Advanced probability theory

Christian Bayer

June 8, 2011

Contents

1 Introduction **4**
- 1.1 Motivation . . . 4
- 1.2 Terminology . . . 7
- 1.3 Some milestones of this course . . . 9
- 1.4 Literature . . . 10

2 Independence **11**
- 2.1 Independent systems . . . 11
- 2.2 Independent random variables . . . 13
- 2.3 The Borel-Cantelli lemma and 0-1-laws . . . 18

3 The strong law of large numbers **22**
- 3.1 The strong law of large numbers . . . 22
- 3.2 Large deviations . . . 24
- 3.3 The ergodic theorem . . . 26
- 3.4 Applications and examples . . . 29
  - 3.4.1 An example of ergodicity . . . 29
  - 3.4.2 Monte Carlo simulation . . . 30

4 Conditional expectations **33**
- 4.1 Conditional expectations . . . 34
- 4.2 Markov kernels and regular conditional distributions . . . 37
- 4.3 Martingales . . . 41
- 4.4 Optional sampling . . . 43

5 Stochastic processes **47**
- 5.1 Examples . . . 47
- 5.2 Poisson process . . . 50
- 5.3 Construction of stochastic processes . . . 52

6 Brownian motion **58**
- 6.1 Construction of the Brownian motion . . . 58
- 6.2 Properties of the Brownian motion . . . 62
- 6.3 Donsker's invariance principle . . . 65

A Collection of results from elementary probability **70**

B Characteristic functions **72**

C Weak convergence and the central limit theorem **76**

Chapter 1

Introduction

1.1 Motivation

Let us start with the fundamental definition.

Definition 1.1. A *probability space* is a measure space $(\Omega, \mathcal{F}, P)$ such that $P(\Omega) = 1$. The elements of the σ-algebra $\mathcal{F}$ are called *events*.

Remark 1.2. The measure space of Definition 1.1 is a model for the axiomatic definition of a probability space. Notice that this model works for every kind of probability, be it subjective, frequentist or other types of probabilities.

In the remainder of this section, I want to motivate this definition, which is in some contrast to the elementary setting – for simplicity, we restrict ourselves to the one-dimensional case. Indeed, in elementary probability theory, one usually considers two quite different classes of models, namely

1. discrete models, where Ω is a finite or countably infinite set and probabilities are given by $P(\{\omega\}) = p_\omega$ for some sequence of non-negative numbers with $\sum_{\omega\in\Omega} p_\omega = 1$;

2. continuous model with a (continuous, or at least Riemann integrable) density p, where probabilities of events A are defined by $\int_A p(x)dx$.

Some unification is achieved by the cumulative *distribution function* $F(x) = P(]-\infty, x])$, but it is not always intuitive to work with the distribution function, and, moreover, some technical problems remain. There are some rather obvious advantages of Definition 1.1:

1. Both discrete and continuous models in the above sense are contained in Definition 1.1. Mixed models, with densities and point masses, are quite messy to deal with in elementary probability, but fit very naturally into the framework of Definition 1.1.

2. Definition 1.1 allows to use the machinery of measure theory, in particular to convergence theorems (for interchanging limits and integrals/sums), which are often much simpler then the corresponding theorems for sums or Riemann integrals.

3. Definition 1.1 encompasses models, which are neither discrete nor continuous nor of the mixed type. Such models can actually appear for instance as limiting cases of elementary models. Moreover, they play an important role by providing counter-examples.

We will now, in some detail, describe a quite innocent probabilistic problem, where we can see how measure theory appears rather naturally. Let us toss one coin. There are two possible outcomes of the experiment, head and tail. Thus, we can choose our probability space to be $\Omega_1 := \{H, T\}$, $\mathcal{F}_1 := \mathcal{P}(\Omega_1)$ and the probability measure $P_1(\{H\}) = p, P_1(\{T\}) = q := 1 - p$ with $0 \leq p \leq 1$. If we want to repeat the experiment n times (in an independent fashion, using the same coin), we can choose the probability space $\Omega_n := \{H, T\}^n$ again with the power set as σ-algebra $\mathcal{F}_n$ and with the probabilities of the elementary events defined as

$$P_n(\{(\omega_1, \ldots, \omega_n)\}) := \prod_{i=1}^{n} P_1(\{\omega_i\}),$$

which is a product of p's and q's. Of course, the probability space $(\Omega_n, \mathcal{F}_n, P_n)$ is nothing but the n-fold product measure space of $(\Omega_1, \mathcal{F}_1, P_1)$, symbolically

$$(\Omega_n, \mathcal{F}_n, P_n) = (\Omega_1, \mathcal{F}_1, P_1)^{\otimes n}.$$

At this stage, we might be interested in certain properties of the experiment. For instance, we might be interested in the number of heads appearing among the first n coin tosses. Principally, we could answer all questions regarding this random number given the notions introduced so far, e.g., the probability that exactly 2 heads appeared among the first four coin tosses is

$$P_4\left(\{(H,H,T,T),(H,T,H,T),(H,T,T,H),(T,H,H,T),(T,H,T,H),(T,T,H,H)\}\right).$$

However, intuition and notation becomes much simpler when we introduce *random variables*, for instance by setting $X_1, \ldots, X_n : \Omega_n \to \{0, 1\}$,

$$X_i(\omega_1, \ldots, \omega_n) := \mathbf{1}_{\{H\}}(\omega_i) = \begin{cases} 1, & \omega_i = H, \\ 0, & \omega_i = T, \end{cases} \qquad i = 1, \ldots, n. \tag{1.1}$$

Then, the number of heads in the first n coin tosses is obviously given by $X_1 + \cdots + X_n$, and all the related probabilities can be expressed in terms of the distribution of the independent random variables X_i, $P(X_i = 1) = p$. Notice an abuse of notation (which is rather typical in probability theory): take two different numbers $n < m$ and consider the k'th coin toss X_k, $k \leq n$. Then, in one case $X_k : \Omega_n \to \{0, 1\}$, in the other case $X_k : \Omega_m \to \{0, 1\}$. Obviously, these functions are different, since their domains of definition are different. However, in both cases, the distribution of the random variable is the same, and in both cases all the random variables X_k, $1 \leq k \leq n$, are independent. Thus, for the sake of our probabilistic inquiry, we may treat both maps as if they were equal. In fact, this is true in much more generality. Indeed, we could have started with a different probability space $(\Omega_1, \mathcal{F}_1, P_1)$. For instance, we could have chosen $\Omega_1 = \mathbb{R}$, $\mathcal{F}_1 = \mathcal{B}(\mathbb{R})$ and P_1 given by a density, e.g., by $f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$, and then define $X_1(\omega) := \mathbf{1}_{[u,\infty[}(\omega)$ for some u such that $P_1(X_1 = 1) = p$. (And define, once again, $(\Omega_n, \mathcal{F}_n, P_n)$ as the product space.) More obvious would be the choice $\Omega_1 = \{0, 1\}$ with $X_1(\omega) = \omega$ (the *canonical* probability space for the random variable X_1). In all these cases, the probabilistic analysis of coin tosses will be completely analogous.

