Lecture notes

Advanced probability theory

Mathias Beiglböck and Yizheng Yuan

Building on earlier notes of Joonas Turunen

Greatly inspired by earlier notes of Nathanaël Berestycki and Alexander Glazman

University of Vienna

Faculty of Mathematics

Contents

0 Introduction **3**
- 0.1 Motivation . . . 3
- 0.2 Very brief history . . . 4

1 Basic notions of measure-theoretic probability **5**
- 1.1 Measure theory recap . . . 5
- 1.2 Probability and measure . . . 8
- 1.3 Random variables . . . 10
- 1.4 Expectation . . . 13
- 1.5 Variance and covariance . . . 16

2 Independence **18**
- 2.1 Sufficient conditions for independence . . . 20
- 2.2 Characterizations for independence . . . 22

3 Sequences of events and limit theorems **29**
- 3.1 Borel-Cantelli lemmas . . . 29
- 3.2 Kolmogorov's zero-one law . . . 32
- 3.3 Infinite product spaces and sequences of independent random variables . . . 33
- 3.4 Convergence in probability and almost sure convergence . . . 34
- 3.5 Laws of large numbers . . . 36
  - 3.5.1 Weak law of large numbers . . . 36
  - 3.5.2 Strong law of large numbers . . . 37
- 3.6 Uniform integrability . . . 40

4 Martingales **46**
- 4.1 Conditional expectation and basic definitions . . . 46
- 4.2 Martingales as fair games and optional stopping . . . 49
- 4.3 Limit theorems for martingales . . . 52
- 4.4 A martingale proof of the strong law of large numbers . . . 55

5 Weak convergence **56**
- 5.1 Weak convergence on Polish spaces . . . 56
- 5.2 Other modes of convergence on metric spaces . . . 60
- 5.3 Convergence in distribution of random variables . . . 61
- 5.4 p-Wasserstein distance on a Hilbert space . . . 67

A Generalities on Polish spaces **72**
A.1 Some basic facts on Polish spaces 72

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 inaccuracies, mistakes and 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 [2, 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 [3] 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 1.1. 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 1.2. It is very useful to notice the following properties of a measure μ:

- $A, B \in \mathcal{A}$, $A \subset B \implies \mu(A) \leq \mu(B)$ (monotonicity)
- $A_i \in \mathcal{A}$, $i \in \mathbb{N} \implies \mu\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} \mu\left(A_i\right)$ (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 1.3 (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 1.4 (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 1.5 (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 1.6 (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 1.7. 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 1.8. 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 1.9 (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 1.10 (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 $p_1 \ldots, p_k$ are positive real numbers and $a_1, \ldots, a_k \in \Omega$, 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 1.11 (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 1.12 (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 1.13 (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 1.14 (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 1.15. *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 1.16. *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 1.17. 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 1.18. 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 1.19. 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 1.20. 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 1.21 (Coin tosses)**.** Let $\Omega = \{0,1\}^n$ (where $n = 1, 2, \dots$ 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 $\{\omega = (\omega_1, \dots, \omega_n) \in \Omega : \omega_1 = 0\}$.

Example 1.22 (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 1.23 (Poisson distribution)**.** Let $\Omega = \{0, 1, 2, \dots\}$, $\mathcal{F} = \mathcal{P}(\Omega)$ 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 1.24 (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 1.25. 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 $\mathbb{P}(\{\omega : X(\omega) \in B\})$, 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 1.26. 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 $a, b \in \mathbb{R}$, $a < b$:

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

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 1.27 (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

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

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

Remark 1.28. It is a straightforward exercise to show that the law μ is a probability measure on $\mathcal{S}$.

Remark 1.29. 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 1.30 (The first outcome of a series of coin tosses)**.** We recall our setting in Example 1.21: 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

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

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 1.31 (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 1.32 (Uniformly distributed random variable)**.** We return to the setting of Example 1.22. 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 1.33 (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 1.36 below). Next, we state and prove some key properties of the distribution function.

Proposition 1.34 (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 1.15 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 1.16 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 1.16 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 1.15, $F(x_n) \xrightarrow[n\to\infty]{} \mu((-\infty, x))$. 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 1.35 (Distribution function of a coin toss)**.** We continue the setting of Example 1.30: $\Omega = \{0,1\}^n$, $\mathcal{F} = \mathcal{P}(\Omega)$ and $\mathbb{P}(\omega) = \frac{1}{|\Omega|} = 2^{-n}$, and define

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

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

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

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 1.34 holds. That is, if a function satisfying the conditions of Proposition 1.34 is given, it is always a distribution function of some random variable.

Theorem 1.36. *Let $F : \mathbb{R} \to [0,1]$ be a function satisfying the conditions (i)-(iii) of Proposition 1.34. 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 $\varepsilon > 0$ such that $F(x + \varepsilon) < u$.

Then, we obtain $F^{-1}(u) = \sup\{y \in \mathbb{R} : F(y) < u\} \geq x + \varepsilon > x$. 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 2.13 or the exercises of the course). Thus it follows from Dynkin's lemma (Proposition 2.18 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 2.25. □

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

Definition 1.37. 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 1.38. 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 1.39 (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 1.40 (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 1.41 (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 1.42. *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 1.43. 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 1.44 (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 1.45 (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 1.46. 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 1.45 often makes computations of expectations of measurable transforms of random variables more straightforward, as the following simple example demonstrates.

Example 1.47. Let U be uniformly distributed on $[0, 1]$. Let us compute its expectation $\mathbb{E}(U^2)$. We recall from Example 1.39 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 1.45 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 1.48 (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 1.42).

Remark 1.49. 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 1.50 (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 1.51 (Chebyshev's inequality)**.** *Let X be a random variable with $\mathbb{E}(X^2) < \infty$. Then for any $a > 0$,*

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

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 $\mathbb{P}(|X - m| \geq a) = \mathbb{P}((X-m)^2 \geq a^2) \leq \frac{\mathbb{E}\left((X-m)^2\right)}{a^2}$. □

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

Definition 1.52 (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 $\mathrm{Cov}(X, Y) := \mathbb{E}\left((X - m_X)(Y - m_Y)\right)$.

Remark 1.53. In the spirit of Remark 1.49, we find a practical formula for the covariance: $\mathrm{Cov}(X, Y) := \mathbb{E}\left((X - m_X)(Y - m_Y)\right) = \mathbb{E}(XY) - m_X\mathbb{E}(Y) - m_Y\mathbb{E}(X) + m_X m_Y = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$. 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 1.54. The *correlation coefficient* of X and Y is defined by

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

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

- If $\mathrm{Cov}(X, Y) \geq 0$, then X and Y are *positively correlated*. That is, on average, $(X - m_X)(Y - m_Y) \geq 0$.
- If $\mathrm{Cov}(X, Y) \leq 0$, then X and Y are *negatively correlated*.
- If $\mathrm{Cov}(X, Y) = 0$, 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 2.1 (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 2.2. 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 2.3 (Memorylessness of an exponential random variable)**.** Let $X \sim \text{Exp}(\lambda)$, 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 $X - t \sim \text{Exp}(\lambda)$ 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 2.4 (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 2.5. 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 2.6 (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 $i_1, \ldots, i_k \in I$ and all $A_1 \in \mathcal{F}_{i_1}$, $A_2 \in \mathcal{F}_{i_2}, \ldots, A_k \in \mathcal{F}_{i_k}$, it holds

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

Remark 2.7 (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 $i_1, \ldots, i_k \in I$,

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

Remark 2.8. 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 $A_1, \ldots, A_n$, 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 2.9 (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 2.10 (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 2.6 and 2.10, we find an alternative definition for independence of random variables, which is usually more practical for applications.

Definition 2.11 (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 that the random variables X_i are *independent* if for all $k \in \mathbb{N}$, all distinct indices $i_1, \ldots, i_k \in I$ and all $B_1, \ldots, B_k \in \mathcal{B}$,

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

Example 2.12 (Independence of two coin tosses)**.** We continue our study of coin tosses as in Example 1.35 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 2.29.

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 2.13 (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 2.14. 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 2.15. 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 2.16 (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 2.17. 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 2.18 (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 2.19. *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 $A_i \in \mathcal{A}_i$, $(i = 1, \dots n)$,*

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

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

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

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

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

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

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 $\mathbb{P}(A^{(k)} \cap A_2 \cap \cdots \cap A_n) = \mathbb{P}(A^{(k)})\mathbb{P}(A_2) \dots \mathbb{P}(A_n)$. We can then apply Lemma 1.15 to both of the sides of this equation, which gives $\mathbb{P}\left(\bigcup_{k=1}^{\infty} A^{(k)} \cap A_2 \cap \cdots \cap A_n\right) = \mathbb{P}\left(\bigcup_{k=1}^{\infty} A^{(k)}\right) \mathbb{P}(A_2) \dots \mathbb{P}(A_n)$, 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 2.20. *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 2.19 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 2.21. 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 2.20 holds by definition if the sigma-algebras $\sigma(X_i)$ are independent.

2.2 Characterizations for independence

As mentioned in the previous remark, Corollary 2.20 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 2.22 (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 2.23 (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 2.24 (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 2.25. 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

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

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 1.9. 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 [1].

Theorem 2.26 (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 $\int_{\mathcal{S}} |f| \, d\mu < \infty$,*

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

Remark 2.27. 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 2.28 (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 2.29 (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 2.30 (Ratio of two independent exponential random variables)**.** We consider the setting in Example 2.3, this time with two independent exponential random variables X, Y with parameter $\lambda > 0$. In particular, both of the random variables have density $f(x) = \lambda e^{-\lambda x}\mathbb{1}_{x\geq 0}$. Let us compute the law of Y/X. By Theorem 2.28, 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 $F_{Y/X}(t) = \mathbb{P}(Y/X \leq t) = \left(1 - \frac{1}{1+t}\right)\mathbb{1}_{t>0} = \frac{t}{1+t}\mathbb{1}_{t>0}$, 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 1.45 and Definition 1.37.

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 2.31 (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 $\mathcal{P}_i := \{A_{i1} \cap \cdots \cap A_{im(i)} : A_{ij} \in \mathcal{F}_{ij} \quad \forall\, j = 1, \dots, m(i)\}$. 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 2.19 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 2.32. *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 2.33. 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 $X_1 + \cdots + X_{n-1}$ 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 2.34. *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 2.32, 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 2.32. The claim follows as before. □

Remark 2.35. 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 1.45. 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 2.36. *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 2.37 (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 1.23 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 2.34 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 2.38. *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 2.32, 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 2.34, 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 1.45) 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 2.39. 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 limit theorems

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 3.1 (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 3.2. 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 3.3. 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 3.4. 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 3.5 (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 $\mathbb{P}(A_n^c \cap \cdots \cap A_{n+k-1}^c) = \mathbb{P}(X_n = 0, \dots, X_{n+k-1} = 0) = \frac{1}{2^k}$, 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 3.6 (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 3.7 (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 1.16, 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 3.8. 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 3.9. 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 3.10. Let $X_1, X_2, \dots$ be a sequence of independent exponentially distributed random variables with parameter $\lambda > 0$. Recall from Example 2.3 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 $\varepsilon > 0$:

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

(ii) $X_n \geq (1-\varepsilon)\frac{\log(n)}{\lambda}$ 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, $\mathbb{P}\left(X_n > (1+\varepsilon)\frac{\log(n)}{\lambda}\right) = \exp\left(-\lambda(1+\varepsilon)\frac{\log(n)}{\lambda}\right) = n^{-(1+\varepsilon)}$ which is summable since $\varepsilon > 0$. Therefore by the 1st Borel-Cantelli, $X_n > (1+\varepsilon)\frac{\log(n)}{\lambda}$ only a finite number of times, which means that eventually $X_n \leq (1+\varepsilon)\frac{\log(n)}{\lambda}$, showing the claim (i). Then, $\mathbb{P}\left(X_n \geq (1-\varepsilon)\frac{\log(n)}{\lambda}\right) = \exp\left(-\lambda(1-\varepsilon)\frac{\log(n)}{\lambda}\right) = n^{-(1-\varepsilon)}$ for $0 < \varepsilon < 1$, which is not summable. Since the random variables $X_1, X_2, \dots$ are independent, so are the events $\left\{X_n \geq (1-\varepsilon)\frac{\log(n)}{\lambda}\right\}$. Hence by the 2nd Borel-Cantelli, $X_n \geq (1-\varepsilon)\frac{\log(n)}{\lambda}$ 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 3.11. 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 3.12. 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 3.13. 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 $\varepsilon > 0$ there exists an $N > 0$ such that for all integers $m, n > N$, $|X_n - X_m| < \varepsilon$. 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 3.14 (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 2.31. 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 2.19 (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 3.15. 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 2.22 and 2.24 the finite product σ-algebras and product measures. The generalization of these concepts to countable infinite setting is based on *cylinder sets*.

Definition 3.16 (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. [1].

Theorem 3.17 (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 2.11, the random variables X_i, $i = 1, 2, \dots$, are hence mutually independent.

Remark 3.18. 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 3.19. 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 $(X_{n,i})_{n \geq 1, i \geq 1}$ 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 $(U_n)_{n \geq 1}$ 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.

3.4 Convergence in probability and almost sure convergence

Definition 3.20 (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

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

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

Remark 3.21. The above definition means that for all $\varepsilon > 0$ and $\delta > 0$, there exist an index $N \in \mathbb{N}$ such that $\mathbb{P}(|X_n - X| \leq \varepsilon) > 1 - \delta$ whenever $n > N$. That is, with arbitrarily high probability, X_n is eventually close to X.

Definition 3.22 (Convergence a.s.)**.** Let us consider the setting of Definition 3.20. 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 3.23 (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,

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

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 3.24. *(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 $\varepsilon > 0$. Define events $E_m := \{|X_m - X| \leq \varepsilon\}$. 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

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

Now by Lemma 1.15, 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 $\varepsilon_k := \frac{1}{k}$ for all $k = 1, 2, \ldots$. By the assumption, $\mathbb{P}(|X_n - X| > \varepsilon_1) \xrightarrow[n\to\infty]{} 0$, and thus there exist a subsequence $(n_k^{(1)})_{k\geq 1}$ such that $\sum_{k=1}^{\infty} \mathbb{P}\left(\left|X_{n_k^{(1)}} - X\right| > \varepsilon_1\right) < \infty$. Now by the 1st Borel-Cantelli lemma, $\left|X_{n_k^{(1)}} - X\right| > \varepsilon_1$ can only occur finitely many times a.s., implying that $\left|X_{n_k^{(1)}} - X\right| \leq \varepsilon_1$ 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 $\left|X_{n_k^{(m+1)}} - X\right| \leq \varepsilon_{m+1}$ 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 3.25. *$X_n \xrightarrow[n\to\infty]{\mathbb{P}} X$ 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 $\varepsilon > 0$ and $\delta > 0$ and a subsequence $(n_k)_{k\geq 1}$ such that $\mathbb{P}(|X_{n_k} - X| > \varepsilon) \geq \delta$ 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, $\mathbb{P}(\left|X_{n_{k_m}} - X\right| > \varepsilon) \xrightarrow[m\to\infty]{} 0$, and hence the subsequence $(n_{k_m})_{m\geq 0}$ contradicts the assumption. □

Remark 3.26 (Convergence in probability is weaker than a.s. convergence)**.** Let $X_1, X_2, \ldots$ 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 $\varepsilon \in (0, 1)$, $\mathbb{P}(X_n > \varepsilon) = \frac{1}{n} \xrightarrow[n\to\infty]{} 0$, 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.

3.5 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 $X_1, X_2, \ldots$ 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).

