed.

Proposition 207. *The following properties allow to compute the conditional expectation.*

(i) If X is $\mathcal{G}$-measurable, then $\mathbb{E}(X|\mathcal{G}) = X$.

(ii) If $\mathcal{G} = \{\emptyset, \Omega\}$, then $\mathbb{E}(X|\mathcal{G}) = \mathbb{E}(X)$.

(iii) If X and $\mathcal{G}$ are independent, then $\mathbb{E}(X|\mathcal{G}) = \mathbb{E}(X)$.

Proof. The claim (i) is a direct consequence of the definition. In all the three claims, the $\mathcal{G}$-measurability of the quantities is also clear. Note also that the claim (ii) follows from (iii).

Let us anyway offer a direct proof of (ii) for illustrational purposes. Thus, assume $\mathcal{G} = \{\emptyset, \Omega\}$ and let $B \in \mathcal{G}$. Then

$$\mathbb{E}(\mathbb{E}(X)\mathbb{1}_B) = \mathbb{E}(X)\mathbb{E}(\mathbb{1}_B) = \mathbb{E}(X)\mathbb{P}(B) = \begin{cases} 0 & (B = \emptyset) \\ 1 & (B = \Omega). \end{cases}$$

The same clearly holds for $\mathbb{E}(X\mathbb{1}_B)$, and the claim (ii) follows then from the definition of the conditional expectation.

Finally, let us just assume X and $\mathcal{G}$ are independent and $B \in \mathcal{G}$. Then $\mathbb{E}(X\mathbb{1}_B) = \mathbb{E}(X)\mathbb{E}(\mathbb{1}_B) = \mathbb{E}(\mathbb{E}(X)\mathbb{1}_B)$ by the independence and the linearity of the expectation. The claim (iii) follows. □

The following property tells us that when conditioning to nested sigma-algebras, the smallest only matters for the conditional expectation. This is also practical in order to evaluate conditional expectations.

Proposition 208 (The tower property of conditional expectation)**.** *If $\mathcal{G}_1 \subset \mathcal{G}_2 \subset \mathcal{F}$, then the following hold:*

(i) $\mathbb{E}\left(\mathbb{E}(X|\mathcal{G}_1)|\mathcal{G}_2\right) = \mathbb{E}(X|\mathcal{G}_1)$

(ii) $\mathbb{E}\left(\mathbb{E}(X|\mathcal{G}_2)|\mathcal{G}_1\right) = \mathbb{E}(X|\mathcal{G}_1)$

Proof. We denote $Z_i := \mathbb{E}(X|\mathcal{G}_i)$ for $i = 1, 2$ and note that, by definition, Z_i is $\mathcal{G}_i$-measurable. Since $\mathcal{G}_1 \subset \mathcal{G}_2$, Z_1 is also $\mathcal{G}_2$-measurable. Thus, $\mathbb{E}(Z_1|\mathcal{G}_2) = Z_1$ by Proposition 207 (i). Hence, we have proven claim (i).

For claim (ii), let $B \in \mathcal{G}_1$, which readily implies $B \in \mathcal{G}_2$. Then $\mathbb{E}(Z_1\mathbb{1}_B) = \mathbb{E}(X\mathbb{1}_B)$ and $\mathbb{E}(Z_2\mathbb{1}_B) = \mathbb{E}(X\mathbb{1}_B)$ by the definitions of the conditional expectations Z_1 and Z_2, respectively. Thus, $\mathbb{E}(Z_1\mathbb{1}_B) = \mathbb{E}(Z_2\mathbb{1}_B)$ for all $B \in \mathcal{G}_1$, which yields $\mathbb{E}(Z_2|\mathcal{G}_1) = Z_1$ by the definition of the conditional expectation. □

Proposition 209 (Further properties of the conditional expectation)**.** *Let X and Y be random variables defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and let $\mathcal{G} \subset \mathcal{F}$ be a sub-σ-algebra.*

(i) If X and Y are integrable and $a, b \in \mathbb{R}$, then $\mathbb{E}(aX + bY|\mathcal{G}) = a\mathbb{E}(X|\mathcal{G}) + b\mathbb{E}(Y|\mathcal{G})$. (Linearity)

(ii) Assume Y is $\mathcal{G}$-measurable and either of the following conditions hold:

(a) $Y \geq 0$ *and* $X \geq 0$

(b) Y*,* X *and* YX *are integrable.*

Then $\mathbb{E}(YX|\mathcal{G}) = Y\mathbb{E}(X|\mathcal{G})$*.*

(iii) If $X \leq Y$*, then* $\mathbb{E}(X|\mathcal{G}) \leq \mathbb{E}(Y|\mathcal{G})$*. (Monotonicity)*

(iv) If X *is integrable then* $|\mathbb{E}(X|\mathcal{G})| \leq \mathbb{E}(|X|\,|\mathcal{G})$*.*