In probability theory, only the distribution of the random variables under investigation matters, but not the underlying probability space.

In addition to questions about the probabilities of events in finite time, one is often also interested in asymptotic questions. For instance, if we do not know the parameter p before – the fairness of the coin – we could try to estimate it using statistics. The obvious estimator certainly is the sample mean, i.e.,

$$\overline{p}^{(n)} := \frac{1}{n} \sum_{i=1}^{n} X_i,$$

and one of the most important problems in statistics is to determine whether $\overline{p}^{(n)}$ converges for $n \to \infty$ to the true value p and in which sense. The *weak law of large numbers* holds if the estimator converges in probability (i.e., in measure), which means that for every $\epsilon > 0$

$$\lim_{n\to\infty} P\left(\left|\overline{p}^{(n)} - p\right| > \epsilon\right) = 0. \tag{1.2}$$

The above statement is not quite precise, because we do not know (yet) how to construct a *common* probability space $(\Omega, \mathcal{F}, P)$, where all the random variables X_n, $n \in \mathbb{N}$, are simultaneously defined. However, in the current situation we might be happy reformulating the statement (1.2) to say

$$\lim_{n\to\infty} P_n\left(\left|\overline{p}^{(n)} - p\right| > \epsilon\right) = 0,$$

which is good enough for our current purpose. Much better than the weak law of large numbers is the *strong law of large numbers*, which implies convergence of the estimator P-almost everywhere – in probability theory, one usually says *"almost surely"* instead of *"almost everywhere"*. This means, we need to show that

$$P\left(\lim_{n\to\infty} \overline{p}^{(n)} = p\right) = 1. \tag{1.3}$$

If we want to prove the strong law of large numbers, then we really need to find a common probability space for all the (infinitely many) coin tosses, i.e., we need to construct an infinite product space.

While this construction was already given in the measure theory course, we use the opportunity to give a short reminder. The candidate for Ω is obvious enough: $\Omega := \{H, T\}^{\mathbb{N}}$, the set of all sequences taking values in $\{H, T\}$.[1] The choice of $\mathcal{F}$ is already not so clear. After all, Ω is now an infinite (even uncountable) set. Therefore, it does not fit into the framework of discrete probability spaces anymore. $P(\{\omega\}) = 0$ for any particular sequence ω of coin tosses, but this does not determine P. In particular, the power set of Ω may very well be too large a σ-algebra. On the other hand, the continuous model also does not fit well. However, since we are doing probability theory, we might argue that we really only care about the random variables X_n, i.e., we only need all the functions $X_n : \Omega \to \{0, 1\}$, $n \in \mathbb{N}$, to be measurable, which leads to

[1] More generally, Ω is the Cartesian product of all the individual Ω_i, $i \in I$, symbolically, $\Omega = \prod_{i\in I} \Omega_i$, if we want to construct $\bigotimes_{i\in I} (\Omega_i, \mathcal{F}_i, P_i)$ for a general index set I.

the choice $\mathcal{F} = \sigma((X_n)_{n\in\mathbb{N}})$.[2] Let now $J \subset \mathbb{N}$ be a finite set, then $X_J := (X_n)_{n\in J} : \Omega \to \{0,1\}^J$ is measurable, implying that the *cylinder sets*

$$Z = \{\, \omega \in \Omega \;\big|\; (\omega_{i_1}, \ldots, \omega_{i_k}) \in A \,\} =: p_J^{-1}(A)$$

are measurable sets, where $J = (i_1, \ldots, i_k) \subset \mathbb{N}$, $k \in \mathbb{N}$, $A \in \{H, T\}^k$ and $p_J((\omega_n)) = (\omega_j)_{j\in J}$. For cylinder sets, however, we *know* how P should look like: we should have

$$P(Z) = P(p_J^{-1}(A)) = P_k(A). \tag{1.4}$$

Finally, one can prove that there is a unique probability measure P having the property (1.4) for all cylinder sets Z.[3]

1.2 Terminology

We start with a very formal definition of random variables.

Definition 1.3. Given a probability space $(\Omega, \mathcal{F}, P)$ and some measurable space $(A, \mathcal{A})$. Then any $\mathcal{F}$-$\mathcal{A}$-measurable map $X : \Omega \to A$ is called *random variable*.

With decreasing level of importance (and frequency), we shall consider random variables taking values in

- $\mathbb{R}$ with the Borel σ-algebra $\mathcal{B}(\mathbb{R})$;
- the Euclidean space $\mathbb{R}^n$ with the Borel σ-algebra $\mathcal{B}(\mathbb{R}^n)$;
- a Banach space E with its Borel σ-algebra $\mathcal{B}(E)$;
- a Polish space[4] M with its Borel σ-algebra $\mathcal{B}(M)$.

Remark 1.4. We may also use random variables taking values, for instance, in $\mathbb{R} \cup \{\pm\infty\}$.

As we have noticed before, in probability theory the underlying probability space usually does not matter, problems are formulated and treated in terms of random variables. Of course, not every probability space can carry any type of random variables. For instance, we cannot define a Gaussian random variable on the probability space $(\Omega_1, \mathcal{F}_1, P_1)$ used in section 1.1. This leads to the notion of the distribution of a random variable.

Definition 1.5. Given a probability space $(\Omega, \mathcal{F}, P)$, a measurable space $(A, \mathcal{A})$ and a random variable $X : \Omega \to A$. The image measure P_X on $(A, \mathcal{A})$ (i.e., the measure defined by $P_X(B) := P(X^{-1}(B))$, $B \in \mathcal{A}$) is called *distribution* of X.