3.5.1 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 3.27 (Weak law of large numbers)**.** *Let $X_1, X_2, \ldots$ 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 $\varepsilon > 0$. Observe that $\operatorname{Var}(S_n) = \sum_{i=1}^n \operatorname{Var}(X_i) = n \operatorname{Var}(X_1)$ 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

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

□

3.5.2 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 3.28. *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,

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

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

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

Now $\sum_{n=1}^\infty \mathbb{P}\left(\left|\frac{S_n}{n}\right| > \varepsilon\right) < \infty$, hence the first Borel-Cantelli lemma gives $\mathbb{P}(E_\varepsilon) = 1$, where $E_\varepsilon := \left\{\left|\frac{S_n}{n}\right| \le \varepsilon \text{ eventually}\right\}$. Therefore, if we take $\varepsilon = \frac{1}{k}$ 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 3.29. 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 $\mathbb{P}(|S_n| > n\varepsilon) \le \frac{\mathrm{Var}(X_1)}{\varepsilon^2}\frac{1}{n}$, 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: $\mathbb{P}(S_n^4 > (n\varepsilon)^4) \le \frac{C}{\varepsilon^4 n^2}$. 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 $\mathbb{P}(|S_n| > n\varepsilon) \le \frac{C}{n^k}$ 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 3.30. *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 $\mathbb{P}\left(\left|\frac{S_n}{n} - m\right| > \varepsilon\right) \leq \frac{\mathrm{Var}(X_1)}{\varepsilon^2}\frac{1}{n}$, which is not summable. However, if we consider the subsequence $(n_k)_{k\geq 0}$ defined by $n_k := k^2$, we have $\mathbb{P}\left(\left|\frac{S_{n_k}}{n_k} - m\right| > \varepsilon\right) \leq \frac{\mathrm{Var}(X_1)}{\varepsilon^2}\frac{1}{k^2}$ 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 3.31 (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 3.30 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 $Y_1, \ldots, Y_n$, 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 $i = 1, 2, \ldots$, hence we can even apply Chebyshev's inequality as follows:

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

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,

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

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

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

Therefore,

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

so the 1st Borel-Cantelli lemma implies that for all $\varepsilon > 0$, a.s. $\left|\frac{T_{n_k}}{n_k} - \frac{\mathbb{E}(T_{n_k})}{n_k}\right| \le \varepsilon$ 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:

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

Observe that since $\mathbb{E}(X_1\mathbb{1}_{X_1>j}) \le \mathbb{E}(X_1) < \infty$, 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 $\varepsilon > 0$ that there exists $N > 0$ large enough such that $a_j < \varepsilon$ 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 $\frac{1}{l}\sum_{j=N+1}^l a_j \le \frac{l-N}{l}\varepsilon$, which can be made arbitrarily small.) Combining this with the previous a.s. convergence, we conclude $\frac{T_{n_k}}{n_k} \xrightarrow[k\to\infty]{a.s.} \mathbb{E}(X_1) = m$.

□

3.6 Uniform integrability

Definition 3.32 (Uniform integrability)**.** Let $(X_i)_{i\in I}$ be a family of integrable random variables. We say that $(X_i)_{i\in I}$ is *uniformly integrable* if

$$\lim_{K\to\infty}\sup_{i\in I}\mathbb{E}\big(|X_i|\,\mathbb{1}_{|X_i|\ge K}\big) = 0. \tag{3.1}$$

Lemma 3.33 (Equivalent characterization)**.** *Let $(X_i)_{i\in I}$ be a family of integrable random variables. Then the following are equivalent:*

(i) $(X_i)_{i\in I}$ *is uniformly integrable.*

(ii) $\sup_{i\in I}\mathbb{E}(|X_i|) < \infty$ *and for every* $\varepsilon > 0$ *there exists* $\delta > 0$ *such that for all* $A \in \mathcal{F}$ *with* $\mathbb{P}(A) \le \delta$*,*

$$\sup_{i\in I}\mathbb{E}\big(|X_i|\,\mathbb{1}_A\big) \le \varepsilon. \tag{3.2}$$

Proof. $(i) \Rightarrow (ii)$: First,

$$\mathbb{E}(|X_i|) = \mathbb{E}(|X_i|\,\mathbb{1}_{|X_i|<K}) + \mathbb{E}(|X_i|\,\mathbb{1}_{|X_i|\ge K}) \le K + \sup_{i\in I}\mathbb{E}(|X_i|\,\mathbb{1}_{|X_i|\ge K}),$$

so taking K large enough gives $\sup_i \mathbb{E}(|X_i|) < \infty$.

Let $\varepsilon > 0$. Choose K such that $\sup_i \mathbb{E}(|X_i|\,\mathbb{1}_{|X_i|\ge K}) \le \varepsilon/2$. Then for any A,

$$\mathbb{E}(|X_i|\,\mathbb{1}_A) \le K\mathbb{P}(A) + \mathbb{E}(|X_i|\,\mathbb{1}_{|X_i|\ge K}),$$

hence

$$\sup_i \mathbb{E}(|X_i|\,\mathbb{1}_A) \le K\mathbb{P}(A) + \varepsilon/2.$$

Choosing $\delta = \varepsilon/(2K)$ yields the claim.

$(ii) \Rightarrow (i)$: Let $\varepsilon > 0$. Choose δ corresponding to $\varepsilon/2$ in (ii). Since $\sup_i \mathbb{E}(|X_i|) < \infty$, Markov's inequality gives

$$\sup_{i\in I}\mathbb{P}(|X_i| \ge K) \le \frac{1}{K}\sup_i \mathbb{E}(|X_i|) \to 0.$$

Hence for large K, $\sup_i \mathbb{P}(|X_i| \ge K) \le \delta$. Applying (ii) with $A = \{|X_i| \ge K\}$ gives

$$\sup_i \mathbb{E}(|X_i|\,\mathbb{1}_{|X_i|\ge K}) \le \varepsilon,$$

which proves uniform integrability. □

Example 3.34.

(i) If X is integrable, then $\{X\}$ is uniformly integrable, since

$$\mathbb{E}(|X|\,\mathbb{1}_{|X|\ge K}) \xrightarrow[K\to\infty]{} 0$$

by dominated convergence.

(ii) Any finite family of integrable random variables is uniformly integrable, since the supremum over a finite set is a maximum.

(iii) If $(X_i)_{i\in I}$ is dominated by an integrable random variable Y, i.e. $|X_i| \le Y$ a.s. for all i, then $(X_i)_{i\in I}$ is uniformly integrable. Indeed,

$$\sup_i \mathbb{E}(|X_i|\,\mathbb{1}_{|X_i|\ge K}) \le \mathbb{E}(Y\mathbb{1}_{Y\ge K}) \xrightarrow[K\to\infty]{} 0.$$

Corollary 3.35. *Let $(X_i)_{i\in I}$ and $(Y_i)_{i\in I}$ be uniformly integrable families. Then the family $(X_i + Y_i)_{i\in I}$ is uniformly integrable.*

Proof. By Lemma 3.33, we have

$$\sup_{i\in I} \mathbb{E}(|X_i|) < \infty, \qquad \sup_{i\in I} \mathbb{E}(|Y_i|) < \infty,$$

and for every $\varepsilon > 0$ there exists $\delta > 0$ such that for all $A \in \mathcal{F}$ with $\mathbb{P}(A) \le \delta$,

$$\sup_{i\in I} \mathbb{E}(|X_i|\, \mathbb{1}_A) \le \varepsilon, \qquad \sup_{i\in I} \mathbb{E}(|Y_i|\, \mathbb{1}_A) \le \varepsilon.$$

Hence

$$\sup_{i\in I} \mathbb{E}(|X_i + Y_i|) \le \sup_{i\in I} \mathbb{E}(|X_i|) + \sup_{i\in I} \mathbb{E}(|Y_i|) < \infty,$$

and

$$\sup_{i\in I} \mathbb{E}(|X_i + Y_i|\, \mathbb{1}_A) \le \sup_{i\in I} \mathbb{E}(|X_i|\, \mathbb{1}_A) + \sup_{i\in I} \mathbb{E}(|Y_i|\, \mathbb{1}_A) \le 2\varepsilon.$$

Thus $(X_i + Y_i)_{i\in I}$ is uniformly integrable by Lemma 3.33. □

Theorem 3.36 (de la Vallée-Poussin criterion)**.** *Let $(X_i)_{i\in I}$ be a family of integrable random variables. Then the following are equivalent:*

(i) *$(X_i)_{i\in I}$ is uniformly integrable.*

(ii) *There exists an increasing function $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ with $\varphi(x)/x \to \infty$ as $x \to \infty$ such that*

$$\sup_{i\in I} \mathbb{E}\big(\varphi(|X_i|)\big) < \infty.$$

Moreover, φ can be chosen convex.

Proof. $(ii) \Rightarrow (i)$: Let $\varepsilon > 0$. Since $\varphi(x)/x \to \infty$, there exists K such that $x \le \varepsilon\varphi(x)$ for all $x \ge K$. Then

$$\mathbb{E}(|X_i|\, \mathbb{1}_{|X_i|\ge K}) \le \varepsilon \mathbb{E}(\varphi(|X_i|)),$$

and taking supremum over i gives uniform integrability.

$(i) \Rightarrow (ii)$: The proof is based on constructing φ piecewise linear such that it grows faster than linear and uses the uniform integrability condition to control expectations. One defines a sequence $K_n \nearrow \infty$ such that

$$\sup_i \mathbb{E}(|X_i|\, \mathbb{1}_{|X_i|\ge K_n}) \le 2^{-n},$$

and then constructs φ increasing and convex with slope increasing on intervals $[K_n, K_{n+1}]$. A standard argument then shows $\sup_i \mathbb{E}(\varphi(|X_i|)) < \infty$. □

Corollary 3.37. *Let $p > 1$ and suppose $\sup_{i\in I} \mathbb{E}(|X_i|^p) < \infty$. Then $(X_i)_{i\in I}$ is uniformly integrable.*

Proof. Take $\varphi(x) = x^p$. Then $\varphi(x)/x = x^{p-1} \to \infty$, and Theorem 3.36 yields the claim. □

Lemma 3.38. *Let $(Y_n)_{n\geq 1}$ be a uniformly integrable family such that $Y_n \xrightarrow[n\to\infty]{\mathbb{P}} 0$. Then*

$$\mathbb{E}(|Y_n|) \xrightarrow[n\to\infty]{} 0.$$

Proof. Let $\varepsilon > 0$. By Lemma 3.33, there exists $\delta > 0$ such that for all $A \in \mathcal{F}$ with $\mathbb{P}(A) \leq \delta$,

$$\sup_{n\geq 1} \mathbb{E}(|Y_n| \mathbb{1}_A) \leq \varepsilon.$$

Since $Y_n \xrightarrow[n\to\infty]{\mathbb{P}} 0$, there exists N such that for all $n \geq N$,

$$\mathbb{P}(|Y_n| > \varepsilon) \leq \delta.$$

Therefore, for all $n \geq N$,

$$\mathbb{E}(|Y_n|) = \mathbb{E}(|Y_n| \mathbb{1}_{|Y_n|\leq\varepsilon}) + \mathbb{E}(|Y_n| \mathbb{1}_{|Y_n|>\varepsilon}) \leq \varepsilon + \varepsilon = 2\varepsilon.$$

This proves $\mathbb{E}(|Y_n|) \to 0$. □

Theorem 3.39. *Let $(X_n)_{n\geq 1}$ and X be integrable random variables. Then the following are equivalent:*

(i) $X_n \xrightarrow[n\to\infty]{L^1} X$, *i.e.* $\mathbb{E}(|X_n - X|) \to 0$.

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

Proof. $(i) \Rightarrow (ii)$: Convergence in L^1 implies convergence in probability by Markov's inequality. Moreover,

$$\mathbb{E}(|X_n|) \leq \mathbb{E}(|X|) + \mathbb{E}(|X_n - X|),$$

so $\sup_n \mathbb{E}(|X_n|) < \infty$. Uniform integrability follows from Lemma 3.33.

$(ii) \Rightarrow (i)$: Apply Lemma 3.38 to $Y_n := X_n - X$. □

Corollary 3.40. *Let $X_1, X_2, \ldots$ be i.i.d. random variables with $\mathbb{E}(|X_1|) < \infty$, and let*

$$\overline{X}_n := \frac{X_1 + \cdots + X_n}{n}, \qquad \mu := \mathbb{E}(X_1).$$

Then

$$\overline{X}_n \xrightarrow[n\to\infty]{L^1} \mu.$$

Proof. By the strong law of large numbers, $\overline{X}_n \xrightarrow[n\to\infty]{a.s.} \mu$, and hence also $\overline{X}_n \xrightarrow[n\to\infty]{\mathbb{P}} \mu$.

Since X_1 is integrable, the singleton family $\{X_1\}$ is uniformly integrable. Thus by Theorem 3.36, there exists an increasing convex function $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ such that $\varphi(x)/x \to \infty$ as $x \to \infty$ and

$$\mathbb{E}(\varphi(|X_1|)) < \infty.$$

Now

$$\left|\overline{X}_n\right| \leq \frac{1}{n} \sum_{k=1}^{n} |X_k|,$$

and therefore, by monotonicity and convexity of φ,

$$\varphi(\big|\overline{X}_n\big|) \le \varphi\Big(\frac{1}{n}\sum_{k=1}^{n}|X_k|\Big) \le \frac{1}{n}\sum_{k=1}^{n}\varphi(|X_k|).$$

Taking expectations and using that the X_k are identically distributed, we obtain

$$\mathbb{E}(\varphi(\big|\overline{X}_n\big|)) \le \frac{1}{n}\sum_{k=1}^{n}\mathbb{E}(\varphi(|X_k|)) = \mathbb{E}(\varphi(|X_1|)).$$

Hence

$$\sup_{n\ge 1}\mathbb{E}(\varphi(\big|\overline{X}_n\big|)) < \infty.$$

By Theorem 3.36, the family $(\overline{X}_n)_{n\ge1}$ is uniformly integrable. Since also $\overline{X}_n \xrightarrow[n\to\infty]{\mathbb{P}} \mu$, Theorem 3.39 yields

$$\overline{X}_n \xrightarrow[n\to\infty]{L^1} \mu.$$

□

Corollary 3.41. *Let $p \in [1,\infty)$ and let $X_1, X_2, \dots$ be i.i.d. random variables with $\mathbb{E}(|X_1|^p) < \infty$. Let*

$$\overline{X}_n := \frac{X_1 + \cdots + X_n}{n}, \qquad \mu := \mathbb{E}(X_1).$$

Then

$$\overline{X}_n \xrightarrow[n\to\infty]{L^p} \mu.$$

Proof. By the strong law of large numbers, $\overline{X}_n \xrightarrow[n\to\infty]{a.s.} \mu$, and hence

$$\big|\overline{X}_n - \mu\big|^p \xrightarrow[n\to\infty]{\mathbb{P}} 0.$$

Since $\mathbb{E}(|X_1-\mu|^p) < \infty$, the singleton family $\{|X_1-\mu|^p\}$ is uniformly integrable. Thus by Theorem 3.36, there exists an increasing convex function $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ such that $\varphi(x)/x \to \infty$ and

$$\mathbb{E}\big(\varphi(|X_1-\mu|^p)\big) < \infty.$$

Now by convexity of $x \mapsto |x|^p$,

$$\big|\overline{X}_n - \mu\big|^p = \left|\frac{1}{n}\sum_{k=1}^{n}(X_k-\mu)\right|^p \le \frac{1}{n}\sum_{k=1}^{n}|X_k-\mu|^p.$$

Hence, by monotonicity and convexity of φ,

$$\varphi\big(\,\big|\overline{X}_n-\mu\big|^p\,\big) \le \varphi\Big(\frac{1}{n}\sum_{k=1}^{n}|X_k-\mu|^p\Big) \le \frac{1}{n}\sum_{k=1}^{n}\varphi\big(\,|X_k-\mu|^p\,\big).$$

Taking expectations gives

$$\sup_{n\ge1}\mathbb{E}\big(\varphi(\big|\overline{X}_n-\mu\big|^p)\big) \le \mathbb{E}\big(\varphi(|X_1-\mu|^p)\big) < \infty.$$

Therefore the family $\left(\left|\overline{X}_n - \mu\right|^p \right)_{n\geq 1}$ is uniformly integrable by Theorem 3.36. Since also $\left|\overline{X}_n - \mu\right|^p \xrightarrow[n\to\infty]{\mathbb{P}} 0$, Theorem 3.39 implies

$$\mathbb{E}\left(\left|\overline{X}_n - \mu\right|^p \right) \xrightarrow[n\to\infty]{} 0.$$

This is exactly $\overline{X}_n \xrightarrow[n\to\infty]{L^p} \mu$. □

Chapter 4

Martingales

4.1 Conditional expectation and basic definitions

Throughout this subsection, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $\mathcal{G} \subset \mathcal{F}$ be a sub-σ-algebra. The intuition is that $\mathcal{G}$ describes the information available to us, and $\mathbb{E}[X \mid \mathcal{G}]$ is the best $\mathcal{G}$-measurable prediction of X.