Proof. The claim (i) follows from linearity of the usual expectation using the definition of the conditional expectation and its uniqueness. The claim (ii) is more delicate. Let us show the positive case (the integrable case follows by decomposing the random variables into positive and negative parts). Assume first that $Y = \mathbb{1}_B$ for some $B \in \mathcal{G}$. Then by definition, $Y\mathbb{E}(X|\mathcal{G})$ is $\mathcal{G}$-measurable. If $A \in \mathcal{G}$, then $\mathbb{E}(Y\mathbb{E}(X|\mathcal{G})\mathbb{1}_A) = \mathbb{E}(\mathbb{1}_B\mathbb{E}(X|\mathcal{G})\mathbb{1}_A) = \mathbb{E}(\mathbb{E}(X|\mathcal{G})\mathbb{1}_{A\cap B}) = \mathbb{E}(X\mathbb{1}_{A\cap B}) = \mathbb{E}(\mathbb{1}_B X\mathbb{1}_A) = \mathbb{E}(YX\mathbb{1}_A)$. Hence by the definition of the conditional expectation, $\mathbb{E}(YX|\mathcal{G}) = Y\mathbb{E}(X|\mathcal{G})$. The claim clearly also holds for linear combinations of indicators with positive coefficients by linearity of the conditional expectation. Hence, the claim holds if Y is a non-negative step function. Finally, if just $Y \geq 0$, there exists an increasing sequence of step functions $Y_1, Y_2, \ldots$ such that $Y_n \xrightarrow[n\to\infty]{a.s.} Y$. If $A \in \mathcal{G}$, we then have $\mathbb{E}(Y_n X\mathbb{1}_A) = \mathbb{E}(Y_n\mathbb{E}(X|\mathcal{G})\mathbb{1}_A)$, and the claim follows by applying the monotone convergence theorem (note that all the random variables are non-negative).

For claim (iii), we note that $Y - X \geq 0$, so the last claim in Theorem 205 implies $\mathbb{E}(Y - X|\mathcal{G}) \geq 0$. Then the claim (iii) follows by linearity if X, Y are integrable, and otherwise after applying Remark 206. Finally, since $-X \leq |X|$ and $X \leq |X|$, the last claim (iv) follows by linearity and monotonicity. Alternatively, one may apply the conditional Jensen's inequality which is stated and proven below. □

Proposition 210 (Conditional Jensen's inequality)**.** *Assume* X *is integrable and let* ϕ *be a convex function. Then*

$$\phi(\mathbb{E}\,(X|\mathcal{G})) \leq \mathbb{E}(\phi(X)|\mathcal{G}).$$

a.s.

Proof. For $x \in \mathbb{R}$, let $E_\phi := \{(a,b) \in \mathbb{Q}^2 : \phi(x) \geq ax + b\}$. Observe that by convexity and the density of $\mathbb{Q}^2$ in $\mathbb{R}^2$, we have $\phi(x) = \sup_{(a,b)\in E_\phi}(ax+b)$. For any $(a,b) \in E_\phi$, we then have $\mathbb{E}(\phi(X)|\mathcal{G}) \geq \mathbb{E}(aX + b|\mathcal{G}) = a\mathbb{E}(X|\mathcal{G}) + b$ where we used the monotonicity and the linearity of the conditional expectation (in this order). Taking the supremum over $(a,b) \in E_\phi$ then yields the claim. □

Finally, let us state analogies of the usual limit theorems for the conditional expectation.

Proposition 211 (Limit theorems for conditional expectation)**.** *Let* $X, Z,$ *and* X_n*,* $n = 1, 2, \ldots,$ *be random variables on a probability space* $(\Omega, \mathcal{F}, \mathbb{P})$ *and let* $\mathcal{G} \subset \mathcal{F}$ *be a sub-*σ*-algebra.*

(i) If $X_n \geq 0$ *and* $X_n \nearrow X$ *as* $n \to \infty$ *a.s., then* $\mathbb{E}(X_n|\mathcal{G}) \nearrow \mathbb{E}(X|\mathcal{G})$ *a.s. (Conditional monotone convergence theorem)*

(ii) If $X_n \geq 0$*, then* $\mathbb{E}(\liminf_{n\to\infty} X_n|\mathcal{G}) \leq \liminf_{n\to\infty}\mathbb{E}(X_n|\mathcal{G})$*. (Conditional Fatou's lemma)*

(iii) Assume that $X_n \xrightarrow[n\to\infty]{} X$ *and* $|X_n| \leq Z$ *a.s. for all* $n \geq 1$ *with* $\mathbb{E}(Z) < \infty$*. Then* $\mathbb{E}(X_n|\mathcal{G}) \xrightarrow[n\to\infty]{} \mathbb{E}(X|\mathcal{G})$ *a.s. and in* L^1*. (Conditional dominated convergence theorem)*

Proof. We prove (i), and leave $(ii)-(iii)$ as an exercise. Thus, assume $X_n \geq 0$ and $X_n \nearrow X$ as $n \to \infty$ a.s. Let $Y_n := X - X_n$. Then by assumption, $Y_n \xrightarrow[n\to\infty]{a.s.} 0$ and the sequence $(Y_n)_{n=1}^{\infty}$ is decreasing. By monotonicity of the conditional expectation, the sequence $(\mathbb{E}(Y_n|\mathcal{G}))_{n=1}^{\infty}$ is then decreasing as well. Thus, there exists a random variable $Z := \lim_{n\to\infty} \mathbb{E}(Y_n|\mathcal{G}) \geq 0$ a.s.