[2] In general, we introduce the projections $p_i : \Omega \to \Omega_i$ defined by $p_i\big((\omega_j)_{j\in I}\big) = \omega_i$ and set $\mathcal{F} = \sigma((p_i)_{i\in I}) = \sigma\big(\bigcup_{i\in I} p_i^{-1}(\mathcal{F}_i)\big)$. Of course, the definition in the dice-throwing example of the text coincides with this definition of $\mathcal{F}$.

[3] In general, cylinder sets have the form $Z = p_J^{-1}(A)$, where p_J is defined as above and $A \in \bigotimes_{i\in J} \mathcal{F}_i$. Then, consider the well-known finite product measure P_J defined on $\bigotimes_{i\in J} \mathcal{F}_i$. Then the requirement (1.4) is recast to say that $P(Z) = P_J(A)$. Again, one can prove existence and uniqueness of such a probability measure on $(\Omega, \mathcal{F})$, just like in the countable case. For a proof in the case of a countable index set I see the lecture notes of F. Hofbauer, for the proof in the general case see Bauer, Satz 9.2.

[4] A Polish space is a separable, complete metric space.

Remark 1.6. Obviously, $(A, \mathcal{A}, P_X)$ is again a probability space. Let us define a map $\mathrm{id}_A : A \to A$ by $\mathrm{id}_A(x) = x$, $x \in A$. We understand id_A as random variable mapping the probability space $(A, \mathcal{A}, P_X)$ into itself. Then the distribution of id_A is P_X, i.e., X and id_A have the same distributions. In particular, this example shows that for every probability measure P one can find a random variable X (defined on a suitable probability space) whose distribution is P. Moreover, the random variable is certainly not unique.

If a random variable X takes values in the Euclidean space $\mathbb{R}^n$, then we can define its integral.

Definition 1.7. Let X be a random variable defined on a probability space $(\Omega, \mathcal{F}, P)$ taking values in $\mathbb{R}$ (with the Borel σ-algebra). If $|X|$ is an integrable function or $X \geq 0$, then we define the *expectation* or *expected value* as its integral and write

$$E[X] \coloneqq \int_\Omega X dP = \int_\Omega X(\omega) P(d\omega).$$

If $X = (X_1, \ldots, X_n)$ takes its values in the space $\mathbb{R}^n$ equipped with its Borel σ-algebra and $\|X\|$ is integrable, then we define $E[X] \coloneqq (E[X_1], \ldots, E[X_n])$.

If, e.g., the negative part X^- is integrable, whereas the positive part is not, we will also write that $E[X] = \infty$. Of course, all the properties of the integral with respect to a measure obtained in measure theory carry over. In particular, we have the following often used formula, sometimes called the "theorem of the intuitive statistician".

Lemma 1.8. *Let X be a real random variable on $(\Omega, \mathcal{F}, P)$.*[5] *Moreover, let $g : \mathbb{R} \to \mathbb{R}$ be measurable such that $|g \circ X|$ is integrable. Then*

$$E[g(X)] \coloneqq E[g \circ X] = \int_\Omega g(X(\omega)) P(d\omega) = \int_{\mathbb{R}} g(x) P_X(dx).$$

In particular, the lemma implies that the expectation of a random variable is the integral of the identity function with respect to the distribution of the random variable.

En passant, we note that the distribution of a real-valued random variable has a uniquely defined *(cumulative) distribution function*

$$F(x) \coloneqq P(X \leq x) \coloneqq P(X^{-1}(]-\infty, x])),$$

which is right-continuous and increasing and satisfies $F(-\infty) = 0$, $F(+\infty) = 1$. Moreover, any such function uniquely determines a distribution on the real line. Then, the integrals of Lemma 1.8 can equivalently be expressed as Lebesgue-Stieltjes integral with respect to the distribution function F

$$E[g(X)] = \int_{\mathbb{R}} g(x) dF(x).$$

Of course, the expectation coincides with the elementary definitions for discrete and continuous distributions. As in the elementary case, we also introduce higher moments.

Definition 1.9. Let $p \geq 1$ and assume that $|X|^p$ is integrable. Then $E[X^p]$ is the *p'th moment* of X. For square-integrable random variables the *variance* is defined by $V[X] \coloneqq E[(X - E[X])^2]$.

[5] By this we mean that X is an $\mathcal{F}$-$\mathcal{B}(\mathbb{R})$-measurable map from Ω to $\mathbb{R}$. In the following we will often omit such obvious qualifications.

Of course, the classical inequalities like Markov's inequality or Chebyshev's inequality immediately carry over from the elementary setting.

In many applications, we use probabilistic models for the evolution of some quantity in time. In its most abstract form, this leads to

Definition 1.10. Given an index set I (usually $I = \mathbb{N}$ or $I = [0, \infty[$ or subsets thereof), a collection of random variables $(X_i)_{i \in I}$ (defined on a common probability space) is called a *stochastic process*.

> Like in Definition 1.10, we will often omit the probability space $(\Omega, \mathcal{F}, P)$ from statements. If the probability space is not explicitly mentioned, it is tacitly assumed that there exists a probability space, where all relevant random variables are (jointly) defined.

1.3 Some milestones of this course

Most topics of the course are somehow related with the concept of a stochastic process. The first major topic regards the asymptotic of sequences of independent, identically distributed random variables $(X_n)_{n \in \mathbb{N}}$. Set $S_n := \sum_{k=1}^{n} X_k$ the corresponding *random walk*. If these random variables are integrable, we know from elementary probability theory that the weak law of large numbers holds,

$$\forall \epsilon > 0 : \lim_{n \to 0} P\left(\left|\frac{1}{n} S_n - E[X_1]\right| > \epsilon\right) = 0.$$

One of the main results of this course will be a substantially stronger result known as strong law of large numbers, which says that

$$P\left(\lim_{n \to \infty} \frac{1}{n} S_n \neq E[X_1]\right) = 0,$$

and we will explore certain versions of this statement. Moreover, it might also be interesting to see, how fast the probability of deviations from the long-term mean $E[X_1]$ decreases to 0, which leads to the realms of "large deviations".