We begin with the simplest case where $\mathcal{G}$ is generated by a finite partition.

Definition 4.1 (Conditional expectation for a finite partition)**.** Assume that $\mathcal{G} = \sigma(A_1, \dots, A_n)$, where $A_1, \dots, A_n$ form a partition of Ω. Let X be an integrable random variable. We define

$$\mathbb{E}[X \mid \mathcal{G}] := \sum_{i=1}^{n} c_i \mathbb{1}_{A_i}, \qquad \text{where} \qquad c_i = \begin{cases} \frac{\mathbb{E}(X \mathbb{1}_{A_i})}{\mathbb{P}(A_i)} & \text{if } \mathbb{P}(A_i) > 0, \\ 0 & \text{if } \mathbb{P}(A_i) = 0. \end{cases}$$

Lemma 4.2. *The random variable defined above is $\mathcal{G}$-measurable, integrable, and satisfies $\int_A \mathbb{E}[X \mid \mathcal{G}] \, d\mathbb{P} = \int_A X \, d\mathbb{P}$ for every $A \in \mathcal{G}$.*

Proof. The function is constant on each atom A_i, hence $\mathcal{G}$-measurable. It is also integrable since $\sum_i |c_i| \, \mathbb{P}(A_i) \le \sum_i \mathbb{E}(|X| \, \mathbb{1}_{A_i}) = \mathbb{E}(|X|)$. Finally, every set $A \in \mathcal{G}$ is a union of some of the atoms A_i, so the identity of integrals follows by summing over those atoms. □

This suggests the general definition.

Definition 4.3 (Conditional expectation)**.** Let X be an integrable random variable. A random variable Y is called a *conditional expectation of X given $\mathcal{G}$* if Y is $\mathcal{G}$-measurable, integrable, and

$$\int_A Y \, d\mathbb{P} = \int_A X \, d\mathbb{P} \qquad \forall A \in \mathcal{G}.$$

Any such random variable is denoted by $\mathbb{E}[X \mid \mathcal{G}]$.

Remark 4.4. The conditional expectation is determined uniquely up to almost sure equality. In these notes we will not emphasize this point further.

Theorem 4.5 (Existence and uniqueness)**.** *Let X be an integrable random variable and let $\mathcal{G} \subset \mathcal{F}$ be a sub-σ-algebra. Then $\mathbb{E}[X \mid \mathcal{G}]$ exists and is unique.*

Proof. Write $X = X^+ - X^-$. The maps $A \mapsto \int_A X^+ \, d\mathbb{P}$ and $A \mapsto \int_A X^- \, d\mathbb{P}$ define finite measures on $(\Omega, \mathcal{G})$ which are absolutely continuous with respect to $\mathbb{P}_{|\mathcal{G}}$. By the Radon–Nikodym theorem, they admit $\mathcal{G}$-measurable densities Y_+ and Y_-. Then $Y := Y_+ - Y_-$ is integrable and satisfies $\int_A Y \, d\mathbb{P} = \int_A X \, d\mathbb{P}$ for all $A \in \mathcal{G}$, so it is a conditional expectation.

For uniqueness, if Y and Y' both satisfy the definition, then $\int_A (Y - Y') \, d\mathbb{P} = 0$ for all $A \in \mathcal{G}$. Applying this to the sets $\{Y > Y'\}$ and $\{Y' < Y\}$ gives $Y = Y'$ almost surely. □

The defining property can also be tested against bounded $\mathcal{G}$-measurable random variables.

Lemma 4.6. *Let X be integrable and let $Y = \mathbb{E}[X \mid \mathcal{G}]$. If Z is bounded and $\mathcal{G}$-measurable, then $\mathbb{E}(ZY) = \mathbb{E}(ZX)$.*

Proof. First prove the claim for indicator functions $Z = \mathbb{1}_A$, where it is exactly the defining identity. By linearity this extends to simple $\mathcal{G}$-measurable functions, and then to bounded $\mathcal{G}$-measurable functions by approximation and dominated convergence. □

We now collect the basic properties of conditional expectation.

Theorem 4.7 (Basic properties of conditional expectation)**.** *Let X and W be integrable random variables, let $a, b \in \mathbb{R}$, and let $\mathcal{H} \subset \mathcal{G} \subset \mathcal{F}$ be sub-σ-algebras. Then:*

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

(ii) $\mathbb{E}[aX + bW \mid \mathcal{G}] = a\mathbb{E}[X \mid \mathcal{G}] + b\mathbb{E}[W \mid \mathcal{G}]$.

(iii) *If $X \geq 0$, then $\mathbb{E}[X \mid \mathcal{G}] \geq 0$. In particular, if $X \leq W$, then $\mathbb{E}[X \mid \mathcal{G}] \leq \mathbb{E}[W \mid \mathcal{G}]$.*

(iv) $\mathbb{E}\big(\mathbb{E}[X \mid \mathcal{G}]\big) = \mathbb{E}(X)$.

(v) *If Z is bounded and $\mathcal{G}$-measurable, then $\mathbb{E}[ZX \mid \mathcal{G}] = Z\mathbb{E}[X \mid \mathcal{G}]$.*

(vi) $\mathbb{E}[\mathbb{E}[X \mid \mathcal{G}] \mid \mathcal{H}] = \mathbb{E}[X \mid \mathcal{H}]$.

(vii) *If X is independent of $\mathcal{G}$, then $\mathbb{E}[X \mid \mathcal{G}] = \mathbb{E}(X)$.*

(viii) $|\mathbb{E}[X \mid \mathcal{G}]| \leq \mathbb{E}[|X| \mid \mathcal{G}]$. *In particular,* $\mathbb{E}(|\mathbb{E}[X \mid \mathcal{G}]|) \leq \mathbb{E}(|X|)$.

Proof. (1) is immediate from the definition.

For (2), the right-hand side is $\mathcal{G}$-measurable and has the same integrals over all $A \in \mathcal{G}$ as $aX + bW$, so uniqueness gives the claim.

For (3), if $Y := \mathbb{E}[X \mid \mathcal{G}]$ and $X \geq 0$, then $\int_A Y \, d\mathbb{P} \geq 0$ for all $A \in \mathcal{G}$. Taking $A = \{Y < 0\}$ gives $Y \geq 0$. Monotonicity follows by applying this to $W - X$.

Part (4) follows by taking $A = \Omega$ in the definition.

For (5), let $Y := \mathbb{E}[X \mid \mathcal{G}]$. Since Z is bounded and $\mathcal{G}$-measurable, ZY is $\mathcal{G}$-measurable and for all $A \in \mathcal{G}$ we have $\mathbb{E}(\mathbb{1}_A ZY) = \mathbb{E}(\mathbb{1}_A ZX)$ by Lemma 4.6. Hence $ZY = \mathbb{E}[ZX \mid \mathcal{G}]$.

Part (6) is the tower property. The random variable $\mathbb{E}[\mathbb{E}[X \mid \mathcal{G}] \mid \mathcal{H}]$ is $\mathcal{H}$-measurable, and for $A \in \mathcal{H} \subset \mathcal{G}$,

$$\int_A \mathbb{E}[\mathbb{E}[X \mid \mathcal{G}] \mid \mathcal{H}] \, d\mathbb{P} = \int_A \mathbb{E}[X \mid \mathcal{G}] \, d\mathbb{P} = \int_A X \, d\mathbb{P}.$$

Thus it equals $\mathbb{E}[X \mid \mathcal{H}]$.

For (7), the constant $\mathbb{E}(X)$ is $\mathcal{G}$-measurable, and for every $A \in \mathcal{G}$, independence gives $\mathbb{E}(X\mathbb{1}_A) = \mathbb{E}(X)\mathbb{P}(A) = \int_A \mathbb{E}(X) \, d\mathbb{P}$. Hence $\mathbb{E}[X \mid \mathcal{G}] = \mathbb{E}(X)$.

Finally, (8) follows from $-|X| \leq X \leq |X|$ and monotonicity. □

The next result is the conditional version of Jensen's inequality.

Theorem 4.8 (Conditional Jensen inequality)**.** *Let X be an integrable random variable, let $\varphi : \mathbb{R} \to \mathbb{R}$ be convex, and assume that $\varphi(X)$ is integrable. Then*

$$\varphi\big(\mathbb{E}\,[X \,|\, \mathcal{G}]\,\big) \leq \mathbb{E}\,[\varphi(X) \,|\, \mathcal{G}]\,.$$

Proof. A standard fact from convex analysis says that a convex function is the supremum of affine functions lying below it. Thus it is enough to prove the inequality for affine functions, where it is just linearity of conditional expectation. Taking the supremum over all such affine minorants yields the result. □

Finally, we present the Hilbert space interpretation of conditional expectation.

Theorem 4.9 (L^2 projection characterization)**.** *Let $X \in L^2(\Omega, \mathcal{F}, \mathbb{P})$. Then $\mathbb{E}\,[X \,|\, \mathcal{G}] \in L^2$, and it is characterized by the following orthogonality property:*

$$\mathbb{E}\big((X - \mathbb{E}\,[X \,|\, \mathcal{G}])Z\big) = 0$$

for every bounded $\mathcal{G}$-measurable random variable Z, and hence also for every square-integrable $\mathcal{G}$-measurable random variable Z.

Equivalently, $\mathbb{E}\,[X \,|\, \mathcal{G}]$ is the unique $\mathcal{G}$-measurable square-integrable random variable minimizing $\mathbb{E}\big((X - Z)^2\big)$ over all $\mathcal{G}$-measurable $Z \in L^2$.

Proof. Since $x \mapsto x^2$ is convex, conditional Jensen gives $\mathbb{E}\,[X \,|\, \mathcal{G}]^2 \leq \mathbb{E}\,[X^2 \,|\, \mathcal{G}]$, hence $\mathbb{E}\,[X \,|\, \mathcal{G}] \in L^2$ after taking expectations.

The orthogonality identity follows from Lemma 4.6: if $Y = \mathbb{E}\,[X \,|\, \mathcal{G}]$ and Z is bounded and $\mathcal{G}$-measurable, then $\mathbb{E}(XZ) = \mathbb{E}(YZ)$, so $\mathbb{E}((X - Y)Z) = 0$.

Conversely, if Y is $\mathcal{G}$-measurable and satisfies $\mathbb{E}((X - Y)Z) = 0$ for all bounded $\mathcal{G}$-measurable Z, then taking $Z = \mathbb{1}_A$ shows that $\int_A Y \, d\mathbb{P} = \int_A X \, d\mathbb{P}$ for all $A \in \mathcal{G}$, hence $Y = \mathbb{E}\,[X \,|\, \mathcal{G}]$.

For the minimization property, write

$$X - Z = (X - Y) + (Y - Z),$$

where $Y = \mathbb{E}\,[X \,|\, \mathcal{G}]$. Since $Y - Z$ is $\mathcal{G}$-measurable, the orthogonality gives $\mathbb{E}((X - Y)(Y - Z)) = 0$. Expanding the square yields

$$\mathbb{E}\big((X - Z)^2\big) = \mathbb{E}\big((X - Y)^2\big) + \mathbb{E}\big((Y - Z)^2\big),$$

so Y is the unique minimizer. □

Remark 4.10. Thus, in L^2, conditional expectation is precisely the orthogonal projection onto the closed subspace of $\mathcal{G}$-measurable random variables.

4.2 Martingales as fair games and optional stopping

In this section we work in discrete time, indexed by $t \in \{0, 1, 2, \dots\}$. The same definitions extend without difficulty to any time set $T \subseteq \mathbb{Z}$.

Definition 4.11 (Filtration and adapted process)**.** A sequence $(\mathcal{F}_t)_{t\geq 0}$ of sub-σ-algebras of $\mathcal{F}$ is called a *filtration* if $\mathcal{F}_s \subset \mathcal{F}_t$ whenever $s \leq t$.

A process $X = (X_t)_{t\geq 0}$ is *adapted* to $(\mathcal{F}_t)_{t\geq 0}$ if X_t is $\mathcal{F}_t$-measurable for every $t \geq 0$.

Remark 4.12. The filtration $(\mathcal{F}_t)$ should be thought of as the information available up to time t. Thus adaptedness means that X_t is observable at time t.

Definition 4.13 (Martingale, submartingale, supermartingale)**.** Let $(\mathcal{F}_t)_{t\geq 0}$ be a filtration. An adapted process $X = (X_t)_{t\geq 0}$ is called

- a *martingale* if $\mathbb{E}(|X_t|) < \infty$ for all t and $\mathbb{E}[X_t \,|\, \mathcal{F}_s] = X_s$ for all $s \leq t$,
- a *submartingale* if $\mathbb{E}(|X_t|) < \infty$ for all t and $\mathbb{E}[X_t \,|\, \mathcal{F}_s] \geq X_s$ for all $s \leq t$,
- a *supermartingale* if $\mathbb{E}(|X_t|) < \infty$ for all t and $\mathbb{E}[X_t \,|\, \mathcal{F}_s] \leq X_s$ for all $s \leq t$.

Remark 4.14. A martingale is the mathematical model of a *fair game*: given the present information, the best prediction for the future value is the current value.

Proposition 4.15 (One-step characterization)**.** *Let $X = (X_t)_{t\geq 0}$ be an adapted integrable process. Then X is a martingale if and only if $\mathbb{E}[X_t \,|\, \mathcal{F}_{t-1}] = X_{t-1}$ for every $t \geq 1$. Analogous one-step characterizations hold for submartingales and supermartingales.*

Proof. If X is a martingale, the one-step identity is immediate. Conversely, if $\mathbb{E}[X_t \,|\, \mathcal{F}_{t-1}] = X_{t-1}$ for every $t \geq 1$, then for $s < t$ repeated use of the tower property gives

$$\mathbb{E}[X_t \,|\, \mathcal{F}_s] = \mathbb{E}[\mathbb{E}[X_t \,|\, \mathcal{F}_{t-1}] \,|\, \mathcal{F}_s] = \mathbb{E}[X_{t-1} \,|\, \mathcal{F}_s] = \cdots = X_s.$$

□

Example 4.16 (Random walk)**.** Let $\xi_1, \xi_2, \dots$ be independent integrable random variables with $\mathbb{E}(\xi_i) = 0$, and define $S_0 := 0$ and $S_t := \xi_1 + \cdots + \xi_t$ for $t \geq 1$. If $\mathcal{F}_t := \sigma(\xi_1, \dots, \xi_t)$, then $(S_t)_{t\geq 0}$ is a martingale. Indeed, S_t is adapted and

$$\mathbb{E}[S_t \,|\, \mathcal{F}_{t-1}] = S_{t-1} + \mathbb{E}[\xi_t \,|\, \mathcal{F}_{t-1}] = S_{t-1} + \mathbb{E}(\xi_t) = S_{t-1},$$

since ξ_t is independent of $\mathcal{F}_{t-1}$.