Assume first that X is integrable. Then $\mathbb{E}(Y_n|\mathcal{G}) \leq \mathbb{E}(X|\mathcal{G})$ for all $n = 1, 2, \ldots$, where $\mathbb{E}(X|\mathcal{G})$ is integrable. Hence by the usual DCT, we find $\mathbb{E}(Y_n) = \mathbb{E}(\mathbb{E}(Y_n|\mathcal{G})) \xrightarrow[n\to\infty]{} \mathbb{E}(Z)$. On the other hand, again by the DCT, we have $\mathbb{E}(Y_n) \xrightarrow[n\to\infty]{} 0$ since $|Y_n| \leq X \in L^1$. Hence we must have $\mathbb{E}(Z) = 0$. Since $Z \geq 0$ a.s., this implies $Z = 0$ a.s. Therefore by linearity, $\mathbb{E}(X|\mathcal{G}) - \mathbb{E}(X_n|\mathcal{G}) \xrightarrow[n\to\infty]{} 0$ a.s., and the final claim is concluded by monotonicity.

Assume finally that $X \geq 0$ is not necessarily integrable. Then for any $M > 0$, we have

$$\begin{aligned}|\mathbb{E}(X|\mathcal{G}) - \mathbb{E}(X_n|\mathcal{G})| \leq{}& |\mathbb{E}(X|\mathcal{G}) - \mathbb{E}(X \wedge M|\mathcal{G})| + |\mathbb{E}(X \wedge M|\mathcal{G}) - \mathbb{E}(X_n \wedge M|\mathcal{G})| \\ &+ |\mathbb{E}(X_n \wedge M|\mathcal{G}) - \mathbb{E}(X_n|\mathcal{G})|\end{aligned}$$

where the first and the last term are small when M is large enough by the definition of the conditional expectation for non-negative random variables, whereas the middle term converges to zero as $n \to \infty$ by integrability. The claim follows. □

9.2 Martingales

Martingales are one of the most key concepts of probability theory. They originate most likely from gambling problems, where they model the expected profit in a fair game. Since then, they have found their applications in various fields such as mathematical finance and statistical physics. We only scratch the surface of the theory in the case of discrete time, and encourage the reader to pursue the study further in the case of stochastic processes and stochastic analysis. For us in this course, martingales illustrate above all the usefulness of conditional expectation, providing explicit and concrete examples of its use. We conclude the notes with the *optional stopping theorem*, which has proven its usefulness in the afore mentioned fields of mathematical finance and statistical mechanics.

Definition 212 (Filtration)**.** A sequence $(\mathcal{F}_n)_{n=0}^{\infty}$ of σ-algebras is called a *filtration* if $\mathcal{F}_n \subset \mathcal{F}_{n+1}$ for all $n \in \mathbb{N}$.

Remark 213. By default, we usually assume that we are given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ such that $\mathcal{F}_n \subset \mathcal{F}$ for all $n \in \mathbb{N}$. Then the interpretation of a filtration is that it is a refinement of the total information $\mathcal{F}$ along time $n = 0, 1, 2, \ldots$. As time passes, the information known about the random system also increases.

Example 214. If $(X_n)_{n=0}^{\infty}$ is a sequence of random variables, then $\mathcal{F}_n = \sigma(X_0, \ldots, X_n)$ defines a filtration. This is perhaps the most common choice for a filtration, and indeed $\mathcal{F}_n$ is the smallest $\sigma-$algebra such that the random variables $X_0, \ldots, X_n$ are measurable with respect to it.

Definition 215 (Martingale)**.** Let $(X_n)_{n=0}^{\infty}$ be a sequence of random variables and $(\mathcal{F}_n)_{n=0}^{\infty}$ a filtration. Then $(X_n)_{n=0}^{\infty}$ is a *martingale* with respect to the filtration $(\mathcal{F}_n)_{n=0}^{\infty}$ if the following hold.

(i) $(X_n)_{n=0}^{\infty}$ is *adapted* to the filtration $(\mathcal{F}_n)_{n=0}^{\infty}$. That is, X_n is measurable w.r.t. $\mathcal{F}_n$ for all $n \in \mathbb{N}$.

(ii) X_n are integrable: $\mathbb{E}(|X_n|) < \infty$ for all $n \in \mathbb{N}$.

(iii) $\mathbb{E}(X_{n+1}|\mathcal{F}_n) = X_n$ for all $n \in \mathbb{N}$.

Definition 216 (Sub- and supermartingale)**.** If we replace *(iii)* by

- $\mathbb{E}(X_{n+1}|\mathcal{F}_n) \geq X_n$ for all $n \in \mathbb{N}$, we call $(X_n)_{n=0}^{\infty}$ a *submartingale*
- $\mathbb{E}(X_{n+1}|\mathcal{F}_n) \leq X_n$ for all $n \in \mathbb{N}$, we call $(X_n)_{n=0}^{\infty}$ a *supermartingale.*

All the other assumptions stay the same.