Already in elementary probability theory we have learned some valuable information about the speed of the above convergence. Indeed, if we multiply the difference by $\sqrt{n}$, then the difference converges in distribution to a normal distribution. More precisely, the *central limit theorem* says that

$$\frac{S_n - nE[X_1]}{\sqrt{n}\sigma} \xrightarrow[n \to \infty]{d} \mathcal{N}(0, 1), \tag{1.5}$$

provided that $\sigma^2 := V[X_1]$ exists, $\xrightarrow{d}$ denotes convergence in distribution and $\mathcal{N}(\mu, \sigma^2)$ denotes the one-dimensional Gaussian distribution with mean μ and variance σ^2. This shows the unique role of the normal distribution as limiting distribution in the finite-dimensional framework. In the infinite-dimensional framework, a similar role is played by a certain stochastic process $(B_t)_{t \geq 0}$, the *Brownian motion*. Indeed, in many physical or social phenomena which are described as functions of time, and which can be explained as the result of many small, independent random perturbations, the Brownian motion is natural probabilistic model. Examples include the position of gas molecules

(which are constantly and from all directions hit by other molecules) or the prices of stocks (where the "shocks" are caused by buy or sell decisions of individual agents). It is not surprising that the marginals B_t of a Brownian motion have normal distribution, but the inter-dependence is less obvious. Brownian motion is not only very important from the point of view of modeling and applications, it is also a very interesting mathematical object in itself, for instance as a source of counter examples in analysis.

Finally, we shall also treat the problem of the construction of a stochastic process. More precisely, we will address the question of how to construct a stochastic process when the finite-dimensional marginal distribution are prescribed – and do not need to be independent.

1.4 Literature

I recommend the following two textbooks on probability theory – both of them have German and English editions:[6]

- Heinz Bauer: *Wahrscheinlichkeitstheorie*, 5. Auflage, Walter de Gruyter, Berlin, 2002.
- Achim Klenke: *Wahrscheinlichkeitstheorie*, 2. Auflage, Springer-Verlag, Berlin, 2008.

The notes are also based on lecture notes "Wahrscheinlichkeitstheorie" by E. Bolthausen, then at TU Berlin, and the lecture notes "Wahrscheinlichkeitstheorie und Statistik" and "Maß- und Integrationstheorie" by F. Hofbauer from University of Vienna.

[6]The editions cited were used for preparation of this text.

Chapter 2

Independence

Let $(\Omega, \mathcal{F}, P)$ denote a fixed probability space. We recall (and slightly extend) the definition of independence for events.

Definition 2.1. A family $(A_i)_{i\in I}$ of events, i.e., $A_i \in \mathcal{F}$ for all $i \in I$, is called *independent* if for every finite subset $J \subset I$ we have

$$P\left(\bigcap_{j\in J} A_j\right) = \prod_{j\in J} P(A_j). \tag{2.1}$$

Remark 2.2. Notice that the notion of independence clearly depends on the probability measure, i.e., if the probability measure P is replaced by some other probability measure Q on $(\Omega, \mathcal{F})$, then $(A_i)_{i\in I}$ may lose its independence.

Example 2.3. Independence of a family of events like in Definition 2.1 is stronger than pairwise independence of the events. Indeed, throw two dice independent of each other, and let A_1 be the event that the outcome of the first die is even, A_2 the event that the outcome of the second die is even, and A_3 the event that the sum of the outcomes is even. Then the A_1 and A_2, A_1 and A_3, A_2 and A_3 are independent (i.e., the events A_1, A_2, A_3 are pairwise independent), but the events A_1, A_2, A_3 are not independent, since $P(A_1 \cup A_2) = P(A_1 \cup A_2 \cup A_3)$.

Recall from elementary probability theory that independence is not destroyed when some of the A_i's are replaced by their complements. In what follows we will generalize the notion of independence to families of σ-algebras and random variables.

2.1 Independent systems

Definition 2.4. Let $(\mathcal{E}_i)_{i\in I}$ a family of systems of sets, $\emptyset \neq \mathcal{E}_i \subset \mathcal{F}$ for $i \in I$. We say that the family is *independent* iff every collection of events $(A_i)_{i\in I}$ with $\forall i \in I : \ A_i \in \mathcal{E}_i$ is independent in the sense of Definition 2.1.

Obviously, a family $(\mathcal{E}_i)_{i\in I}$ is independent if and only if every finite sub-family is independent. Moreover, independence of a family of systems of sets is certainly preserved when we pass to a family of subsystems $\left(\mathcal{E}'_i\right)_{i\in I}$ with $\mathcal{E}'_i \subset \mathcal{E}_i$. In what follows, we are mostly interested in the question of how we can preserve independence when we enlarge the set-systems.

Lemma 2.5. *Let $(\mathcal{E}_i)_{i\in I}$ be independent. Then the family of the generated* Dynkin systems *$(\delta(\mathcal{E}_i)_{i\in I}$ is also independent.*

Proof. We may assume that I is finite. We show that for any individual index $i_0 \in I$ we may replace $\mathcal{E}_{i_0}$ by its generated Dynkin system $\delta(\mathcal{E}_{i_0})$ without losing independence, which shows the claim. Let us define a system of sets $\mathcal{D}_{i_0}$ by setting

$$\mathcal{D}_{i_0} := \{\, A \in \mathcal{F} \;\big|\; (\{A\}, \mathcal{E}_i :\ i \in I \setminus \{i_0\}) \text{ is independent} \,\}.$$

We show that $\mathcal{D}_{i_0}$ is a Dynkin system. The set Ω is contained in $\mathcal{D}_{i_0}$, since for any collection $A_i \in \mathcal{E}_i$, $i \in I \setminus \{i_0\}$, and for every subset $\{i_1, \dots, i_n\} \subset I \setminus \{i_0\}$ we have

$$P(\Omega \cap A_{i_1} \cap \cdots \cap A_{i_n}) = P(A_{i_1} \cap \cdots \cap A_{i_n}) = P(A_1) \cdots P(A_{i_n}) = P(\Omega)P(A_{i_1}) \cdots P(A_{i_n}).$$