Example 4.17 (Doob martingale)**.** Let Y be an integrable random variable and let $(\mathcal{F}_t)_{t\geq 0}$ be a filtration. Define $M_t := \mathbb{E}[Y \,|\, \mathcal{F}_t]$. Then $(M_t)_{t\geq 0}$ is a martingale. Indeed, by the tower property,

$$\mathbb{E}[M_t \,|\, \mathcal{F}_s] = \mathbb{E}[\mathbb{E}[Y \,|\, \mathcal{F}_t] \,|\, \mathcal{F}_s] = \mathbb{E}[Y \,|\, \mathcal{F}_s] = M_s \qquad \text{for } s \leq t.$$

This is called the *Doob martingale* associated with Y and $(\mathcal{F}_t)$.

To describe gains and losses from trading, we introduce predictable processes.

Definition 4.18 (Predictable process)**.** A process $H = (H_t)_{t\geq 1}$ is called *predictable* with respect to $(\mathcal{F}_t)_{t\geq 0}$ if H_t is $\mathcal{F}_{t-1}$-measurable for every $t \geq 1$.

It is called *bounded* if there exists $C < \infty$ such that $|H_t| \leq C$ for all $t \geq 1$.

Remark 4.19. The interpretation is that H_t is chosen at time $t-1$, before the increment $X_t - X_{t-1}$ is observed.

Definition 4.20 (Gain from trading)**.** Let $X = (X_t)_{t\geq 0}$ be an adapted process and let $H = (H_t)_{t\geq 1}$ be predictable. The *gain process* (or *stochastic integral*) of the strategy H against X is defined by

$$(H \cdot X)_0 := 0, \qquad (H \cdot X)_t := \sum_{i=1}^{t} H_i(X_i - X_{i-1}), \qquad t \geq 1.$$

Remark 4.21. The quantity H_i is interpreted as the number of units of the asset held during the time interval $(i-1, i]$. The corresponding gain on that interval is then $H_i(X_i - X_{i-1})$.

Lemma 4.22. *Let $X = (X_t)_{t\geq 0}$ be a martingale and let $H = (H_t)_{t\geq 1}$ be bounded and predictable. Then $(H \cdot X)_t$ is a martingale.*

Proof. The process $(H \cdot X)_t$ is adapted. Since H is bounded and $X_t - X_{t-1}$ is integrable, $(H \cdot X)_t$ is integrable for every t. Moreover,

$$\mathbb{E}\left[(H \cdot X)_t \,|\, \mathcal{F}_{t-1}\right] = (H \cdot X)_{t-1} + H_t\, \mathbb{E}\left[X_t - X_{t-1} \,|\, \mathcal{F}_{t-1}\right] = (H \cdot X)_{t-1},$$

because H_t is $\mathcal{F}_{t-1}$-measurable and X is a martingale. Thus Proposition 4.15 applies. □

The next result says that a process is a martingale precisely when no bounded predictable trading strategy has positive expected gain. It formalizes the intuition that martingales are *fair games*.

Theorem 4.23 (Martingale characterization by trading strategies)**.** *Let $X = (X_t)_{t\geq 0}$ be an adapted integrable process. Then the following are equivalent:*

(i) X is a martingale.

(ii) For every bounded predictable process H and every $t \geq 0$, $\mathbb{E}((H \cdot X)_t) = 0$.

Proof. If X is a martingale, then Lemma 4.22 shows that $H \cdot X$ is a martingale, and since $(H \cdot X)_0 = 0$, we get $\mathbb{E}((H \cdot X)_t) = 0$ for all t.

Conversely, assume (2). Fix $t \geq 1$ and let $A \in \mathcal{F}_{t-1}$. Define a predictable process H by $H_t = \mathbb{1}_A$ and $H_i = 0$ for $i \neq t$. Then

$$0 = \mathbb{E}((H \cdot X)_t) = \mathbb{E}(\mathbb{1}_A(X_t - X_{t-1})).$$

Since this holds for every $A \in \mathcal{F}_{t-1}$, it follows that $\mathbb{E}\left[X_t - X_{t-1} \,|\, \mathcal{F}_{t-1}\right] = 0$, that is, $\mathbb{E}\left[X_t \,|\, \mathcal{F}_{t-1}\right] = X_{t-1}$. By Proposition 4.15, X is a martingale. □

We now turn to stopping times, which model random times determined by the information observed so far.

Definition 4.24 (Stopping time)**.** A map $\tau : \Omega \to \{0, 1, 2, \dots\} \cup \{\infty\}$ is called a *stopping time* with respect to $(\mathcal{F}_t)_{t\geq 0}$ if $\{\tau \leq t\} \in \mathcal{F}_t$ for every $t \geq 0$.

Remark 4.25. The condition means that by time t one can decide whether or not the stopping time has already occurred. In discrete time this is equivalent to $\{\tau = t\} \in \mathcal{F}_t$ for every t.

Definition 4.26 (Stopped process)**.** Let $X = (X_t)_{t\geq 0}$ be a process and let τ be a stopping time. The *stopped process* $X^\tau = (X_t^\tau)_{t\geq 0}$ is defined by $X_t^\tau := X_{t\wedge\tau}$.

Remark 4.27 (Buy and hold until τ)**.** Let τ be a stopping time and define $H_t := \mathbb{1}_{\tau\geq t}$ for $t \geq 1$. Then H is predictable, since $\{\tau \geq t\} = \{\tau \leq t-1\}^c \in \mathcal{F}_{t-1}$. This strategy means: buy one unit at time 0, hold it until time τ, and then sell it. A telescoping sum gives

$$(H \cdot X)_t = \sum_{i=1}^{t} \mathbb{1}_{\tau\geq i}(X_i - X_{i-1}) = X_{t\wedge\tau} - X_0 = X_t^\tau - X_0.$$

Thus $X_t^\tau = X_0 + (H \cdot X)_t$.

Theorem 4.28 (Optional stopping theorem for bounded stopping times)**.** *Let $X = (X_t)_{t\geq 0}$ be an adapted integrable process. Then the following are equivalent:*

(i) X is a martingale.

(ii) For every bounded stopping time τ, we have

$$\mathbb{E}(X_\tau) = \mathbb{E}(X_0).$$

Moreover, in this case, for every bounded stopping time τ, the stopped process

$$X^\tau = (X_{t\wedge\tau})_{t\geq 0}$$

is a martingale.

Proof. Assume first that X is a martingale, and let τ be a bounded stopping time. By Remark 4.27, if $H_t := \mathbb{1}_{\tau\geq t}$, then H is bounded and predictable and

$$X_t^\tau = X_0 + (H \cdot X)_t.$$

Since X is a martingale, Lemma 4.22 implies that $H \cdot X$ is a martingale, and hence so is X^τ. In particular, if $\tau \leq N$ almost surely, then $X_\tau = X_N^\tau$, so

$$\mathbb{E}(X_\tau) = \mathbb{E}(X_N^\tau) = \mathbb{E}(X_0^\tau) = \mathbb{E}(X_0).$$

Conversely, assume that $\mathbb{E}(X_\tau) = \mathbb{E}(X_0)$ for every bounded stopping time τ. Fix $t \geq 1$ and let $A \in \mathcal{F}_{t-1}$. Define

$$\tau := (t-1)\mathbb{1}_A + t\mathbb{1}_{A^c}.$$

Then τ is a bounded stopping time, so by assumption

$$\mathbb{E}(X_\tau) = \mathbb{E}(X_0).$$

Also, the constant stopping time $t-1$ is bounded, hence

$$\mathbb{E}(X_{t-1}) = \mathbb{E}(X_0).$$

Subtracting gives

$$\mathbb{E}(X_t\mathbb{1}_{A^c} + X_{t-1}\mathbb{1}_A) = \mathbb{E}(X_{t-1}),$$

and therefore

$$\mathbb{E}\big((X_t - X_{t-1})\mathbb{1}_{A^c}\big) = 0.$$

Since $A \in \mathcal{F}_{t-1}$ was arbitrary, this implies

$$\mathbb{E}(X_t - X_{t-1} \mid \mathcal{F}_{t-1}) = 0,$$

that is,

$$\mathbb{E}(X_t \mid \mathcal{F}_{t-1}) = X_{t-1}.$$

By Proposition 4.15, X is a martingale.

The final statement was already proved in the first part. □

Remark 4.29. The boundedness assumption on τ is essential for the above short proof. More general forms of the optional stopping theorem require additional assumptions on the martingale or on the stopping time.

4.3 Limit theorems for martingales

We continue with discrete-time martingales $(X_t)_{t\geq 0}$ adapted to a filtration $(\mathcal{F}_t)_{t\geq 0}$.

A basic tool is the number of upcrossings of an interval.

Definition 4.30 (Upcrossings)**.** Let $a < b$ and let $x_0, \dots, x_n \in \mathbb{R}$. We define recursively the stopping indices

$$\sigma_1 := \inf\{k \leq n : x_k \leq a\}, \qquad \tau_1 := \inf\{k \geq \sigma_1 : x_k \geq b\},$$

and then $\sigma_{m+1} := \inf\{k \geq \tau_m : x_k \leq a\}$, $\tau_{m+1} := \inf\{k \geq \sigma_{m+1} : x_k \geq b\}$, with the convention that the infimum of the empty set is ∞. The number of completed upcrossings of $[a, b]$ by the finite path $(x_0, \dots, x_n)$ is

$$U_n[a, b] := \sup\{m \geq 0 : \tau_m \leq n\}.$$

If $X = (X_t)_{t\geq 0}$ is a process, we write $U_n[a, b](\omega)$ for the corresponding number of upcrossings of the sample path $(X_0(\omega), \dots, X_n(\omega))$.

The following inequality is the key estimate.

Theorem 4.31 (Doob's upcrossing inequality)**.** *Let $X = (X_t)_{t\geq 0}$ be a submartingale and let $a < b$. Then for every $n \geq 0$,*

$$(b - a)\,\mathbb{E}\big(U_n[a, b]\big) \leq \mathbb{E}\big((X_n - a)^+\big).$$

Proof. The idea is to use a buy-and-hold strategy: buy one unit when the process is at or below a, hold it until it reaches b, then sell it, and repeat. More precisely, one defines a predictable process $H = (H_t)_{t\geq 1}$ with values in $\{0, 1\}$ which is equal to 1 exactly during the intervals in which one is holding the asset.

Then $(H \cdot X)_n$ is the total gain of the strategy up to time n. Since H is bounded and predictable and X is a submartingale, the same computation as in Lemma 4.22 shows that $H \cdot X$ is a submartingale, so $\mathbb{E}((H \cdot X)_n) \geq 0$.

On the other hand, every completed upcrossing contributes at least $b-a$ to the gain, and the only possible error comes from the last incomplete trade, which is bounded below by $-(X_n-a)^+$. Therefore

$$(H\cdot X)_n \le (X_n-a)^+ - (b-a)\,U_n[a,b].$$

Taking expectations yields the claim. □

We now use the upcrossing inequality to prove almost sure convergence.

Theorem 4.32 (Almost sure convergence of L^1-bounded martingales)**.** *Let $X=(X_t)_{t\ge 0}$ be a martingale such that $\sup_{t\ge 0}\mathbb{E}(|X_t|)<\infty$. Then there exists an integrable random variable X_∞ such that $X_t\to X_\infty$ almost surely.*

Proof. Fix rational numbers $a<b$. Since X is a martingale, it is in particular a submartingale, and Theorem 4.31 gives

$$(b-a)\,\mathbb{E}\big(U_n[a,b]\big) \le \mathbb{E}\big((X_n-a)^+\big) \le \mathbb{E}(|X_n|)+|a|\,.$$

Hence $\sup_n \mathbb{E}(U_n[a,b])<\infty$, so $U_n[a,b]$ is almost surely bounded in n. Therefore the total number of upcrossings of $[a,b]$ is almost surely finite.

Since this holds for every rational pair $a<b$, it follows that almost surely the path $t\mapsto X_t(\omega)$ cannot oscillate between two distinct limit points. Thus $\lim_{t\to\infty} X_t$ exists almost surely. Denote the limit by X_∞.

Finally, Fatou's lemma gives $\mathbb{E}(|X_\infty|)\le \liminf_t \mathbb{E}(|X_t|)<\infty$, so X_∞ is integrable. □

The next notion tells when a martingale can be completed by a terminal value at infinity.

Definition 4.33 (Closable martingale)**.** A martingale $X=(X_t)_{t\ge 0}$ is called *closable* if there exists an integrable random variable X_∞ such that

$$X_t = \mathbb{E}\,[X_\infty \,|\, \mathcal{F}_t] \qquad \text{for every } t\ge 0.$$

Theorem 4.34 (Closable if and only if uniformly integrable)**.** *Let $X=(X_t)_{t\ge 0}$ be a martingale. Then X is closable if and only if the family $(X_t)_{t\ge 0}$ is uniformly integrable. In this case X_∞ can be chosen such that $X_t\to X_\infty$ a.s. and in L^1.*

Proof. Assume first that X is closable, so $X_t=\mathbb{E}\,[X_\infty\,|\,\mathcal{F}_t]$ for some integrable X_∞. By the de la Vallée–Poussin criterion, there exists an increasing convex function $\varphi:\mathbb{R}_+\to\mathbb{R}_+$ with $\varphi(x)/x\to\infty$ and $\mathbb{E}(\varphi(|X_\infty|))<\infty$. Conditional Jensen gives

$$\varphi(|X_t|) = \varphi\big(\,|\mathbb{E}\,[X_\infty\,|\,\mathcal{F}_t]|\,\big) \le \mathbb{E}\,[\varphi(|X_\infty|)\,|\,\mathcal{F}_t]\,,$$

hence $\sup_t \mathbb{E}(\varphi(|X_t|))\le \mathbb{E}(\varphi(|X_\infty|))<\infty$. Another application of the de la Vallée–Poussin criterion shows that $(X_t)_{t\ge 0}$ is uniformly integrable.

Conversely, assume that $(X_t)_{t\ge 0}$ is uniformly integrable. Then in particular $\sup_t \mathbb{E}(|X_t|)<\infty$, so Theorem 4.32 yields almost sure convergence $X_t\to X_\infty$ for some integrable X_∞. Since almost sure convergence implies convergence in probability, and (X_t) is uniformly integrable, Theorem 3.39 implies $X_t\to X_\infty$ in L^1.

Now fix $s\ge 0$ and $A\in\mathcal{F}_s$. For all $t\ge s$ we have $\mathbb{E}(X_t\mathbb{1}_A)=\mathbb{E}(X_s\mathbb{1}_A)$ because X is a martingale. Passing to the limit and using L^1 convergence gives $\mathbb{E}(X_\infty\mathbb{1}_A)=\mathbb{E}(X_s\mathbb{1}_A)$. Hence $X_s=\mathbb{E}\,[X_\infty\,|\,\mathcal{F}_s]$, so X is closable. □

Theorem 4.35 (L^2-closable martingales)**.** *Let $X = (X_t)_{t\geq 0}$ be a closable martingale with final element X_∞, and assume $X_\infty \in L^2$. Then $(X_t)_{t\geq 0}$ is bounded in L^2 and*

$$X_t \xrightarrow[t\to\infty]{L^2} X_\infty.$$

Proof. Since $X_t = \mathbb{E}[X_\infty \,|\, \mathcal{F}_t]$, conditional Jensen gives $\mathbb{E}(X_t^2) \leq \mathbb{E}(X_\infty^2)$ for all t, so (X_t) is L^2-bounded. Also,

$$\mathbb{E}\big((X_\infty - X_t)^2\big) = \mathbb{E}(X_\infty^2) - \mathbb{E}(X_t^2),$$

because $\mathbb{E}(X_t X_\infty) = \mathbb{E}(X_t\,\mathbb{E}[X_\infty \,|\, \mathcal{F}_t]) = \mathbb{E}(X_t^2)$. Since closable martingales are uniformly integrable by Theorem 4.34, we have $X_t \to X_\infty$ almost surely. Fatou's lemma then yields $\mathbb{E}(X_\infty^2) \leq \liminf_t \mathbb{E}(X_t^2)$, while already $\mathbb{E}(X_t^2) \leq \mathbb{E}(X_\infty^2)$. Hence $\mathbb{E}(X_t^2) \to \mathbb{E}(X_\infty^2)$, and therefore $\mathbb{E}((X_\infty - X_t)^2) \to 0$. □

Corollary 4.36 (L^2-bounded martingales are closable)**.** *Let $X = (X_t)_{t\geq 0}$ be a martingale such that $\sup_{t\geq 0} \mathbb{E}(X_t^2) < \infty$. Then there exists $X_\infty \in L^2$ such that $X_t = \mathbb{E}[X_\infty \,|\, \mathcal{F}_t]$ for all $t \geq 0$. In particular,*

$$X_t \xrightarrow[t\to\infty]{a.s.} X_\infty \qquad \text{and} \qquad X_t \xrightarrow[t\to\infty]{L^2} X_\infty.$$

Proof. Since $2 > 1$, Corollary 3.37 implies that $(X_t)_{t\geq 0}$ is uniformly integrable. Hence Theorem 4.34 shows that X is closable, so there exists an integrable X_∞ with $X_t = \mathbb{E}[X_\infty \,|\, \mathcal{F}_t]$. By Theorem 4.32, we also have $X_t \to X_\infty$ almost surely.