The concept of sub- and supermartingale can be viewed as how the expected gain behaves in a game, or alternatively, how fair the game is. If the process $(X_n)_{n=0}^{\infty}$ is a submartingale, then the expected payoff to the player is larger or equal than the current payoff as the game proceeds with new rounds. In popular forms of gamblings, for example lotteries, the game is designed to be a supermartingale, so that it remains profitable to the organizer. Martingale is then the boundary case between these two regimes, representing a *fair game.* The following example gives some illustration to this.

Example 217. Let $X_1, X_2, \dots$ be i.i.d. integrable random variables. The random variable X_n can be viewed as the gain (or the loss) in the n:th round of some game. We consider $S_n := X_1 + \cdots + X_n$ which represents the total gain in n rounds of the game, i.e. the *cumulative fortune* by time n. Let $\mathcal{F}_n := \sigma(X_1, \dots, X_n) = \sigma(S_1, \dots, S_n)$. We immediately see that S_n is measurable with respect to $\mathcal{F}_n$, and the integrability follows by the triangle inequality since $\mathbb{E}(|X_1|) < \infty$. By the fact that S_n is $\mathcal{F}_n$-measurable and X_{n+1} is independent of $\mathcal{F}_n$, we then derive the following *fairness condition* using the properties of the conditional expectation:

$$\mathbb{E}(S_{n+1}|\mathcal{F}_n) = \mathbb{E}(S_n|\mathcal{F}_n) + \mathbb{E}(X_{n+1}|\mathcal{F}_n) = S_n + \mathbb{E}(X_{n+1}) \quad \begin{cases} \leq S_n & \text{if } \mathbb{E}(X_{n+1}) = \mathbb{E}(X_1) \leq 0 \\ = S_n & \text{if } \mathbb{E}(X_{n+1}) = \mathbb{E}(X_1) = 0 \\ \geq S_n & \text{if } \mathbb{E}(X_{n+1}) = \mathbb{E}(X_1) \geq 0. \end{cases}$$

That is, $(S_n)_{n=0}^{\infty}$ is a submartingale if $\mathbb{E}(X_1) \leq 0$, a supermartingale if $\mathbb{E}(X_1) \geq 0$ and a martingale if $\mathbb{E}(X_1) = 0$.

Example 218. Let us consider the setting of the previous example, with the extra assumption that $\mathbb{E}(X_1) = 0$ and $\mathbb{E}(X_1^2) = \sigma^2 \in (0, \infty)$. Let $M_n := S_n^2 - n\sigma^2$ and $\mathcal{F}_n := \sigma(X_1, \dots, X_n)$. Then clearly M_n is measurable w.r.t. $\mathcal{F}_n$. Moreover, we have $\mathbb{E}(|M_n|) = \mathbb{E}(|S_n^2 - n\sigma^2|) \leq \mathbb{E}(S_n^2) + n\sigma^2 < \infty$ by the second moment assumption $\mathbb{E}(X_1^2) < \infty$ and independence (by expanding the square S_n^2). Finally, the fairness condition becomes

$$\begin{aligned} \mathbb{E}(M_{n+1}|\mathcal{F}_n) &= \mathbb{E}((S_n + X_{n+1})^2|\mathcal{F}_n) - (n+1)\sigma^2 = \mathbb{E}(S_n^2 + 2S_nX_{n+1} + X_{n+1}^2|\mathcal{F}_n) - (n+1)\sigma^2 \\ &= \mathbb{E}(S_n^2|\mathcal{F}_n) + \mathbb{E}(2S_nX_{n+1}|\mathcal{F}_n) + \mathbb{E}(X_{n+1}^2|\mathcal{F}_n) - (n+1)\sigma^2 \\ &= S_n^2 + 2S_n\mathbb{E}(X_{n+1}|\mathcal{F}_n) + \mathbb{E}(X_{n+1}^2) - (n+1)\sigma^2 \\ &= S_n^2 + 2S_n\mathbb{E}(X_{n+1}) + \mathbb{E}(X_1^2) - (n+1)\sigma^2 = S_n^2 - n\sigma^2 = M_n. \end{aligned}$$

Above we used the facts that S_n^2, $2S_n$, X_{n+1} and $2S_nX_{n+1}$ are integrable and measurable with respect to $\mathcal{F}_n$, as well as X_{n+1} and X_{n+1}^2 are independent of $\mathcal{F}_n$, together with linearity, the moment assumptions and the fact that the random variables X_i are i.i.d. We have hence shown that $(M_n)_{n=1}^\infty$ is a martingale.

Remark 219. If $(S_n)_{n=0}^\infty$ is a martingale, then there always exists a sequence of random variables $(Q_n)_{n=0}^\infty$ such that $(S_n^2 - Q_n)_{n=0}^\infty$ is a martingale. The sequence $(Q_n)_{n=0}^\infty$ is called the *quadratic variation* of $(S_n)_{n=0}^\infty$. We do not pursue this further in this course, but quadratic variation plays an important role in stochastic analysis (in particular in continuous time).