Given a sets $A \in \mathcal{D}_{i_0}$, then

$$\begin{aligned} P(A^c \cap A_{i_1} \cap \cdots \cap A_{i_n}) &= P(\Omega \cap A_{i_1} \cap \cdots \cap A_{i_n}) - P(A \cap A_{i_1} \cap \cdots \cap A_{i_n}) \\ &= P(\Omega)P(A_{i_1}) \cdots P(A_{i_n}) - P(A)P(A_{i_1}) \cdots P(A_{i_n}) = P(A^c)P(A_{i_1}) \cdots P(A_{i_n}), \end{aligned}$$

showing that $A^c \in \mathcal{D}_{i_0}$. Similarly, we show that any countably union of disjoint sets from $\mathcal{D}_{i_0}$ is again in $\mathcal{D}_{i_0}$. Therefore, $\mathcal{D}_{i_0}$ is a Dynkin system, which contains the smallest Dynkin system $\delta(\mathcal{D}_{i_0})$ generated by $\mathcal{E}_{i_0}$. □

If a set system $\mathcal{E}$ is stable under intersections of sets, then $\delta(\mathcal{E}) = \sigma(\mathcal{E})$. Therefore, Lemma 2.5 immediately implies the

Corollary 2.6 (Hofbauer, MT, Satz 4.9)**.** *Let $(\mathcal{E}_i)_{i\in I}$ be an independent family of systems of sets, which are closed under intersections. Then the family of generated σ-algebras $(\sigma(\mathcal{E}_i))_{i\in I}$ is independent.*

The following result is a preparation for important results concerning independence of random variables.

Theorem 2.7. *Let $(\mathcal{E}_i)_{i\in I}$ be an independent family of subsets $\emptyset \neq \mathcal{E}_i \subset \mathcal{F}$ which are closed under intersections. Moreover, let $(I_j)_{j\in J}$ be a partition of the index set I. Then the family $\left(\mathcal{F}_j\right)_{j\in J}$ is independent, where $\mathcal{F}_j := \sigma\left(\bigcup_{i\in I_j} \mathcal{E}_i\right)$, $j \in J$.*

Proof. For $j \in J$ define a system $\mathcal{G}_j$ by

$$\mathcal{G}_j := \left\{ A_{i_1} \cap \cdots \cap A_{i_n} \;\middle|\; n \in \mathbb{N},\ \{i_1, \dots, i_n\} \subset I_j,\ A_{i_1} \in \mathcal{E}_{i_1}, \dots, A_{i_n} \in \mathcal{E}_{i_n} \right\}.$$

Obviously, $\mathcal{G}_j$ is closed under intersections. Moreover, $(\mathcal{G}_j)_{j\in J}$ is independent by independence of $(\mathcal{E}_i)_{i\in I}$. Therefore, the claim follows by Corollary 2.6, since $\sigma(\mathcal{G}_j) = \sigma\left(\bigcup_{i\in I_j} \mathcal{E}_i\right)$. □

As usual, in probability theory, we want to build our intuition on properties of random variables and their distributions, not on the underlying probability spaces. Therefore, we want to arrive at a notion of independence for random variables. Let us first link independence of systems and events by showing that independence of systems is really a more general notion than independence of events. Indeed, given a family $(A_i)_{i\in I}$ of events, the corresponding family of systems $(\{A_i\})_{i\in I}$ is trivially closed under intersections. Thus, independence of the events is equivalent to independence of the system and to independence of the generated σ-algebras $\mathcal{F}_i = \sigma(\{A_i\}) = \{\emptyset, A_i, A_i^c, \Omega\}$, which could naturally serve as definition of the random variable $\mathbf{1}_{A_i}$. This is indeed the approach followed next, which will lead to a notion of independence equivalent to the elementary one.

2.2 Independent random variables

Definition 2.8. Given a family $(X_i)_{i\in I}$ of random variables defined on a probability space $(\Omega, \mathcal{F}, P)$. The family is called *independent*, iff the family of σ-algebras $(\sigma(X_i))_{i\in I}$ is independent in the sense of Definition 2.4.

Note that we do not require the random variables to take values in the same measurable space. Indeed, we may assume that for each $i \in I$, X_i takes values in the measurable space $(\Omega_i, \mathcal{F}_i)$. Then the product formula (2.1) can be expressed like follows: for every finite subset $J = \{ i_1, \dots, i_n \} \subset I$ and any $A_j \in \mathcal{F}_j$, $j \in J$, we have

$$P(X_{i_1} \in A_{i_1}, \dots, X_{i_n} \in A_{i_n}) = P(X_{i_1} \in A_{i_1}) \cdots P(X_{i_n} \in A_{i_n}). \tag{2.2}$$

By Corollary 2.6 we may restrict ourselves to sets A_j forming generators of $\mathcal{F}_j$ closed under intersections. Indeed, we have the

Corollary 2.9. *Let $(\mathcal{E}_i)_{i\in I}$ be a family of systems $\mathcal{E}_i \subset \mathcal{F}_i$ which are closed under intersections and generate the σ-algebras, i.e., $\forall i \in I : \mathcal{F}_i = \sigma(\mathcal{E}_i)$. Then the family of $\mathcal{F}$-$\mathcal{F}_i$-measurable random variables X_i, $i \in I$, are independent if and only if the system $(\mathcal{G}_i)_{i\in I}$ is independent with $\mathcal{G}_i \coloneqq X_i^{-1}(\mathcal{E}_i)$.*

Spelled out, this means that the product formula (2.2) holds for every finite subset $\{ i_1, \dots, i_n \} \subset I$ and $A_{i_j} \in \mathcal{E}_{i_j}$, $j = 1, \dots, n$. For real random variables (i.e., random variables taking values in $\mathbb{R}$ equipped with the Borel σ-algebra), this means that independence is characterized by the distribution function.

Theorem 2.10 (Hofbauer, MT, Definition)**.** *Let $(X_i)_{i\in I}$ be a family of real random variables. For any finite subset $\{ i_1, \dots, i_n \} = J \subset I$ let $F_J : \mathbb{R}^{|J|} \to [0, 1]$ denote the corresponding cumulative distribution functions, i.e.,*

$$F_J(x_1, \dots, x_n) = P(X_{i_1} \le x_1, \dots, X_{i_n} \le x_n).$$

Then the family $(X_i)_{i\in I}$ is independent if and only if for every such finite subset J we have the product formula

$$F_J(x_1, \dots, x_n) = \prod_{j=1}^{n} F_{\{ i_j \}}(x_j).$$

In particular, for discrete and continuous random variables, the notion of independence as defined in Definition 2.8 coincides with the elementary definition.