Since $X_t^2 \to X_\infty^2$ almost surely, Fatou's lemma gives $\mathbb{E}(X_\infty^2) \leq \liminf_t \mathbb{E}(X_t^2) < \infty$, so $X_\infty \in L^2$. The L^2 convergence now follows from Theorem 4.35. □

We finally consider martingales indexed by the negative integers.

Definition 4.37 (Backward martingale)**.** Let $(\mathcal{F}_t)_{t\in -\mathbb{N}}$ be a filtration indexed by $-\mathbb{N} = \{0, -1, -2, \dots\}$, so that $\mathcal{F}_s \subset \mathcal{F}_t$ whenever $s < t \leq 0$. An adapted integrable process $(X_t)_{t\in -\mathbb{N}}$ is called a *backward martingale* if

$$\mathbb{E}[X_t \,|\, \mathcal{F}_s] = X_s \qquad \text{for all } s < t \leq 0.$$

Remark 4.38. Equivalently, for every $t \leq 0$ we have $X_t = \mathbb{E}[X_0 \,|\, \mathcal{F}_t]$.

Theorem 4.39 (Backward martingales are closable)**.** *Let $(X_t)_{t\in -\mathbb{N}}$ be a backward martingale, and set $\mathcal{F}_{-\infty} := \bigcap_{t\in -\mathbb{N}} \mathcal{F}_t$. Then there exists an integrable random variable $X_{-\infty}$ such that*

$$X_t = \mathbb{E}[X_0 \,|\, \mathcal{F}_t] \qquad \text{for all } t \in -\mathbb{N}, \qquad \text{and} \qquad X_t \to X_{-\infty} \quad \text{a.s. and in } L^1,$$

where $X_{-\infty} = \mathbb{E}[X_0 \,|\, \mathcal{F}_{-\infty}]$.

Proof. Since $X_t = \mathbb{E}[X_0 \,|\, \mathcal{F}_t]$, the same conditional Jensen argument as in the proof of Theorem 4.34 shows that the family $(X_t)_{t\in -\mathbb{N}}$ is uniformly integrable. Thus one may repeat the previous argument with $t \to -\infty$ in place of $t \to \infty$: the upcrossing argument gives almost sure convergence, and Theorem 3.39 upgrades this to L^1 convergence. Denote the limit by $X_{-\infty}$.

Since $\mathcal{F}_{-\infty} \subset \mathcal{F}_t$ for all $t \in -\mathbb{N}$, passing to the limit in the identity $X_t = \mathbb{E}[X_0 \,|\, \mathcal{F}_t]$ gives $X_{-\infty} = \mathbb{E}[X_0 \,|\, \mathcal{F}_{-\infty}]$. Thus the backward martingale has a genuine first element at $-\infty$. □

Remark 4.40. The last theorem is often called the *backward martingale convergence theorem*. It is the natural counterpart of the forward convergence theorem above.

4.4 A martingale proof of the strong law of large numbers

We now give another proof of the strong law of large numbers, based on the backward martingale convergence theorem.

Let $Y_1, Y_2, \dots$ be i.i.d. integrable random variables, and write $S_n := Y_1 + \cdots + Y_n$ for $n \geq 1$. For $n \geq 1$, define

$$\mathcal{F}_{-n} := \sigma(S_n, S_{n+1}, S_{n+2}, \dots).$$

Since $S_m - S_n = Y_{n+1} + \cdots + Y_m$, this is the same as $\sigma(S_n, Y_{n+1}, Y_{n+2}, \dots)$. In particular, $(\mathcal{F}_{-n})_{n\geq 1}$ is a decreasing filtration indexed by the negative integers.

Lemma 4.41. *For every $n \geq 1$, we have*

$$\mathbb{E}\left[Y_1 \,|\, \mathcal{F}_{-n}\right] = \frac{S_n}{n}.$$

Hence the process $X_{-n} := \frac{S_n}{n}$, $n \geq 1$, is a backward martingale.

Proof. The σ-algebra $\mathcal{F}_{-n}$ is unchanged under permutations of $Y_1, \dots, Y_n$. Therefore, by symmetry,

$$\mathbb{E}\left[Y_1 \,|\, \mathcal{F}_{-n}\right] = \mathbb{E}\left[Y_2 \,|\, \mathcal{F}_{-n}\right] = \cdots = \mathbb{E}\left[Y_n \,|\, \mathcal{F}_{-n}\right].$$

Summing these equalities and using linearity of conditional expectation gives

$$n\,\mathbb{E}\left[Y_1 \,|\, \mathcal{F}_{-n}\right] = \mathbb{E}\left[Y_1 + \cdots + Y_n \,|\, \mathcal{F}_{-n}\right] = \mathbb{E}\left[S_n \,|\, \mathcal{F}_{-n}\right] = S_n,$$

since S_n is $\mathcal{F}_{-n}$-measurable. This proves the claim. □

Theorem 4.42 (Strong law of large numbers via backward martingales)**.** *Let $Y_1, Y_2, \dots$ be i.i.d. random variables with $\mathbb{E}(|Y_1|) < \infty$. Then*

$$\frac{S_n}{n} \xrightarrow[n\to\infty]{a.s.} \mathbb{E}(Y_1).$$

In fact, the convergence also holds in L^1.

Proof. By Lemma 4.41, the process $X_{-n} = S_n/n$ is a backward martingale. Hence Theorem 4.39 yields the existence of an integrable random variable $X_{-\infty}$ such that

$$\frac{S_n}{n} \xrightarrow[n\to\infty]{a.s.} X_{-\infty} \qquad \text{and} \qquad \frac{S_n}{n} \xrightarrow[n\to\infty]{L^1} X_{-\infty}.$$

It remains to identify $X_{-\infty}$. Since changing finitely many of the variables Y_i does not affect the limit of the averages S_n/n, the random variable $X_{-\infty}$ is measurable with respect to the tail σ-algebra $\bigcap_{m\geq 1} \sigma(Y_m, Y_{m+1}, \dots)$. By Kolmogorov's 0-1 law, this tail σ-algebra is trivial, so $X_{-\infty}$ is almost surely constant.

Finally, by L^1 convergence,

$$\mathbb{E}(X_{-\infty}) = \lim_{n\to\infty} \mathbb{E}\left(\frac{S_n}{n}\right) = \mathbb{E}(Y_1),$$

so the constant must be $\mathbb{E}(Y_1)$. Thus $S_n/n \to \mathbb{E}(Y_1)$ almost surely, and also in L^1. □

Remark 4.43. This proof is conceptually very clean: the averages S_n/n form a backward martingale, so convergence is automatic from the backward martingale theorem, and Kolmogorov's 0-1 law identifies the limit.

Chapter 5

Weak convergence

So far, most of our notions of convergence have been formulated directly for random variables. This is a very natural point of view: random variables live in linear spaces, and this vector space structure allows one to define and study quantities such as expectation, variance, covariance, and conditional expectation in a particularly clean way. However, convergence of random variables is often too strong for asymptotic questions. In many important limit theorems, one does not expect convergence of the underlying random variables on the original probability space, but only convergence of their laws. A fundamental example is the central limit theorem, where the normalized sums converge in distribution to a Gaussian law, but not as random variables on the same space in any obvious sense. This motivates the introduction of weak convergence of probability measures, which is one of the central concepts of modern probability theory.

In this chapter, we introduce weak convergence on general Polish spaces. We then derive its main equivalent formulations, study the particularly important case of the real line, and finally discuss tightness and Prokhorov's theorem. These results provide the basic compactness theory for probability measures and form the basis for many later limit theorems.

5.1 Weak convergence on Polish spaces

Throughout this section, let (M, d) be a Polish space, i.e. a complete separable metric space, equipped with its Borel σ-algebra $\mathcal{B}(M)$. We denote by $\mathcal{P}(M)$ the set of all Borel probability measures on M.

Definition 5.1 (Weak convergence)**.** Let $\mu, \mu_1, \mu_2, \dots \in \mathcal{P}(M)$. We say that μ_n converges *weakly* to μ, and write

$$\mu_n \Rightarrow \mu,$$

if

$$\int_M f \, d\mu_n \to \int_M f \, d\mu$$

for every bounded continuous function $f : M \to \mathbb{R}$.

Remark 5.2. If X_n and X are M-valued random variables with laws μ_n and μ, respectively, then one also says that X_n converges *in distribution* or *in law* to X if $\mu_n \Rightarrow \mu$.

Example 5.3. If $x_n, x \in M$ and $x_n \to x$, then the Dirac measures satisfy $\delta_{x_n} \Rightarrow \delta_x$. This follows immediately from continuity of bounded continuous test functions.

The following theorem gives the most useful equivalent formulations of weak convergence.

Theorem 5.4 (Portmanteau theorem)**.** *Let $\mu, \mu_1, \mu_2, \dots \in \mathcal{P}(M)$. Then the following are equivalent:*

(i) $\mu_n \Rightarrow \mu$.

(ii) $\int f \, d\mu_n \to \int f \, d\mu$ *for every bounded uniformly continuous function* $f : M \to \mathbb{R}$.

(iii) For every closed set $F \subset M$,

$$\limsup_{n\to\infty} \mu_n(F) \le \mu(F).$$

(iv) For every open set $G \subset M$,

$$\liminf_{n\to\infty} \mu_n(G) \ge \mu(G).$$

(v) For every Borel set $A \subset M$ *such that* $\mu(\partial A) = 0$,

$$\mu_n(A) \to \mu(A).$$

Proof. We only sketch the main ideas.

(1) $\Rightarrow$ (2) is immediate. To show (1) $\Rightarrow$ (3), approximate the indicator of a closed set F from above by bounded continuous functions such as $f_m(x) := (1 - m\, d(x, F))^+$. Applying weak convergence to f_m and then letting $m \to \infty$ yields the inequality for closed sets. The implication (1) $\Rightarrow$ (4) follows by passing to complements.

To obtain (4) $\Rightarrow$ (5), apply the open-set inequality to A° and the closed-set inequality to $\overline{A}$, and use the assumption $\mu(\partial A) = 0$.

Finally, (5) $\Rightarrow$ (1) follows by approximating a bounded continuous function from above and below by simple functions supported on continuity sets of μ. This is standard and gives back convergence against bounded continuous test functions. □

The case of the real line is particularly important, since then weak convergence can be read off from distribution functions.

Theorem 5.5 (Weak convergence on $\mathbb{R}$)**.** *Let $\mu, \mu_1, \mu_2, \dots \in \mathcal{P}(\mathbb{R})$, and let*

$$F_n(x) := \mu_n((-\infty, x]), \qquad F(x) := \mu((-\infty, x])$$

be their distribution functions. Then $\mu_n \Rightarrow \mu$ if and only if

$$F_n(x) \to F(x)$$

for every continuity point x of F.

Proof. Assume first that $\mu_n \Rightarrow \mu$. Since $(-\infty, x]$ is closed and $(-\infty, x)$ is open, the Portmanteau theorem gives

$$\mu((-\infty, x)) \le \liminf_{n\to\infty} F_n(x) \le \limsup_{n\to\infty} F_n(x) \le \mu((-\infty, x]).$$

If x is a continuity point of F, then $\mu(\{x\}) = 0$, so the left and right hand sides coincide, yielding $F_n(x) \to F(x)$.

Conversely, assume $F_n(x) \to F(x)$ at every continuity point of F. To prove $\mu_n \Rightarrow \mu$, it suffices by the Portmanteau theorem to show convergence on intervals and then on all continuity sets of μ. This is standard on $\mathbb{R}$, because intervals generate the Borel σ-algebra and the boundary of an interval consists of finitely many points. Thus the assumed convergence of the distribution functions implies convergence of $\mu_n(A)$ to $\mu(A)$ for every Borel set A with $\mu(\partial A) = 0$, and the Portmanteau theorem yields $\mu_n \Rightarrow \mu$. □

The next notion is the correct compactness condition for families of probability measures.

Definition 5.6 (Tightness)**.** A family $\mathcal{K} \subset \mathcal{P}(M)$ is called *tight* if for every $\varepsilon > 0$ there exists a compact set $K \subset M$ such that

$$\mu(K) \ge 1 - \varepsilon \qquad \text{for all } \mu \in \mathcal{K}.$$

A sequence $(\mu_n)_{n\ge1}$ is tight if the family $\{\mu_n : n \ge 1\}$ is tight.

Remark 5.7. Tightness says that the mass of the measures cannot escape to infinity. It is the analogue of boundedness for families of probability measures.

The fundamental compactness theorem is Prokhorov's theorem.

Theorem 5.8 (Prokhorov's theorem)**.** *Let M be a Polish space and let $\mathcal{K} \subset \mathcal{P}(M)$. Then the following are equivalent:*

(i) $\mathcal{K}$ is tight.

(ii) Every sequence in $\mathcal{K}$ has a weakly convergent subsequence.

In other words, relative compactness in the weak topology is equivalent to tightness.

Proof. We only indicate the idea. If $\mathcal{K}$ is relatively compact, then it must be tight; otherwise one could construct a sequence whose mass escapes every compact set, contradicting weak compactness via the Portmanteau theorem.

Conversely, assume $\mathcal{K}$ is tight and let (μ_n) be a sequence in $\mathcal{K}$. Tightness gives enough compact containment to construct a subsequence which converges on a countable convergence-determining class of test functions. By separability of M, one can choose such a countable class, and the limiting functional is then represented by a probability measure. This yields a weakly convergent subsequence. □

A very important consequence is that the weak topology on $\mathcal{P}(M)$ can be metrized in a natural way.

Definition 5.9 (Prokhorov metric)**.** For $\mu, \nu \in \mathcal{P}(M)$, define

$$\pi(\mu, \nu) := \inf\left\{\varepsilon > 0 : \mu(A) \le \nu(A^\varepsilon) + \varepsilon \text{ for all Borel sets } A \subset M\right\},$$

where

$$A^\varepsilon := \{x \in M : d(x, A) < \varepsilon\}.$$

Remark 5.10. One can show that π is a metric on $\mathcal{P}(M)$ and that the topology induced by π is exactly the weak topology.

Theorem 5.11. *If M is Polish, then $\mathcal{P}(M)$ endowed with the weak topology is again a Polish space. More precisely, the Prokhorov metric makes $\mathcal{P}(M)$ into a Polish space.*

Proof. The fact that the Prokhorov metric induces the weak topology is standard. Separability of $\mathcal{P}(M)$ follows from separability of M by approximating probability measures with finitely supported rational measures on a countable dense subset of M.

To prove completeness, let (μ_n) be a Cauchy sequence in the Prokhorov metric. Then the family $\{\mu_n : n \geq 1\}$ is tight, hence by Prokhorov's theorem every subsequence has a weakly convergent subsequence. Since (μ_n) is Cauchy, all such subsequential limits must coincide, and therefore the whole sequence converges weakly. Hence $\mathcal{P}(M)$ is complete. This proves that $\mathcal{P}(M)$ is Polish. □

Remark 5.12. The previous theorem is extremely useful: once one works on a Polish state space, the corresponding space of probability measures has again a good topological structure. In particular, one can iterate the construction, for example when studying random probability measures.