Example 220. Let Z be an integrable random variable (with respect to some probability space $(\Omega, \mathcal{F}, \mathbb{P})$) and let $(\mathcal{F}_n)_{n=0}^\infty$ be a filtration (such that $\mathcal{F}_n \subset \mathcal{F}$ for all $n \in \mathbb{N}$). Define $M_n := \mathbb{E}(Z|\mathcal{F}_n)$. Then by the definition of the conditional expectation, $(M_n)_{n=0}^\infty$ is adapted to the filtration $(F_n)_{n=0}^\infty$, and moreover M_n is integrable by Theorem 205. Furthermore, the tower property of the conditional expectation yields $\mathbb{E}(M_{n+1}|\mathcal{F}_n) = \mathbb{E}(\mathbb{E}(Z|\mathcal{F}_{n+1})|\mathcal{F}_n) = \mathbb{E}(Z|\mathcal{F}_n) = M_n$. Hence, $(M_n)_{n=0}^\infty$ is a martingale with respect to the filtration $(F_n)_{n=0}^\infty$.

The following proposition tells us that the expected gain in a martingale always stays the same, regardless how many time units one proceeds.

Proposition 221. *Let $(M_n)_{n=0}^\infty$ be a martingale with respect to a filtration $(\mathcal{F}_n)_{n=0}^\infty$. Then the following hold.*

(i) For all $n \in \mathbb{N}$, we have $\mathbb{E}(M_n) = \mathbb{E}(M_0)$.

(ii) For all $n, k \in \mathbb{N}$, we have $\mathbb{E}(M_{n+k}|\mathcal{F}_n) = M_n$.

Proof. Since $\mathcal{F}_n \subset \mathcal{F}_{n+1} \subset \cdots \subset \mathcal{F}_{n+k}$ for all $n, k \in \mathbb{N}$, the definition of the conditional expectation with the martingale assumption yield $\mathbb{E}(M_{n+k}\mathbb{1}_A) = \mathbb{E}(M_{n+k-1}\mathbb{1}_A) = \cdots = \mathbb{E}(M_n\mathbb{1}_A)$ for all $A \in \mathcal{F}_n$. Choosing $A = \Omega$ and $n = 0$ gives claim (i). The claim (ii) follows by the definition of the conditional expectation with respect to $\mathcal{F}_n$ for any $n \in \mathbb{N}$. □

Remark 222. If $(M_n)_{n=0}^\infty$ a martingale with respect to some filtration $(\mathcal{F}_n)_{n=0}^\infty$, then it is always a martingale w.r.t. the filtration $(\sigma(M_0, \ldots, M_n))_{n\geq 0}$. That is, $(\sigma(M_0, \ldots, M_n))_{n\geq 0} \subset \mathcal{F}_n$ for all $n \in \mathbb{N}$. For this reason, one often just says by convention that $(M_n)_{n=0}^\infty$ is a martingale if it is a martingale with respect to this canonical filtration.

9.2.1 Martingale transform

In this brief subsection, we rephrase first the definition of a martingale $(M_n)_{n=0}^\infty$ in terms of its increments $\Delta_n := M_n - M_{n-1}$. Then we notice that this representation can be generalized by reweighting the coefficients of the corresponding telescopic sum by certain random variables, yielding a way to generate a new martingale from a given martingale as a transform.

Lemma 223. *Let $(M_n)_{n=0}^\infty$ be a discrete-time stochastic process, i.e. a sequence of random variables, and let $(\mathcal{F}_n)_{n=0}^\infty$ be a filtration. Define $\Delta_0 := M_0$ and $\Delta_n := M_n - M_{n-1}$ for all $n \geq 1$. Then $(M_n)_{n=0}^\infty$ is a martingale if and only if all of the following hold:*

(i) $(\Delta_n)_{n=0}^\infty$ is adapted to $(\mathcal{F}_n)_{n=0}^\infty$, i.e. Δ_n is $\mathcal{F}_n$ measurable for all $n \in \mathbb{N}$.

(ii) Δ_n is integrable for all $n \in \mathbb{N}$.

(iii) $\mathbb{E}(\Delta_{n+1}|\mathcal{F}_n) = 0$ *for all* $n \in \mathbb{N}$ *("fairness condition").*

Proof. For all $n \in \mathbb{N}$, we write $M_n = \sum_{k=0}^{n} \Delta_k$. From this, it is easy to see that M_n is integrable if and only if Δ_k is integrable for all $k \leq n$, and M_n is $\mathcal{F}_n$ measurable if and only if Δ_k is $\mathcal{F}_k$ (and hence $\mathcal{F}_n$) measurable for all $k \leq n$. Since $M_{n+1} = \Delta_{n+1} + \sum_{k=0}^{n} \Delta_k = \Delta_{n+1} + M_n$, the fairness condition (iii) is obviously equivalent with the martingale condition $\mathbb{E}(M_{n+1}|\mathcal{F}_n)$. □

Next, we would like to reweight the above telescopic sum by *predictable* random variables. In gambling, these would correspond the varying stakes of a player for each round in a game.

Definition 224 (Predictable process)**.** A discrete-time stochastic process $(W_n)_{n\geq 1}$ is *predictable* if W_n is $\mathcal{F}_{n-1}$-measurable for all $n \geq 1$.