Example 2.11. Given a finite sequence $(X_1, \dots, X_n)$ of random variables, each taking values in a finite (or countably infinite) set E_k, $k = 1, \dots, n$. Then $X_1, \dots, X_n$ are independent if and only if for every $(e_1, \dots, e_n) \in E_1 \times \cdots \times E_n$

$$P(X_1 = e_1, \dots, X_n = e_n) = P(X_1 = e_1) \cdots P(X_n = e_n).$$

Example 2.12. Given a finite sequence of real random variables $(X_1, \dots, X_n)$ with joint density $f : \mathbb{R}^n \to \mathbb{R}$. Then $X_1, \dots, X_n$ are independent if and only if f factorizes, i.e., there are functions $f_1, \dots, f_n : \mathbb{R} \to \mathbb{R}$ with the property that for any $x = (x_1, \dots, x_n) \in \mathbb{R}^n$

$$f(x) = f_1(x_1) \cdots f_n(x_n).$$

As a probabilistic property, independence of random variables only depends on the distributions, but not on the underlying probability space. However, the marginal distributions P_{X_i}, $i \in I$, cannot determine the dependence structure, which is a property of the *joint distribution*. Let $(X_i)_{i\in I}$ be a family of random variables defined on the probability space $(\Omega, \mathcal{F}, P)$ and taking values in the measurable spaces $(\Omega_i, \mathcal{F}_i)$, $i \in I$, respectively. Consider the product space

$$\left(\widetilde{\Omega}, \widetilde{\mathcal{F}}\right) := \bigotimes_{i\in I} (\Omega_i, \mathcal{F}_i).^1 \tag{2.3}$$

By definition of the product space, $X : \Omega \to \widetilde{\Omega}$, $\omega \mapsto (X_i(\omega))_{i\in I}$, is an $\mathcal{F}$-$\widetilde{\mathcal{F}}$-measurable map: indeed, we need to show that for every $i \in I$, $p_i \circ X : \Omega \to \Omega_i$ is $\mathcal{F}$-$\mathcal{F}_i$-measurable. But this is true since $p_i \circ X = X_i$. Therefore, we can define the *distribution* of the family $(X_i)_{i\in I}$ as the image measure P_X of P under X, which gives us the probability space $(\widetilde{\Omega}, \widetilde{\mathcal{F}}, P_X)$.

Theorem 2.13. *A family of random variables $(X_i)_{i\in I}$ is independent if and only if the joint distribution is the product of the marginal distributions, i.e.,*

$$\left(\widetilde{\Omega}, \widetilde{\mathcal{F}}, P_X\right) = \bigotimes_{i\in I} (\Omega_i, \mathcal{F}_i, P_{X_i}).$$

Proof. Let W denote the product measure of the probability measures P_{X_i}, $i \in I$. We need to show that $P_X = W$. By the construction of infinite product measures, this is equivalent to equality between all finite-dimensional image measures: for finite subsets $J \subset I$ denote p_J the corresponding projection $p_J((\omega_i)_{i\in I}) = (\omega_i)_{i\in J}$ and consider the image measures W_{p_J} and $(P_X)_{p_J} = P_{p_J \circ X} = P_{(X_i)_{i\in J}}$. But by construction of the product measure $W = \bigotimes_{i\in I} P_{X_i}$, W_{p_J} is the product measure of P_{X_i}, $i \in J$.

Now assume that $(X_i)_{i\in I}$ are independent. Then, in particular, $(X_i)_{i\in J}$ is independent. For any $A_i \in \mathcal{F}_i$, $i \in J$, we have

$$P_{(X_i)_{i\in J}}\left(\prod_{i\in J} A_i\right) = P\left((X_i)_{i\in J} \in \prod_{i\in J} A_i\right) = \prod_{i\in J} P(X_i \in A_i) = \prod_{i\in J} P_{X_i}(A_i) = W_{p_J}\left(\prod_{i\in J} A_i\right).$$

This shows that, indeed, $P_{(X_i)_{i\in J}} = W_{p_J}$.

Now, to the contrary, assume that $P_{(X_i)_{i\in J}} = W_{p_J}$. Then, for any $A_i \in \mathcal{F}_i$, $i \in J$,

$$P\left((X_i)_{i\in J} \in \prod_{i\in J} A_i\right) = P_{(X_i)_{i\in J}}\left(\prod_{i\in J} A_i\right) = W_{p_J}\left(\prod_{i\in J} A_i\right) = \prod_{i\in J} P(X_i \in A_i),$$

which implies independence of $(X_i)_{i\in J}$, and, a fortiori, independence of $(X_i)_{i\in I}$. □

Remark 2.14. Any family of independent random variables $(X_i)_{i\in I}$ can be realized on the corresponding probability space $\bigotimes_{i\in I}(\Omega_i, \mathcal{F}_i, P_{X_i})$ in the sense that if we set $Y_i := p_i$, then $(Y_i)_{i\in I}$ are independent and have the same distributions as $(X_i)_{i\in I}$.

If we have a family of independent random variables $(X_i)_{i\in I}$ and transform them to $(Y_i)_{i\in I}$ with $Y_i = f_i(X_i)$, $i \in I$, then one would naturally expect that the transformed random variables are again independent.

[1] Recall that $\widetilde{\Omega} := \prod_{i\in I} \Omega_i$, while $\widetilde{\mathcal{F}} := \sigma(p_i; i \in I)$ is generated by the projections $p_j : \widetilde{\Omega} \to \Omega_j$, $\widetilde{\omega} = (\omega_i)_{i\in I} \mapsto \omega_j$, $j \in I$.

Theorem 2.15. *Let $(X_i)_{i\in I}$ be a family of independent random variables taking values in $(\Omega_i, \mathcal{F}_i)$, $i \in I$. Moreover, let $(I_j)_{j\in J}$ be a partition of the index set I consisting of non-empty sets and $(A_j, \mathcal{A}_j)$ be measurable spaces, $j \in J$. For $j \in J$ let $f_j : \prod_{i\in I_j} \Omega_i \to A_j$ be $\bigotimes_{i\in I_j}$-$\mathcal{A}_j$-measurable. Set*

$$Y_j \coloneqq f_j\left((X_i)_{i\in I_j}\right), \quad j \in J.$$

Then the collection $(Y_j)_{j\in J}$ is independent.