Definition 5.13. Let (M, ρ) be a Polish space, let $x_0 \in M$, and let $p \in [1, \infty)$. We define

$$\mathcal{P}_p(M) := \left\{ \mu \in \mathcal{P}(M) : \int_M \rho(x, x_0)^p \, d\mu(x) < \infty \right\}.$$

For $\mu, \mu_1, \mu_2, \cdots \in \mathcal{P}_p(M)$, we say that μ_n converges to μ in the *p-weak sense* if

$$\int_M f \, d\mu_n \to \int_M f \, d\mu$$

for every continuous function $f : M \to \mathbb{R}$ such that for some constant $C < \infty$,

$$|f(x)| \leq C\big(1 + \rho(x, x_0)^p\big) \qquad \text{for all } x \in M.$$

Remark 5.14. This notion is independent of the choice of the reference point x_0. It is also the topology usually associated with the p-Wasserstein distance, a point of view which we discuss below.

Theorem 5.15. *Let $\mu, \mu_1, \mu_2, \cdots \in \mathcal{P}_p(M)$. Then the following are equivalent:*

(i) μ_n converges to μ in the p-weak sense.

(ii) $\mu_n \Rightarrow \mu$ and

$$\int_M \rho(x, x_0)^p \, d\mu_n(x) \to \int_M \rho(x, x_0)^p \, d\mu(x).$$

Proof. Assume first (1). Since every bounded continuous function satisfies the growth condition in Definition 5.13, we immediately get $\mu_n \Rightarrow \mu$. Taking $f(x) = \rho(x, x_0)^p$, which is continuous and has admissible growth, also gives convergence of the p-th moments.

Assume now (2), and let $f : M \to \mathbb{R}$ be continuous with

$$|f(x)| \leq C\big(1 + \rho(x, x_0)^p\big).$$

We must show that $\int f\,d\mu_n \to \int f\,d\mu$.

Set $r(x) := \rho(x, x_0)^p$, and for $R > 0$ define the bounded continuous truncation

$$r_R(x) := r(x) \wedge R.$$

Since $\mu_n \Rightarrow \mu$, we have $\int r_R\,d\mu_n \to \int r_R\,d\mu$ for every $R > 0$. Therefore

$$\limsup_{n\to\infty} \int_M (r - r_R)\,d\mu_n = \limsup_{n\to\infty} \left(\int_M r\,d\mu_n - \int_M r_R\,d\mu_n \right) = \int_M r\,d\mu - \int_M r_R\,d\mu.$$

Letting $R \to \infty$ and using monotone convergence on the right-hand side, we obtain

$$\lim_{R\to\infty} \limsup_{n\to\infty} \int_{\{r>R\}} r\,d\mu_n = 0.$$

Since the p-th moments converge, they are bounded, so also $\sup_n \int r\,d\mu_n < \infty$.

Next choose a continuous cut-off function $\chi_R : M \to [0,1]$ such that $\chi_R(x) = 1$ when $\rho(x, x_0) \le R$ and $\chi_R(x) = 0$ when $\rho(x, x_0) \ge R + 1$, and define $f_R := f\chi_R$. Then f_R is bounded and continuous, so weak convergence gives

$$\int_M f_R\,d\mu_n \to \int_M f_R\,d\mu.$$

It remains to control the tails. Since $1 - \chi_R$ vanishes on $\{\rho(\cdot, x_0) \le R\}$,

$$|f - f_R| \le C(1 + r)\mathbf{1}_{\{r>R^p\}}.$$

On the set $\{r > R^p\}$ we have $1 \le R^{-p} r$, hence

$$(1 + r)\mathbf{1}_{\{r>R^p\}} \le (1 + R^{-p})\, r\, \mathbf{1}_{\{r>R^p\}}.$$

Therefore

$$\sup_n \int_M |f - f_R|\,d\mu_n \le C(1 + R^{-p}) \sup_n \int_{\{r>R^p\}} r\,d\mu_n \to 0 \qquad \text{as } R \to \infty.$$

The same argument gives $\int |f - f_R|\,d\mu \to 0$.

Hence

$$\left| \int_M f\,d\mu_n - \int_M f\,d\mu \right| \le \int_M |f - f_R|\,d\mu_n + \left| \int_M f_R\,d\mu_n - \int_M f_R\,d\mu \right| + \int_M |f_R - f|\,d\mu.$$

First let $n \to \infty$, so the middle term vanishes, and then let $R \to \infty$, so the two tail terms vanish. This proves $\int f\,d\mu_n \to \int f\,d\mu$. □

5.2 Other modes of convergence on metric spaces

Let (M, ρ) be a Polish space, and let $X, X_1, X_2, \ldots$ be M-valued random variables defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The following definitions are most natural in the case where M is a $\mathbb{R}^d$ or a Hilbertspace but we note the definitions are meaningful in the generality of Polish spaces and the proofs of the basic results below are in fact not more complicated.

Definition 5.16. We say that $X_n \xrightarrow[n\to\infty]{a.s.} X$ if

$$\rho(X_n, X) \xrightarrow[n\to\infty]{} 0 \qquad \text{almost surely.}$$

We say that $X_n \xrightarrow[n\to\infty]{\mathbb{P}} X$ if for every $\varepsilon > 0$,

$$\mathbb{P}(\rho(X_n, X) > \varepsilon) \xrightarrow[n\to\infty]{} 0.$$

If $p \in [1, \infty)$, we say that $X_n \xrightarrow[n\to\infty]{L^p} X$ if

$$\mathbb{E}\big(\rho(X_n, X)^p\big) \xrightarrow[n\to\infty]{} 0.$$

Remark 5.17. If one fixes a point $x_0 \in M$, one may define

$$L^p(\Omega, \mathcal{F}, \mathbb{P}; M) := \Big\{ X : \Omega \to M \text{ measurable} : \ \mathbb{E}(\rho(X, x_0)^p) < \infty \Big\}.$$

This definition does not depend on the choice of x_0. Unlike in the real-valued case, this is in general not a vector space unless M itself carries a linear structure. What matters here is only the metric notion of L^p convergence.

Proposition 5.18. *The following implications hold:*

$$X_n \xrightarrow[n\to\infty]{a.s.} X \Longrightarrow X_n \xrightarrow[n\to\infty]{\mathbb{P}} X, \qquad X_n \xrightarrow[n\to\infty]{L^p} X \Longrightarrow X_n \xrightarrow[n\to\infty]{\mathbb{P}} X.$$

In general, neither converse holds.

Proof. The proof is the same as on the real line. If $X_n \to X$ almost surely, then the indicators $\mathbb{1}_{\rho(X_n,X)>\varepsilon}$ converge almost surely to 0 for every $\varepsilon > 0$, and are bounded by 1, so dominated convergence gives convergence in probability.

If $X_n \to X$ in L^p, then Markov's inequality gives

$$\mathbb{P}(\rho(X_n, X) > \varepsilon) \le \frac{\mathbb{E}(\rho(X_n, X)^p)}{\varepsilon^p},$$

which tends to 0.

The converses already fail for real-valued random variables, and hence also in the present generality. □

5.3 Convergence in distribution of random variables

Let (M, ρ) be a Polish space.

Definition 5.19 (Convergence in distribution)**.** Let $X, X_1, X_2, \dots$ be M-valued random variables. We say that X_n converges *in distribution* to X, and write

$$X_n \xrightarrow[n\to\infty]{d} X,$$

if

$$\mathcal{L}(X_n) \Rightarrow \mathcal{L}(X).$$

Equivalently,

$$\mathbb{E}(f(X_n)) \to \mathbb{E}(f(X))$$

for every bounded continuous function $f : M \to \mathbb{R}$.

Remark 5.20. Convergence in distribution only depends on the laws of the random variables, not on the probability space on which they are defined.

Proposition 5.21. *If* $X_n \xrightarrow[n\to\infty]{\mathbb{P}} X$*, then* $X_n \xrightarrow[n\to\infty]{d} X$*.*

Proof. Let $f : M \to \mathbb{R}$ be bounded and continuous. Then $f(X_n) \xrightarrow[n\to\infty]{\mathbb{P}} f(X)$. Since the family $(f(X_n))_{n\geq 1}$ is uniformly bounded, it is uniformly integrable. Hence $f(X_n) \to f(X)$ in L^1, so $\mathbb{E}(f(X_n)) \to \mathbb{E}(f(X))$. □

An analogous result holds for p-weak convergence:

Theorem 5.22. *Let* (M, ρ) *be a Polish space, let* $x_0 \in M$*, let* $p \in [1, \infty)$*, and let* $X, X_1, X_2, \ldots$ *be* M*-valued random variables on the same probability space such that*

$$\mathbb{E}\big(\rho(X_n, X)^p\big) \xrightarrow[n\to\infty]{} 0.$$

Then the laws $\mathcal{L}(X_n)$ *and* $\mathcal{L}(X)$ *belong to* $\mathcal{P}_p(M)$*, and*

$$\mathcal{L}(X_n) \to \mathcal{L}(X)$$

in the p*-weak sense.*

Proof. Since $X_n \xrightarrow[n\to\infty]{L^p} X$, Proposition 5.18 gives $X_n \xrightarrow[n\to\infty]{\mathbb{P}} X$. Hence Proposition 5.21 yields

$$\mathcal{L}(X_n) \Rightarrow \mathcal{L}(X).$$

It remains to show convergence of the p-th moments. First, since

$$\rho(X_n, x_0) \leq \rho(X_n, X) + \rho(X, x_0),$$

we have

$$\mathbb{E}\big(\rho(X_n, x_0)^p\big) < \infty$$

for all n, so indeed $\mathcal{L}(X_n), \mathcal{L}(X) \in \mathcal{P}_p(M)$.

Next, by the reverse triangle inequality in L^p,

$$\big|\|\rho(X_n, x_0)\|_{L^p} - \|\rho(X, x_0)\|_{L^p}\big| \leq \|\rho(X_n, x_0) - \rho(X, x_0)\|_{L^p}.$$

Since

$$|\rho(X_n, x_0) - \rho(X, x_0)| \leq \rho(X_n, X),$$

it follows that

$$\|\rho(X_n, x_0) - \rho(X, x_0)\|_{L^p} \leq \|\rho(X_n, X)\|_{L^p} \xrightarrow[n\to\infty]{} 0.$$

Therefore

$$\|\rho(X_n, x_0)\|_{L^p} \xrightarrow[n\to\infty]{} \|\rho(X, x_0)\|_{L^p},$$

and hence

$$\mathbb{E}\big(\rho(X_n, x_0)^p\big) \xrightarrow[n\to\infty]{} \mathbb{E}\big(\rho(X, x_0)^p\big).$$

By Theorem 5.15, this implies that $\mathcal{L}(X_n) \to \mathcal{L}(X)$ in the p-weak sense. □

The next theorem is the most important basic tool.

Theorem 5.23 (Continuous mapping theorem)**.** *Let X_n, X be M-valued random variables such that $X_n \xrightarrow[n\to\infty]{d} X$, and let $g : M \to N$ be a measurable map into another Polish space N. Assume that the set of discontinuity points of g has X-probability zero. Then*

$$g(X_n) \xrightarrow[n\to\infty]{d} g(X).$$

In particular, if g is continuous, then $g(X_n) \xrightarrow[n\to\infty]{d} g(X)$.

Proof. If $f : N \to \mathbb{R}$ is bounded and continuous, then $f \circ g$ is bounded and measurable, and it is continuous at every point where g is continuous. The claim therefore follows from the Portmanteau theorem. □

We record next the standard algebraic consequence on Euclidean spaces.

Theorem 5.24 (Slutsky's theorem)**.** *Let $X_n \xrightarrow[n\to\infty]{d} X$ in $\mathbb{R}^d$ and let $Y_n \xrightarrow[n\to\infty]{\mathbb{P}} c$ in $\mathbb{R}^m$, where $c \in \mathbb{R}^m$ is deterministic. Then*

$$(X_n, Y_n) \xrightarrow[n\to\infty]{d} (X, c).$$

Consequently, if $g : \mathbb{R}^{d+m} \to \mathbb{R}^k$ is continuous, then

$$g(X_n, Y_n) \xrightarrow[n\to\infty]{d} g(X, c).$$

In particular,

$$X_n + Y_n \xrightarrow[n\to\infty]{d} X + c, \qquad X_n - Y_n \xrightarrow[n\to\infty]{d} X - c,$$

and in one dimension,

$$X_n Y_n \xrightarrow[n\to\infty]{d} cX.$$

If moreover $c \neq 0$, then

$$\frac{X_n}{Y_n} \xrightarrow[n\to\infty]{d} \frac{X}{c}.$$

Proof. Let $f : \mathbb{R}^{d+m} \to \mathbb{R}$ be bounded and continuous. Since $X_n \xrightarrow[n\to\infty]{d} X$, the family of laws of X_n is tight. Fix $\varepsilon > 0$. Choose a compact set $K \subset \mathbb{R}^d$ such that $\mathbb{P}(X_n \in K) \geq 1 - \varepsilon$ for all n. Since f is uniformly continuous on the compact set $K \times \overline{B}(c, 1)$, there exists $\delta > 0$ such that

$$|f(x, y) - f(x, c)| < \varepsilon$$

whenever $x \in K$ and $|y - c| < \delta$. It follows that

$$\big|\mathbb{E}(f(X_n, Y_n)) - \mathbb{E}(f(X_n, c))\big|$$

is bounded by

$$\varepsilon + 2\|f\|_\infty \, \mathbb{P}(X_n \notin K) + 2\|f\|_\infty \, \mathbb{P}(|Y_n - c| \geq \delta),$$

which tends to at most $(1 + 2\|f\|_\infty)\varepsilon$ as $n \to \infty$. Since ε was arbitrary, we get

$$\mathbb{E}(f(X_n, Y_n)) - \mathbb{E}(f(X_n, c)) \to 0.$$

On the other hand, the map $x \mapsto f(x, c)$ is bounded and continuous, so

$$\mathbb{E}(f(X_n, c)) \to \mathbb{E}(f(X, c)).$$

Hence $(X_n, Y_n) \xrightarrow[n\to\infty]{d} (X, c)$, and the remaining claims follow from the continuous mapping theorem. □

We now turn to a useful representation theorem on the real line.

Definition 5.25 (Quantile function)**.** Let $\mu \in \mathcal{P}(\mathbb{R})$ with distribution function

$$F(x) := \mu((-\infty, x]).$$

Its *quantile function* is the generalized inverse

$$Q(u) := \inf\{x \in \mathbb{R} : F(x) \geq u\}, \qquad 0 < u < 1.$$

Lemma 5.26. *Let U be uniformly distributed on $(0, 1)$, and let Q be the quantile function associated with a probability measure μ on $\mathbb{R}$. Then $Q(U)$ has law μ.*

Proof. Let F be the distribution function of μ. For every $x \in \mathbb{R}$,

$$Q(u) \leq x \Leftrightarrow u \leq F(x).$$

Hence

$$\mathbb{P}(Q(U) \leq x) = \mathbb{P}(U \leq F(x)) = F(x),$$

so $Q(U)$ has distribution function F, i.e. law μ. □

Theorem 5.27 (Skorokhod representation theorem on the real line)**.** *Let X_n, X be real-valued random variables such that*

$$X_n \xrightarrow[n\to\infty]{d} X.$$

Then there exist real-valued random variables $\widetilde{X}_n, \widetilde{X}$ on the probability space $((0, 1), \mathcal{B}((0, 1)), \lambda)$ such that

$$\mathcal{L}(\widetilde{X}_n) = \mathcal{L}(X_n), \qquad \mathcal{L}(\widetilde{X}) = \mathcal{L}(X),$$

and

$$\widetilde{X}_n \xrightarrow[n\to\infty]{a.s.} \widetilde{X}.$$

Proof. Let F_n and F be the distribution functions of X_n and X, and let Q_n and Q be the corresponding quantile functions. By Lemma 5.26, if we define

$$\widetilde{X}_n(u) := Q_n(u), \qquad \widetilde{X}(u) := Q(u), \qquad 0 < u < 1,$$

then $\widetilde{X}_n$ and $\widetilde{X}$ have the desired laws.