Definition 225 (Martingale transform)**.** Let $(M_n)_{n=0}^{\infty}$ be a martingale and $(W_n)_{n\geq 1}$ be a predictable process with respect to the same filtration $(F_n)_{n=0}^{\infty}$. Denote $\Delta_0 := M_0$ and $\Delta_n := M_n - M_{n-1}$ for all $n \geq 1$. Then the *martingale transform* of $(M_n)_{n=0}^{\infty}$ by $(W_n)_{n\geq 1}$ is the process

$$(W \cdot M)_n := M_0 + \sum_{k=1}^{n} W_k \Delta_k.$$

Remark 226. The martingale transform is the discrete version of the *stochastic integral* $\int_0^s W_t dM_t$, which is a central object in the continuum theory of *stochastic analysis* and which falls out of the scope of this course. The reader is very much encouraged to pursue in the continuum theory after reading these notes, eg. by following a course on stochastic analysis.

The next proposition tells intuitively that one cannot alter a fair game by simply varying the stakes, as long as the stakes stay bounded. In other words, the martingale transform preserves the martingale structure if the reweighting by the predictable process is bounded.

Proposition 227. *Let* $(M_n)_{n=0}^{\infty}$ *and* $(W_n)_{n\geq 1}$ *be as in Definition 225. Assume furthermore that* $|W_n| \leq C$ *for some* $C \in (0, \infty)$ *and all* $n = 1, 2, \ldots$. *Then* $((W \cdot M)_n)_{n=0}^{\infty}$ *is a martingale.*

Proof. We check the conditions of Lemma 223. For $n \geq 1$, denote $\Delta'_n := W_n \Delta_n$, where $\Delta_n = M_n - M_{n-1}$. Since W_n is $\mathcal{F}_{n-1}$ measurable, it is also $\mathcal{F}_n$ measurable, as Δ_n is. Thus, Δ'_n is $\mathcal{F}_n$ measurable. Moreover, since $|\Delta'_n| = |W_n| \, |\Delta_n| \leq C \, |\Delta_n|$, where Δ_n is integrable (by Lemma 223), also $|\Delta'_n|$ is integrable. Finally, $\mathbb{E}(\Delta'_{n+1}|\mathcal{F}_n) = \mathbb{E}(W_{n+1}\Delta_{n+1}|\mathcal{F}_n) = W_{n+1}\mathbb{E}(\Delta_{n+1}|\mathcal{F}_n) = 0$ where we used the assumption that W_{n+1} is $\mathcal{F}_n$ measurable. Hence, the claim follows from Lemma 223. □

9.2.2 Optional stopping

Intuitively, optional stopping is about quitting the game at the right moment. When the game is fair, i.e. a martingale, the "right moment" is given by some random time. In this section, we provide the classical form of optional stopping theorem for martingales. Throughout this section, we are given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a filtration $(\mathcal{F}_n)_{n=0}^{\infty}$ such that $\mathcal{F}_n \subset \mathcal{F}$.

Definition 228 (Stopping time)**.** Let $\tau : \Omega \to \mathbb{N} \cup \{\infty\}$ be a random variable (possibly with $\mathbb{P}(\tau = \infty) > 0$). Then τ is a *stopping time* with respect to the filtration $(\mathcal{F}_n)_{n=0}^{\infty}$ if $\{\tau \leq n\} \in \mathcal{F}_n$ for all $n \in \mathbb{N}$. Equivalently, τ is a *stopping time* w.r.t. $(\mathcal{F}_n)_{n=0}^{\infty}$ if $\{\tau = n\} \in \mathcal{F}_n$ for all $n \in \mathbb{N}$.

Remark 229. Intuitively, the definition of the stopping time reads: a random time τ is a stopping time if and only if at time n and given the information $\mathcal{F}_n$, we can determine whether the time τ has passed or not.

Example 230 (First hitting time)**.** Let $(X_n)_{n=0}^{\infty}$ be a sequence of random variables adapted to a filtration $(\mathcal{F}_n)_{n=0}^{\infty}$. Let $A \in \mathcal{B}$, and define $\tau := \inf\{n \geq 0 : X_n \in A\}$. This is called the *first hitting time* of A by X_n. It is not hard to see that τ is a stopping time. Indeed, given the information $\mathcal{F}_n$, one knows whether X_i has already visited A for some $i = 1, \dots, n$ or not. Turning this into a formal proof is left as an extra exercise to the reader.

Note that $\sup\{n \geq 0 : X_n \in A\}$ is *not* a stopping time, since there is no way to determine whether a visit to A was the last one at any given finite time n.

Proposition 231. *Let τ_1 and τ_2 be two stopping times w.r.t the same filtration $(\mathcal{F}_n)_{n=0}^{\infty}$. Then $\tau_1 \wedge \tau_2 := \min\{\tau_1, \tau_2\}$ and $\tau_1 \vee \tau_2 := \max\{\tau_1, \tau_2\}$ are stopping times w.r.t $(\mathcal{F}_n)_{n=0}^{\infty}$. In particular, if $N \in \mathbb{N}$ is fixed, then $\tau \wedge N$ is a bounded stopping time.*

Proof. We write $\{\tau_1 \wedge \tau_2 \leq n\} = \{\tau_1 \leq n \text{ or } \tau_2 \leq n\} = \{\tau_1 \leq n\} \cup \{\tau_2 \leq n\} \in \mathcal{F}_n$. Similarly, $\{\tau_1 \vee \tau_2 \leq n\} = \{\tau_1 \leq n \text{ and } \tau_2 \leq n\} = \{\tau_1 \leq n\} \cap \{\tau_2 \leq n\} \in \mathcal{F}_n$. □

The following proposition forms the basis for the optional stopping theorem.