Proof. Let $Z_j \coloneqq (X_i)_{i\in I_j}$, $j \in J$, and notice that $\sigma(Y_j) \subset \sigma(Z_j)$. Thus, we are left with proving independence of the family $(Z_j)_{j\in J}$. By construction of the product σ-algebra, we have

$$\sigma(Z_j) = \sigma(X_i : i \in I_j) = \sigma\left(\bigcup_{i\in I_j} \sigma(X_i)\right).$$

Therefore, independence of the σ-algebras $\sigma(Z_j)$, $j \in J$, follows from independence of the σ-algebras $\sigma(X_i)$, $i \in I$, by Theorem 2.7. □

Example 2.16. Let $(X_{m,n})_{n,m\in\mathbb{N}}$ be a collection of independent random variables distributed according to the Bernoulli distribution with parameter $0 < p < 1$. Define

$$Y_m \coloneqq \inf\{\, n \in \mathbb{N} \mid X_{n,m} = 1 \,\} - 1, \ m \in \mathbb{N}.$$

Then the sequence of random variables $(Y_m)_{m\in\mathbb{N}}$ is independent, identically distributed according to the geometric distribution with parameter p, i.e.,

$$P(Y_m = k) = p(1-p)^k.$$

Indeed, consider the set $f_m : \{0,1\}^{\mathbb{N}} \to \mathbb{N}$ defined by $f_m(x) \coloneqq \inf\{\, n \in \mathbb{N} \mid x_n = 1 \,\}$ – we disregard the event $\{\, n \in \mathbb{N} \mid x_n = 1 \,\} = \emptyset$, which only has probability 0. Then f_m is $\mathcal{P}(\mathbb{N})$-$\mathcal{P}(\{0,1\})^{\otimes\mathbb{N}}$-measurable: by the structure of the power set, we only need to check that $f_m^{-1}(\{k\}) \in \mathcal{P}(\{0,1\})^{\otimes\mathbb{N}}$ for every $k \in \mathbb{N}$. By the construction of the product space, we have

$$f_m^{-1}(\{k\}) = p_{\{1,\dots,k\}}^{-1}(\{(0,\dots,0,1)\}) \in \mathcal{P}(\{0,1\})^{\otimes\mathbb{N}}.$$

Thus, independence of $(Y_m)_{m\in\mathbb{N}}$ follows from Theorem 2.15. Regarding the distribution, note that

$$P(Y_m > k) = P(X_{m,1} = \cdots = X_{m,k+1} = 0) = (1-p)^{k+1},$$

implying that $P(Y_m = k) = P(Y_m > k-1) - P(Y_m > k) = p(1-p)^k$.

From Theorem 2.13 one can derive, just like in the elementary case,[2] the

Corollary 2.17 (Hofbauer, MT, Satz 4.10 for $n = 2$)**.** *Given n independent positive or integrable random variables $X_1, \dots, X_n$. Then*

$$E\left[\prod_{i=1}^{n} X_i\right] = \prod_{i=1}^{n} E[X_i]. \tag{2.4}$$

In particular, if $X_1, \dots, X_n$ are integrable and independent, then $X_1 \cdots X_n$ is also integrable.

[2] Indeed, one first assumes the random variable to only take finitely many values. Then the distribution is actually discrete and the result is known. Then on approximates the general random variables $X_1, \dots, X_n$ by random variables taking finitely many values only (step functions) and concludes by the construction of the integral.

Proof. We give an alternative proof. Assume that $X_1, \dots, X_n$ are integrable. By Theorem 2.13, the joint distribution of $Z \coloneqq (X_1, \dots, X_n)$ is given by

$$P_Z = \bigotimes_{i=1}^{n} P_{X_i}.$$

Thus, Fubini's theorem implies that

$$\begin{aligned} E\left[|X_1 \cdots X_n|\right] &= \int_{\mathbb{R}^n} |x_1 \cdots x_n| \, P_Z(dx_1, \dots, dx_n) = \int_{\mathbb{R}^n} |x_1| \cdots |x_n| \, P_{X_1}(dx_1) \cdots P_{X_n}(dx_n) \\ &= \prod_{i=1}^{n} \int_{\mathbb{R}} |x_i| \, P_{X_i}(dx_i) = \prod_{i=1}^{n} E\left[|X_i|\right] < \infty. \end{aligned}$$

Having established integrability of the product, the very same argument without $|\cdot|$ shows the formula. The proof for non-negative random variables works in the same way. □

The converse is, of course, not true: given random variables $X_1, \dots, X_n$ satisfying equation (2.4), they do not have to be independent. On the other hand, the product formula (2.4) implies that, as in elementary probability, independent random variables are uncorrelated. In particular, for all square-integrable, independent random variables $X_1, \dots, X_n$, we have *Bienaymé's equality*

$$V[X_1 + \cdots + X_n] = V[X_1] + \cdots + V[X_n]. \tag{2.5}$$

There is one important case, however, where the product formula (2.4) indeed implies independence.

Corollary 2.18. *Given an n-dimensional Gaussian random vector $X = (X_1, \dots, X_n)$. Then $X_1, \dots, X_n$ are independent if and only if they are uncorrelated.*

We postpone the proof of Corollary 2.18 to the end of the section.

Remark 2.19. It is essential that $X_1, \dots, X_n$ are *jointly* normally distributed. Indeed, it is possible to construct two Gaussian random variables X and Y, such that (X, Y) is not Gaussian, X and Y are uncorrelated, but not independent.

Exercise 2.20. Given a standard Gaussian random variable X and a constant $c > 0$, define a random variable Y_c by

$$Y_c(\omega) \coloneqq \begin{cases} X(\omega), & |X(\omega)| \le c, \\ -X(\omega), & |X(\omega)| > c. \end{cases}$$

Show that $Y_c \sim \mathcal{N}(0, 1)$ again. Moreover, show that there is a positive real number c such that $\mathrm{Cov}[X, Y_c] = 0$ – for instance, one can apply the intermediate value theorem to the continuous function $c \mapsto \mathrm{Cov}[X, Y_c]$. However, X and Y_c are not independent (why not?). Is this a contradiction to Corollary 2.18?

Next we consider the distribution of sums of independent random variables.