It remains to prove pointwise convergence for almost every u. Let $u \in (0,1)$ be a continuity point of Q. We claim that $Q_n(u) \to Q(u)$. Fix $\varepsilon > 0$. Since Q is continuous at u, we may choose continuity points x_-, x_+ of F such that

$$Q(u) - \varepsilon < x_- < Q(u) < x_+ < Q(u) + \varepsilon$$

and

$$F(x_-) < u < F(x_+).$$

By Theorem 5.5, we have $F_n(x_\pm) \to F(x_\pm)$, and hence for all large n,

$$F_n(x_-) < u < F_n(x_+).$$

By the definition of the quantile function, this implies

$$x_- < Q_n(u) \le x_+.$$

Therefore $|Q_n(u) - Q(u)| < \varepsilon$ for all large n, so $Q_n(u) \to Q(u)$.

Since Q is monotone, it has at most countably many discontinuity points. Hence $Q_n(u) \to Q(u)$ for Lebesgue-a.e. $u \in (0,1)$, which proves the theorem. □

Remark 5.28. The previous theorem gives an additional characterization of weak convergence on the real line: if $\mu_n, \mu \in \mathcal{P}(\mathbb{R})$ have quantile functions Q_n, Q, then

$$\mu_n \Rightarrow \mu$$

if and only if

$$Q_n(u) \to Q(u)$$

for Lebesgue-a.e. $u \in (0,1)$, equivalently, for every continuity point u of Q.

Theorem 5.29 (Skorokhod representation theorem)**.** *Let M be a Polish space, and let X_n, X be M-valued random variables such that*

$$X_n \xrightarrow[n\to\infty]{d} X.$$

Then there exist M-valued random variables $\widetilde{X}_n, \widetilde{X}$ on a common probability space such that

$$\mathcal{L}(\widetilde{X}_n) = \mathcal{L}(X_n), \qquad \mathcal{L}(\widetilde{X}) = \mathcal{L}(X),$$

and

$$\widetilde{X}_n \xrightarrow[n\to\infty]{a.s.} \widetilde{X}.$$

Proof. We first prove the result on the Hilbert cube

$$K := [0,1]^{\mathbb{N}},$$

equipped with the metric

$$d(x,y) := \sum_{m=1}^{\infty} 2^{-m} |x_m - y_m|.$$

Let $\nu_n, \nu \in \mathcal{P}(K)$ and assume $\nu_n \Rightarrow \nu$. For every $k \geq 1$, choose an integer m_k so large that $\sum_{m>m_k} 2^{-m} < 2^{-k-1}$. For each $1 \leq m \leq m_k$, choose a finite partition of $[0,1]$ into half-open intervals with mesh smaller than 2^{-k-1}, all whose endpoints are continuity points of the mth marginal of ν, and such that the partitions refine as k increases. These one-dimensional partitions induce a finite partition $\mathcal{P}_k$ of K into half-open cylinders depending only on the first m_k coordinates. The atoms of $\mathcal{P}_k$ are continuity sets of ν, the partition $\mathcal{P}_{k+1}$ refines $\mathcal{P}_k$, and every atom of $\mathcal{P}_k$ has diameter at most 2^{-k}.

Now fix $\sigma \in \mathcal{P}(K)$. Enumerate the atoms of $\mathcal{P}_k$ as $A_{k,1}, \ldots, A_{k,N_k}$ in a way compatible with refinement, and split $[0,1)$ into half-open intervals $I^\sigma_{k,1}, \ldots, I^\sigma_{k,N_k}$ of lengths $\sigma(A_{k,j})$, again compatibly with refinement. For $u \in [0,1)$, let $A^\sigma_k(u)$ be the unique atom such that $u \in I^\sigma_{k,j}$. Since the partitions refine, the sets $A^\sigma_k(u)$ are nested, and since their diameters tend to zero, their intersection contains exactly one point. Denote this point by $T_\sigma(u)$. The map $T_\sigma : [0,1) \to K$ is measurable, and by construction

$$\lambda(T_\sigma \in A) = \sigma(A)$$

for every atom A of every partition $\mathcal{P}_k$. Since these atoms generate the Borel σ-algebra of K, it follows that $\mathcal{L}(T_\sigma) = \sigma$.

We now set $\widetilde{Y}_n := T_{\nu_n}$ and $\widetilde{Y} := T_\nu$ on the probability space $([0,1), \mathcal{B}([0,1)), \lambda)$. Then $\mathcal{L}(\widetilde{Y}_n) = \nu_n$ and $\mathcal{L}(\widetilde{Y}) = \nu$. It remains to show that $\widetilde{Y}_n \to \widetilde{Y}$ almost surely. Fix $k \geq 1$. Since every atom of $\mathcal{P}_k$ is a continuity set of ν, we have $\nu_n(A) \to \nu(A)$ for each atom A of $\mathcal{P}_k$. Hence the endpoints of the intervals $I^{\nu_n}_{k,j}$ converge to those of $I^\nu_{k,j}$. Therefore, for every $u \in [0,1)$ outside the finite set of endpoints of the intervals $I^\nu_{k,j}$, the atom $A^{\nu_n}_k(u)$ eventually equals $A^\nu_k(u)$. Since this holds for every fixed k, and the union of all these endpoint sets is countable, we obtain that for Lebesgue-a.e. u and every k,

$$d(\widetilde{Y}_n(u), \widetilde{Y}(u)) \leq 2^{-k}$$

for all large enough n. Hence $\widetilde{Y}_n(u) \to \widetilde{Y}(u)$ for Lebesgue-a.e. u.

We now pass to a general Polish space M. By the classical Alexandrov theorem, there exist a G_δ subset $G \subset K$ and a homeomorphism $\varphi : M \to G$. Let

$$\nu_n := \varphi_\# \mathcal{L}(X_n), \qquad \nu := \varphi_\# \mathcal{L}(X).$$

Since φ is continuous, we have $\nu_n \Rightarrow \nu$. By the cube case, there exist K-valued random variables $\widetilde{Y}_n, \widetilde{Y}$ on a common probability space such that $\mathcal{L}(\widetilde{Y}_n) = \nu_n$, $\mathcal{L}(\widetilde{Y}) = \nu$, and $\widetilde{Y}_n \to \widetilde{Y}$ almost surely. As $\nu_n(G) = \nu(G) = 1$, we have $\widetilde{Y}_n, \widetilde{Y} \in G$ almost surely. Defining

$$\widetilde{X}_n := \varphi^{-1}(\widetilde{Y}_n), \qquad \widetilde{X} := \varphi^{-1}(\widetilde{Y}),$$

we obtain $\mathcal{L}(\widetilde{X}_n) = \mathcal{L}(X_n)$, $\mathcal{L}(\widetilde{X}) = \mathcal{L}(X)$, and $\widetilde{X}_n \to \widetilde{X}$ almost surely, since φ^{-1} is continuous on G. □

Remark 5.30. On the real line, the proof becomes particularly simple: the copies may be taken to be the quantile functions on the unit interval, as in Theorem 5.27. The general theorem shows that the same almost sure representation is available on every Polish space.

5.4 p-Wasserstein distance on a Hilbert space

So far, we have considered two complementary points of view: random variables on the one hand, and their laws on the other. The p-Wasserstein distance is one of the places where these two perspectives fit together especially well. On the measure side, it is defined by minimizing a transportation cost over all couplings of two probability measures. Later we will see that the same quantity can equivalently be written as an infimum over all pairs of random variables X and Y with prescribed laws, of the L^p-distance between X and Y.

Throughout this section, let H be a separable Hilbert space with norm $\|\cdot\|$. In fact, in most applications one is interested in the case $H = \mathbb{R}^d$, but assuming this would not make the proofs of the results below easier.[1] We also fix an auxiliary atomless Polish probability space $(\Omega, \mathcal{F}, \mathbb{P})$ in the background, which will be useful later when we pass to random variable representations.[2]

For $p \in [1, \infty)$, we write

$$\mathcal{P}_p(H) := \left\{ \mu \in \mathcal{P}(H) : \int_H \|x\|^p \, d\mu(x) < \infty \right\}.$$

Definition 5.31 (Transport plans)**.** Let $\mu, \nu \in \mathcal{P}(H)$. A *transport plan*, or *coupling*, of μ and ν is a probability measure $\pi \in \mathcal{P}(H \times H)$ such that

$$\pi(A \times H) = \mu(A), \qquad \pi(H \times B) = \nu(B)$$

for all Borel sets $A, B \subset H$. The set of all such couplings is denoted by $\Pi(\mu, \nu)$.

Remark 5.32. The set $\Pi(\mu, \nu)$ is never empty, since the product measure $\mu \otimes \nu$ belongs to it.

Lemma 5.33. *For every $\mu, \nu \in \mathcal{P}(H)$, the set $\Pi(\mu, \nu)$ is compact for the weak topology on $\mathcal{P}(H \times H)$.*

Proof. Since H is Polish, so is $H \times H$. We first show that $\Pi(\mu, \nu)$ is tight. Let $\varepsilon > 0$. Since μ and ν are tight, there exist compact sets $K, L \subset H$ such that $\mu(K) \geq 1 - \varepsilon/2$ and $\nu(L) \geq 1 - \varepsilon/2$. Then for every $\pi \in \Pi(\mu, \nu)$,

$$\pi(K \times L) \geq 1 - \pi(K^c \times H) - \pi(H \times L^c) = 1 - \mu(K^c) - \nu(L^c) \geq 1 - \varepsilon.$$

Thus $\Pi(\mu, \nu)$ is tight, hence relatively compact by Prokhorov's theorem.

It remains to show that $\Pi(\mu, \nu)$ is closed. Let $\pi_n \in \Pi(\mu, \nu)$ and assume $\pi_n \Rightarrow \pi$. If $f : H \to \mathbb{R}$ is bounded and continuous, then $(x, y) \mapsto f(x)$ and $(x, y) \mapsto f(y)$ are bounded and continuous on $H \times H$. Therefore

$$\int_{H \times H} f(x) \, d\pi = \lim_{n \to \infty} \int_{H \times H} f(x) \, d\pi_n = \int_H f \, d\mu,$$

and similarly $\int_{H \times H} f(y) \, d\pi = \int_H f \, d\nu$. Hence the first marginal of π is μ and the second marginal is ν, so $\pi \in \Pi(\mu, \nu)$. Thus $\Pi(\mu, \nu)$ is closed, and therefore compact. □

[1] In fact, the same results with same proofs can also be established for a Polish metric space instead of H.

[2] While one usually does not fix the probability space in the background, it is useful to do so in the present context, due to the interplay of $L^p(\Omega, \mathcal{F}, \mathbb{P}; H)$ and $\mathcal{P}^p(\Omega, \mathcal{F}, \mathbb{P}; H)$.

Definition 5.34 (p-Wasserstein distance)**.** Let $p \in [1,\infty)$ and let $\mu,\nu \in \mathcal{P}_p(H)$. The *$p$-Wasserstein distance* between μ and ν is defined by

$$W_p(\mu,\nu) := \left(\inf_{\pi\in\Pi(\mu,\nu)} \int_{H\times H} \|x-y\|^p \, d\pi(x,y)\right)^{1/p}.$$

Remark 5.35. The quantity above is finite. Indeed, $\mu\otimes\nu \in \Pi(\mu,\nu)$, and

$$\|x-y\|^p \le 2^{p-1}\big(\|x\|^p + \|y\|^p\big),$$

so

$$\int_{H\times H} \|x-y\|^p \, d(\mu\otimes\nu)(x,y) \le 2^{p-1}\left(\int_H \|x\|^p \, d\mu(x) + \int_H \|y\|^p \, d\nu(y)\right) < \infty.$$

Proposition 5.36. *For fixed $p \in [1,\infty)$ and $\mu,\nu \in \mathcal{P}_p(H)$, the map*

$$\pi \mapsto \int_{H\times H} \|x-y\|^p \, d\pi(x,y)$$

is lower semicontinuous on $\mathcal{P}(H\times H)$ for the weak topology. Consequently, the infimum in the definition of $W_p(\mu,\nu)$ is attained: there exists $\pi^* \in \Pi(\mu,\nu)$ such that

$$W_p(\mu,\nu)^p = \int_{H\times H} \|x-y\|^p \, d\pi^*(x,y).$$

Proof. Let $c(x,y) := \|x-y\|^p$. Then c is continuous and non-negative on $H\times H$. For $R>0$, define the truncation $c_R(x,y) := c(x,y)\wedge R$. Each c_R is bounded and continuous, and $c_R \uparrow c$ as $R\to\infty$. Hence $\int c\,d\pi = \sup_{R>0}\int c_R\,d\pi$. Since $\pi \mapsto \int c_R\,d\pi$ is continuous for every R, it follows that $\pi\mapsto\int c\,d\pi$ is lower semicontinuous as a supremum of continuous maps.

By Lemma 5.33, the set $\Pi(\mu,\nu)$ is compact. Therefore the lower semicontinuous functional $\pi\mapsto\int c\,d\pi$ attains its minimum on $\Pi(\mu,\nu)$. □

Definition 5.37 (Independent randomization)**.** Let $X:\Omega\to H$ be a random variable. We say that X *allows for independent randomization* if there exists a random variable $U:\Omega\to[0,1]$ such that U is uniformly distributed on $[0,1]$ and independent of X.

Lemma 5.38 (Realizing couplings by independent randomization)**.** *Let $\mu,\nu\in\mathcal{P}(H)$, let $X:\Omega\to H$ be a random variable with law μ, and assume that X allows for independent randomization. Then for every $\pi\in\Pi(\mu,\nu)$ there is a random variable $Y:\Omega\to H$ such that*

$$\mathcal{L}(X,Y) = \pi.$$

Proof. Let U be a $[0,1]$-uniform random variable independent of X. By disintegration, there exists a probability kernel $K: H\times\mathcal{B}(H)\to[0,1]$ such that

$$\pi(dx,dy) = \mu(dx)\,K(x,dy).$$

Choose a Borel isomorphism $\varphi: H\to[0,1]$. For each $x\in H$, let η_x be the pushforward of $K(x,\cdot)$ under φ, that is,

$$\eta_x(A) := K(x,\varphi^{-1}(A)), \qquad A\in\mathcal{B}([0,1]).$$

Let

$$F(x,t) := \eta_x([0,t]), \qquad (x,t) \in H \times [0,1],$$

and define the corresponding quantile map by

$$Q(x,u) := \inf\{t \in [0,1] : F(x,t) \geq u\}, \qquad (x,u) \in H \times [0,1].$$

Since $x \mapsto F(x,t)$ is Borel for every fixed t, the map Q is Borel measurable. Now define

$$Y := \varphi^{-1}(Q(X,U)).$$

Fix $x \in H$. Since U is uniform on $[0,1]$, the quantile construction on the real line shows that the conditional law of $Q(x,U)$ is η_x. Therefore the conditional law of Y given $X = x$ is $K(x,\cdot)$. □

Proposition 5.39. *Let $p \in [1,\infty)$, let $\mu, \nu \in \mathcal{P}_p(H)$, and let $X : \Omega \to H$ be a random variable with law μ. Assume that X allows for independent randomization. Then*

$$W_p(\mu,\nu) = \min\left\{\|X - Y\|_{L^p(\Omega;H)} : \ Y : \mathcal{L}(Y) = \nu\right\}.$$

Proof. Let $Y : \Omega \to H$ be any random variable with law ν. Then the law of (X,Y) is a coupling of μ and ν. So by the definition of W_p, $W_p(\mu,\nu)^p \leq \inf\left\{\|X - Y\|_{L^p(\Omega;H)} : \ Y : \mathcal{L}(Y) = \nu\right\}$. Conversely, by Proposition 5.36, there exists an optimal coupling $\pi^* \in \Pi(\mu,\nu)$ such that $W_p(\mu,\nu)^p = \int_{H\times H} \|x-y\|^p \, d\pi^*(x,y)$. Since X allows for independent randomization, Lemma 5.38 yields a random variable $Y : \Omega \to H$ such that $\mathcal{L}(X,Y) = \pi^*$. Hence $\mathcal{L}(Y) = \nu$ and $\mathbb{E}\big(\|X-Y\|^p\big) = \int_{H\times H} \|x-y\|^p \, d\pi^*(x,y) = W_p(\mu,\nu)^p$. This proves the reverse inequality and shows that the infimum is in fact attained. □

Remark 5.40. Thus, whenever one may keep the first marginal fixed and still realize arbitrary couplings by independent randomization, the Wasserstein distance can be viewed directly as the smallest possible L^p-distance to a random variable with the prescribed target law.