Proposition 232. *Let $(M_n)_{n=0}^{\infty}$ be a martingale with respect to a filtration $(\mathcal{F}_n)_{n=0}^{\infty}$ and let τ be a stopping time with respect to $(\mathcal{F}_n)_{n=0}^{\infty}$. Consider the process $(M_{n\wedge\tau})_{n\geq 0}$ defined as*

$$M_{n\wedge\tau} = \begin{cases} M_n & \text{if } n \leq \tau \\ M_\tau & \text{if } n > \tau \end{cases} .$$

Then $(M_{n\wedge\tau})_{n\geq 0}$ is a martingale, which we call the stopped martingale.

The claim also holds in the form where we replace the martingales by submartingales.

Proof. We define $W_n := \mathbb{1}_{\tau \geq n}$ for all $n \in \mathbb{N}$. Since $\{\tau \geq n\} = \{\tau \leq n-1\}^c \in \mathcal{F}_{n-1}$, the process $(W_n)_{n=0}^{\infty}$ is predictable. We also have $|W_n| \leq 1$ for all $n \in \mathbb{N}$. Thus by Proposition 227, $((W \cdot M)_n)_{n=0}^{\infty}$ is a martingale. We find

$$(W \cdot M)_n = M_0 + \sum_{k=1}^{n} W_k \Delta_k = M_0 + \sum_{k=1}^{n} \mathbb{1}_{\tau \geq k} \Delta_k = M_0 + \sum_{k=1}^{\tau\wedge n} (M_k - M_{k-1}) = M_{\tau\wedge n}$$

showing the claim for martingales.

The claim for submartingales follows by noting that $W_n \geq 0$, and the proof of Proposition 227 still applies by replacing some of the equalities by inequalities. We leave this extension to the reader. □

Theorem 233 (The optional stopping theorem)**.** *Let $(M_n)_{n=0}^{\infty}$ be a martingale with respect to a filtration $(\mathcal{F}_n)_{n=0}^{\infty}$ and let τ be a stopping time with respect to $(\mathcal{F}_n)_{n=0}^{\infty}$. Assume that either* *(i)* $\tau \leq C$ *a.s. for some* $C < \infty$*, or*

(ii) $\tau < \infty$ *a.s. and* $(M_{n\wedge\tau})_{n\geq 0}$ *is uniformly integrable.*

Then $\mathbb{E}(M_\tau) = \mathbb{E}(M_0)$*.*

If $(M_n)_{n=0}^\infty$ *is a submartingale and the other assumptions remain the same, then the claim holds in the form* $\mathbb{E}(M_\tau) \geq \mathbb{E}(M_0)$

Proof. Assume first (i) holds. Without loss of generality, we may assume $\tau \leq N$ a.s. for some $N \in \mathbb{N}$. Since $(M_{n\wedge\tau})_{n\geq 0}$ is a martingale by the previous proposition, we have $\mathbb{E}(M_{n\wedge\tau}) = \mathbb{E}(M_{0\wedge\tau}) = \mathbb{E}(M_0)$. Substituting $n = N$ then yields $\mathbb{E}(M_\tau) = \mathbb{E}(M_{N\wedge\tau}) = \mathbb{E}(M_0)$.

Then, assume (ii) holds. For any $n \in \mathbb{N}$ fixed, $\tau \wedge n$ is a bounded stopping time, and thus $\mathbb{E}(M_{n\wedge\tau}) = \mathbb{E}(M_0)$. But since $\tau < \infty$ a.s., we have $n \wedge \tau \to \tau$ a.s. as $n \to \infty$, and further $M_{n\wedge\tau} \xrightarrow[n\to\infty]{a.s.} M_\tau$. But since $(M_{n\wedge\tau})_{n\geq 0}$ is uniformly integrable, the almost sure convergence implies convergence in L^1 (via convergence in probability). Hence, $|\mathbb{E}(M_{n\wedge\tau}) - \mathbb{E}(M_\tau)| \leq \mathbb{E}(|M_{n\wedge\tau} - M_\tau|) \xrightarrow[n\to\infty]{} 0$. This shows $\mathbb{E}(M_\tau) = \mathbb{E}(M_0)$. □

Let us finally look at some examples.

Example 234 (Simple random walk on $\mathbb{Z}$)**.** We consider i.i.d. random variables $X_1, X_2, \ldots$ taking values in $\{-1, +1\}$ with probability $\mathbb{P}(X_1 = \pm 1) = \frac{1}{2}$. Let $S_n := X_1 + \cdots + X_n$ for $n = 1, 2, \ldots$ and $S_0 = 0$. Then $(S_n)_{n=0}^\infty$ is the *simple random walk* on $\mathbb{Z}$. Recall that by Example 217 it is a martingale, since $\mathbb{E}(X_1) = 0$.