Theorem 5.41. *For every $p \in [1,\infty)$, the W_p defines a complete separable metric on $\mathcal{P}_p(H)$.*

Proof. Non-negativity and symmetry are immediate. If $W_p(\mu,\nu) = 0$, then by Proposition 5.36 there is an optimal coupling $\pi^* \in \Pi(\mu,\nu)$ with $\int \|x-y\|^p \, d\pi^* = 0$. Since the integrand is non-negative, π^* is concentrated on the diagonal, and therefore $\mu = \nu$.

For the triangle inequality, let $\mu,\nu,\eta \in \mathcal{P}_p(H)$, and choose a random variable Y with law ν which allows for independent randomization. By Proposition 5.39, applied twice with this fixed middle variable Y, there exist random variables X and Z with laws μ and η such that $\|X - Y\|_{L^p} = W_p(\mu,\nu)$ and $\|Y - Z\|_{L^p} = W_p(\nu,\eta)$. Since the law of (X,Z) is a coupling of μ and η, we have $W_p(\mu,\eta) \leq \|X - Z\|_{L^p}$. The triangle inequality for L^p then yields

$$W_p(\mu,\eta) \leq W_p(\mu,\nu) + W_p(\nu,\eta).$$

We next prove completeness. Let (μ_n) be a W_p-Cauchy sequence. Passing to a subsequence, we may assume that $W_p(\mu_n,\mu_{n+1}) \leq 2^{-n}$ for all n. Choose X_1 with law μ_1. Recursively, Proposition 5.39 yields random variables X_n with law μ_n such that $\|X_n - X_{n+1}\|_{L^p} = W_p(\mu_n,\mu_{n+1}) \leq 2^{-n}$. Hence (X_n) is Cauchy in $L^p(\Omega;H)$, so by completeness of $L^p(\Omega;H)$

it converges in L^p to some random variable X. Let $\mu := \mathcal{L}(X)$. Since the law of (X_n, X) is a coupling of μ_n and μ, we get $W_p(\mu_n, \mu) \le \|X_n - X\|_{L^p} \to 0$. Thus the subsequence converges to μ in W_p, and since the original sequence was W_p-Cauchy, the whole sequence converges to μ.

Finally, separability follows in the usual way: if $D = \{x_1, x_2, \dots\}$ is a countable dense subset of H, then the finitely supported probability measures with rational weights on points of D form a countable dense subset of $\mathcal{P}_p(H)$.

Therefore W_p is a complete separable metric on $\mathcal{P}_p(H)$. □

Remark 5.42 (Wasserstein distance as an extension of the Hilbert space distance)**.** The Wasserstein distance extends the original metric structure of the Hilbert space H to the space $\mathcal{P}_p(H)$ of probability measures. Indeed, if we identify a point $x \in H$ with the Dirac mass δ_x, then for all $x, y \in H$ we have

$$W_p(\delta_x, \delta_y) = \|x - y\|.$$

Thus $\mathcal{P}_p(H)$ may be viewed as a nonlinear enlargement of H in which points are replaced by probability distributions.

This point of view also shows that there are natural geodesics in Wasserstein space. Let $\mu_0, \mu_1 \in \mathcal{P}_p(H)$, and choose random variables X_0, X_1 such that $\mathcal{L}(X_0) = \mu_0$, $\mathcal{L}(X_1) = \mu_1$, and

$$\|X_1 - X_0\|_{L^p} = W_p(\mu_0, \mu_1).$$

For $t \in [0, 1]$, define

$$\mu_t := \mathcal{L}\big((1-t)X_0 + tX_1\big).$$

Then $(\mu_t)_{t\in[0,1]}$ is a constant-speed geodesic from μ_0 to μ_1. More precisely, for all $0 \le s \le t \le 1$,

$$W_p(\mu_s, \mu_t) = (t - s)\, W_p(\mu_0, \mu_1).$$

Indeed, the law of the pair $((1-s)X_0 + sX_1, (1-t)X_0 + tX_1)$ is a coupling of μ_s and μ_t, so

$$W_p(\mu_s, \mu_t) \le \|(t-s)(X_1 - X_0)\|_{L^p} = (t-s)W_p(\mu_0, \mu_1).$$

On the other hand, the triangle inequality gives

$$W_p(\mu_0, \mu_1) \le W_p(\mu_0, \mu_s) + W_p(\mu_s, \mu_t) + W_p(\mu_t, \mu_1),$$

and the two outer terms are bounded above by $sW_p(\mu_0, \mu_1)$ and $(1-t)W_p(\mu_0, \mu_1)$ by the same coupling argument. Hence equality must hold everywhere.

In this sense, Wasserstein space inherits not only a metric structure from the Hilbert space, but also a rich geometry. In fact, one can go much further and exploit the linear and differential structure of H to develop a kind of differential calculus on $\mathcal{P}_p(H)$. This leads to the so-called Otto calculus and, in a related direction, to the Lions derivative. These ideas have played a central role in many developments of the last two decades.

Theorem 5.43. *Let $p \in [1, \infty)$ and let $\mu, \mu_1, \mu_2, \dots \in \mathcal{P}_p(H)$. Then*

$$W_p(\mu_n, \mu) \xrightarrow[n\to\infty]{} 0$$

if and only if $\mu_n \to \mu$ in the p-weak sense. In particular, the metric W_p metrizes the p-weak topology on $\mathcal{P}_p(H)$.

Proof. The implication $W_p(\mu_n, \mu) \to 0 \Longrightarrow \mu_n \to \mu$ in the p-weak sense follows from Proposition 5.39.

Conversely, assume that $\mu_n \to \mu$ in the p-weak sense. By the Skorokhod representation theorem, there exist H-valued random variables X_n, X on a common probability space such that $\mathcal{L}(X_n) = \mu_n$, $\mathcal{L}(X) = \mu$, and $X_n \to X$ almost surely.

By Theorem 5.15, we have $\mu_n \Rightarrow \mu$ and

$$\int_H \|x\|^p \, d\mu_n(x) \to \int_H \|x\|^p \, d\mu(x).$$

As in the proof of Theorem 5.15, this implies the uniform tail estimate

$$\lim_{R\to\infty} \sup_n \int_{\{\|x\|>R\}} \|x\|^p \, d\mu_n(x) = 0.$$

Hence the family $(\|X_n\|^p)_{n\geq 1}$ is uniformly integrable. Since $\|X\|^p$ is integrable, it follows that the family

$$\big(\|X_n - X\|^p\big)_{n\geq 1}$$

is also uniformly integrable, because

$$\|X_n - X\|^p \leq 2^{p-1}\big(\|X_n\|^p + \|X\|^p\big).$$

Now $X_n \to X$ almost surely, so in particular $\|X_n - X\|^p \to 0$ in probability. By Lemma 3.38, we obtain

$$\mathbb{E}\big(\|X_n - X\|^p\big) \xrightarrow[n\to\infty]{} 0.$$

Since the law of (X_n, X) is a coupling of μ_n and μ, the definition of W_p gives

$$W_p(\mu_n, \mu)^p \leq \mathbb{E}\big(\|X_n - X\|^p\big) \xrightarrow[n\to\infty]{} 0.$$

Therefore $W_p(\mu_n, \mu) \to 0$. □

Appendix A

Generalities on Polish spaces

A.1 Some basic facts on Polish spaces

We collect here a few standard facts on Polish spaces which are used at various points in these notes.

Definition A.1. A topological space M is called *Polish* if there exists a metric d on M such that (M, d) is complete and separable, and the topology induced by d is the given topology.

Example A.2. The spaces $\mathbb{R}^d$, every separable Banach space, every compact metric space, and every open or closed subset of $\mathbb{R}^d$ are Polish.

The following theorem characterizes the Polish subsets of a Polish space.

Theorem A.3. *Let M be a Polish space and let $A \subset M$. Then A, equipped with the subspace topology, is Polish if and only if A is a G_δ subset of M.*

Proof. Assume first that $A = \bigcap_{n=1}^{\infty} U_n$, where each U_n is open in M. Let $F_n := M \setminus U_n$, so F_n is closed. Fix a complete metric d on M bounded by 1, and define on A

$$\rho(x, y) := d(x, y) + \sum_{n=1}^{\infty} 2^{-n} \left| \frac{1}{d(x, F_n)} - \frac{1}{d(y, F_n)} \right| \wedge 1.$$

Since $A \cap F_n = \emptyset$, the quantities $d(x, F_n)$ are positive on A, so ρ is well defined. It is standard that ρ induces the same topology on A as the subspace topology. We claim that ρ is complete.

Let (x_k) be a ρ-Cauchy sequence in A. Then it is also d-Cauchy, hence converges in M to some x. Moreover, for each fixed n, the sequence $d(x_k, F_n)$ cannot converge to 0, because then $1/d(x_k, F_n)$ could not be Cauchy. Hence $x \notin F_n$ for every n, so $x \in A$. Thus (x_k) converges in A, proving that A is Polish.

Conversely, assume that A is Polish in the subspace topology. Let ρ be a complete metric on A generating that topology. For every $n \geq 1$ and every $x \in A$, choose an open set $V_n(x) \subset M$ such that

$$x \in V_n(x) \qquad \text{and} \qquad V_n(x) \cap A \subset B_\rho(x, 2^{-n}).$$

Set

$$U_n := \bigcup_{x \in A} V_n(x).$$

Then each U_n is open in M and clearly $A \subset \bigcap_{n=1}^{\infty} U_n$.

To prove the reverse inclusion, let $y \in \bigcap_{n=1}^{\infty} U_n$. For each n choose $x_n \in A$ such that $y \in V_n(x_n)$. If $m \geq n$, then $x_m \in V_n(x_n) \cap A$, hence $\rho(x_m, x_n) < 2^{-n}$. Thus (x_n) is ρ-Cauchy, so it converges in A to some $x \in A$. Since the ρ-topology agrees with the subspace topology, we also have $x_n \to x$ in M. On the other hand, by construction $y \in V_n(x_n)$ for every n, and the sets $V_n(x_n)$ become arbitrarily small in the subspace topology, so necessarily $y = x$. Hence $y \in A$, and therefore $A = \bigcap_{n=1}^{\infty} U_n$. □

Remark A.4. Related to the previous result, every Polish space is homeomorphic to a G_δ subset of the Polish space $[0,1]^{\mathbb{N}}$; this is not hard to see.

Remark A.5. A set which is both Polish and a subspace of a Polish space is therefore automatically Borel, since every G_δ set is Borel.

We will also use the following classical facts from descriptive set theory.

Proposition A.6. *Let M be a Polish space.*

(i) There exists a Borel subset $B \subset [0,1]$ and a Borel isomorphism from M onto B.

(ii) If M is uncountable, then M is Borel isomorphic to $[0,1]$.

(iii) Any two uncountable Borel subsets of Polish spaces are Borel isomorphic.

Remark A.7. We will use this proposition in the next theorem. A proof may be found in any standard text on descriptive set theory.

Definition A.8. Let $(M, \mathcal{B}(M))$ and $(N, \mathcal{B}(N))$ be measurable spaces. A bijection $T : M \to N$ is called a *Borel isomorphism* if both T and T^{-1} are Borel measurable.

A probability measure μ on M is called *continuous* if

$$\mu(\{x\}) = 0 \qquad \text{for every } x \in M.$$

Equivalently, μ has no atoms.

The next theorem says that every atomless Borel probability measure on a Polish space is, from the measurable point of view, the same as Lebesgue measure on the unit interval.

Theorem A.9. *Let M be a Polish space and let μ be a continuous probability measure on M. Then there exists a Borel isomorphism*

$$T : M \to [0,1]$$

such that

$$T_{\#}\mu = \lambda,$$

where λ denotes Lebesgue measure on $[0,1]$.

Proof. By Proposition A.6, there exists a Borel isomorphism from M onto a Borel subset $B \subset [0,1]$. Replacing M by B and μ by its pushforward, we may assume from the beginning that $M = B \subset [0,1]$ is a Borel set and μ is a continuous probability measure on B.

Let $S := \operatorname{supp}(\mu)$ be the support of μ. Then S is a closed subset of $[0,1]$, hence Polish, and $\mu(B \setminus S) = 0$. Define

$$F(x) := \mu(S \cap [0,x]), \qquad x \in S.$$

Since μ is continuous, F is continuous on S. Moreover, F is strictly increasing on S: indeed, if $x < y$ are in S, then $(x,y) \cap S \neq \emptyset$, and by definition of the support every nonempty open subset of S has positive μ-measure. Hence $\mu(S \cap (x,y)) > 0$, so $F(y) > F(x)$. Since $F(\min S) = 0$ and $F(\max S) = 1$, it follows that

$$F : S \to [0,1]$$

is a homeomorphism. In particular, it is a Borel isomorphism, and for every $t \in [0,1]$,

$$\mu(F^{-1}([0,t])) = t,$$

so $F_{\#}(\mu_{|S}) = \lambda$.

It remains to extend this to all of B. Let $N := B \setminus S$. Then N is Borel and $\mu(N) = 0$. We now choose a Borel null set $D \subset [0,1]$ such that there exists a Borel isomorphism from $N \cup F^{-1}(D)$ onto D. This is possible by Proposition A.6: if N is countable, choose D countable; if N is uncountable, choose D to be an uncountable null Borel set, for instance a Cantor-type set.

Set

$$C := F^{-1}(D) \subset S.$$

Then C is Borel and $\mu(C) = \lambda(D) = 0$. Also, F restricts to a Borel isomorphism

$$F : S \setminus C \to [0,1] \setminus D.$$

Choose a Borel isomorphism

$$h : N \cup C \to D.$$

Now define

$$T(x) := \begin{cases} F(x), & x \in S \setminus C, \\ h(x), & x \in N \cup C. \end{cases}$$

Then $T : B \to [0,1]$ is a Borel bijection, and its inverse is again Borel since both pieces are Borel isomorphisms on Borel sets. Thus T is a Borel isomorphism.

Finally, let $A \subset [0,1]$ be Borel. Since $\mu(N \cup C) = 0$ and $\lambda(D) = 0$, we have

$$\mu(T^{-1}(A)) = \mu\big(F^{-1}(A \setminus D)\big) = \lambda(A \setminus D) = \lambda(A).$$

Hence $T_{\#}\mu = \lambda$, as claimed. □

Remark A.10. The previous theorem says that every atomless Borel probability measure on a Polish space is measurably indistinguishable from Lebesgue measure on the unit interval. In probability theory, this is one of the basic manifestations of the fact that atomless standard probability spaces are all isomorphic.

Bibliography

[1] Fifth edition, 2019.

[2] G. Grimmett and D. Stirzaker. *Probability and Random Processes*. OUP, 3rd edn edition, 2001.

[3] G. Shafer and V. Vovk. The Sources of Kolmogorov's Grundbegriffe. *Statistical Science*, 21(1):70–98, 2006.