Fix integers $a < 0 < b$ and define the hitting times $\tau_a := \inf\{n \in \mathbb{N} : S_n = a\}$ and $\tau_b := \inf\{n \in \mathbb{N} : S_n = b\}$. By Example 230, these are both stopping times w.r.t. the filtration $(\mathcal{F}_n)_{n=0}^\infty$ given by $\mathcal{F}_n := \sigma(X_1, \ldots, X_n)$. Then by Proposition 231, the minimum $\tau := \tau_a \wedge \tau_b$ is again a stopping time.

We want to compute the probability $\mathbb{P}(\tau_a < \tau_b)$. Let us apply the optional stopping theorem as follows. Observe that the random walk exits the interval $[a+1, b-1]$ for sure if its steps are $+1$ at least $b - a$ time units in a row. To use this observation, we see that $\mathbb{P}(X_1 = 1, \ldots, X_{b-a} = 1) = \mathbb{P}(X_1 = 1)^{b-a} = 2^{-(b-a)}$ since $X_1, X_2, \ldots$ are i.i.d. In general, $\mathbb{P}(X_{N(b-a)+1} = 1, \ldots, X_{(N+1)(b-a)} = 1) = 2^{-(b-a)}$ for all $N \in \mathbb{N}$, where the events $\{X_{N(b-a)+1} = 1, \ldots, X_{(N+1)(b-a)} = 1\}$ are independent when $N = 0, 1, \ldots$. Therefore, since $\sum_{N=0}^\infty \mathbb{P}(X_{N(b-a)+1} = 1, \ldots, X_{(N+1)(b-a)} = 1) = \sum_{N=0}^\infty 2^{-(b-a)} = \infty$, the second Borel-Cantelli lemma implies that the event $\{X_{N(b-a)+1} = 1, \ldots, X_{(N+1)(b-a)} = 1\}$ happens for infinitely many $N \in \mathbb{N}$ almost surely. Therefore, we must have $\tau < \infty$ a.s.

Furthermore, we observe $|S_{n\wedge\tau}| \leq \max\{-a, b\} < \infty$, so $(S_{n\wedge\tau})_{n\geq 0}$ is UI (by Proposition 189). Hence by the optional stopping theorem, $\mathbb{E}(S_\tau) = \mathbb{E}(S_0) = 0$. On the other hand, we have $\mathbb{E}(S_\tau) = a\mathbb{P}(\tau_a < \tau_b) + b\mathbb{P}(\tau_a \geq \tau_b) = b + (a-b)\mathbb{P}(\tau_a < \tau_b)$. Solving this gives $\mathbb{P}(\tau_a < \tau_b) = \frac{b}{b-a}$.

Example 235 (Expectation of a stopping time)**.** We continue from the previous example with the same setting. Now, let us study $\mathbb{E}(\tau)$. In order to approach this, recall from Example 218 that $M_n := S_n^2 - n$ defines a martingale (since here $\sigma^2 = \mathbb{E}(X_1^2) = 1$). Now by Proposition 232, $(M_{n\wedge\tau})_{n\geq 0}$ is a also martingale. Hence by Proposition 221, $\mathbb{E}(M_{n\wedge\tau}) = \mathbb{E}(M_0) = 0$. On the other hand, $\mathbb{E}(M_{n\wedge\tau}) = \mathbb{E}(S_{n\wedge\tau}^2) - \mathbb{E}(n\wedge\tau)$, therefore $\mathbb{E}(n\wedge\tau) = \mathbb{E}(S_{n\wedge\tau}^2)$. But since $n \wedge \tau \nearrow \tau$ as $n \to \infty$, it follows by the monotone convergence theorem that

$\mathbb{E}(n \wedge \tau) \xrightarrow[n\to\infty]{} \mathbb{E}(\tau)$. Moreover, since $S_{n\wedge\tau}^2 \leq \max\{-a, b\}^2 < \infty$ and $S_{n\wedge\tau} \xrightarrow[n\to\infty]{a.s.} S_\tau$, the dominated convergence yields

$$\mathbb{E}(S_{n\wedge\tau}^2) \xrightarrow[n\to\infty]{} \mathbb{E}(S_\tau^2) = a^2\mathbb{P}(\tau_a < \tau_b) + b^2(1 - \mathbb{P}(\tau_a < \tau_b)) = a^2\frac{b}{b-a} - b^2\frac{a}{b-a} = -ab.$$

Bibliography

[1] H. Duminil-Copin. Introduction to Bernoulli percolation. Lecture notes, https://www.unige.ch/~duminil/publi/2017percolation.pdf, October 2018.

[2] R. Durrett. *Probability: theory and examples*, volume 31 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, fifth edition, 2019.

[3] G. Grimmett and D. Stirzaker. *Probability and Random Processes*. OUP, 3rd edn edition, 2001.

[4] J.-F. Le Gall. Intégration, probabilités et processus aléatoires. Lecture notes, https://www.imo.universite-paris-saclay.fr/~jean-francois.le-gall/IPPA2.pdf, September 2006.

[5] G. Shafer and V. Vovk. The Sources of Kolmogorov's Grundbegriffe. *Statistical Science*, 21(1):70–98, 2006.