Definition 2.21. Given to probability measures μ and ν on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$. Let $T(x, y) \coloneqq x + y$, $x, y \in \mathbb{R}^n$, then the *convolution* of μ and ν is defined as the image measure of $\mu \otimes \nu$ under the map T, symbolically

$$\mu * \nu \coloneqq (\mu \otimes \nu)_T.$$

So if X and Y are independent random variables taking values in $\mathbb{R}^n$, then the distribution of $X + Y$ is given by the convolution of the distributions of X and Y, i.e., $P_{X+Y} = P_X * P_Y$. We obtain the following formulas for distribution functions and densities:

Theorem 2.22. *Let X and Y be $\mathbb{R}^n$-valued independent random variables with distribution functions F_X and F_Y, respectively. Then their sum has distribution function F_{X+Y} with*

$$F_{X+Y}(x) = \int_{\mathbb{R}^n} F_Y(x-y)dF_X(y) = \int_{\mathbb{R}^n} F_X(x-y)dF_Y(y).$$

Moreover, assume that X and Y have densities f_X and f_Y. Then $X + Y$ has a density f_{X+Y}, given by the convolution of the densities f_X and f_Y, i.e.,

$$f_{X+Y}(x) = f_X * f_Y(x) = \int_{\mathbb{R}^n} f_X(x-y)f_Y(y)dy = \int_{\mathbb{R}^n} f_X(y)f_Y(x-y)dy. \tag{2.6}$$

Proof. Let $\mu := P_X$ and $\nu := P_Y$. Then Fubini's theorem implies that

$$\begin{aligned} F_{X+Y}(t) &= \int_{\mathbb{R}^2} \mathbf{1}_{\{x+y\le t\}}(\mu\otimes\nu)(d(x,y)) \\ &= \int_{\mathbb{R}} \left(\int_{\mathbb{R}} \mathbf{1}_{]-\infty,t-y]}(x)\mu(dx)\right)\nu(dy) \\ &= \int_{\mathbb{R}} F_X(t-y)dF_Y(y). \end{aligned}$$

In the case of X and Y having densities, we can insert the densities and apply Fubini's theorem again, to obtain

$$\begin{aligned} F_{X+Y}(t) &= \int_{\mathbb{R}} \left(\int_{-\infty}^{t-y} f_X(x)dx\right) f_Y(y)dy \\ &= \int_{\mathbb{R}} \left(\int_{-\infty}^{t} f_X(x-y)dx\right) f_Y(y)dy \\ &= \int_{-\infty}^{t} \left(\int_{\mathbb{R}} f_X(x-y)f_Y(y)dy\right) dx. \end{aligned}$$

□

Exercise 2.23. Given two probability measures μ and ν defined on $(\mathbb{Z}, \mathcal{P}(\mathbb{Z}))$. Derive a formula for their convolution $\mu * \nu$. Apply the formula to find out the distribution of the sum of two independent random variables $X \sim \mathrm{Poi}_\lambda$ and $Y \sim \mathrm{Poi}_\mu$.

Given two random variables X and Y such that $P_{X+Y} = P_X * P_Y$, can we infer that X and Y need to be independent? As the following example shows, the answer is "no" in general.

Example 2.24. Let γ_α denote the Cauchy-distribution with parameter $\alpha > 0$, i.e., the distribution with density

$$f_\alpha(x) = \frac{1}{\pi\alpha\,(1 + (x/\alpha)^2)}$$

and characteristic function

$$\widehat{\gamma_\alpha}(u) = e^{-\alpha|u|}.$$

Then one can easily show that $\gamma_\alpha * \gamma_\beta = \gamma_{\alpha+\beta}$. On the other hand, let $T_\beta(x) := \beta x, \beta > 0$. Then $(\gamma_\alpha)_{T_\beta} = \gamma_{\alpha\beta}$. In particular, let $X \sim \gamma_1$. Then $P_{2X} = \gamma_2 = P_X * P_X$, but X is not independent of itself.

As a probabilistic notion, independence can be expressed in terms of the characteristic function. We use the notation $\varphi_X(u) = E[e^{i\langle u, X\rangle}]$ for the characteristic function of an $\mathbb{R}^n$-valued random variable X and $\widehat{\mu}$ for the Fourier transform of the probability measure μ, i.e., $\widehat{P_X} = \varphi_X$.

Theorem 2.25.

(i) The real random variables $X_1, \dots, X_n$ are independent if and only if the characteristic function of the random vector $X = (X_1, \dots, X_n)$ is the tensor product of the characteristic functions of $X_1, \dots, X_n$, i.e.,

$$\varphi_X(u) = \varphi_{X_1}(u_1) \cdots \varphi_{X_n}(u_n), \quad \forall u = (u_1, \dots, u_n) \in \mathbb{R}^n.$$

(ii) For two probability measure μ and ν on $\mathbb{R}^n$ we have

$$\widehat{\mu * \nu} = \widehat{\mu}\,\widehat{\nu}.$$

Proof. If $X_1, \dots, X_n$ are independent, then the formula for the characteristic function follows from

$$e^{i\langle u, X\rangle} = e^{i\sum_{j=1}^n u_j X_j} = e^{iu_1 X_1} \cdots e^{iu_n X_n}$$

and Corollary 2.17. The converse direction follows from the fact that the characteristic function characterizes a distribution and Theorem 2.13.

For the second part, we construct independent random variables X and Y with $P_X = \mu$ and $P_Y = \nu$. Then

$$\widehat{\mu * \nu}(u) = E\left[e^{i\langle u, X+Y\rangle}\right] = E\left[e^{i\langle u, X\rangle} e^{i\langle u, Y\rangle}\right] = E\left[e^{i\langle u, X\rangle}\right] E\left[e^{i\langle u, Y\rangle}\right] = \widehat{\mu}(u)\widehat{\nu}(u). \qquad \square$$

Proof of Corollary 2.18. By Lemma A.3, the characteristic function φ_X is of the form $e^{i\langle \mu, u\rangle - \frac{1}{2}\langle u, \Sigma u\rangle}$, where $\mu = E[X]$ and $\Sigma = \mathrm{Cov}[X]$. If $X_1, \dots, X_n$ are uncorrelated, then $\Sigma = \mathrm{diag}(\sigma_1^2, \dots, \sigma_n^2)$ for $\sigma_i^2 = V[X_i]$. Thus,

$$\varphi_X(u) = e^{i\sum_{j=1}^n \mu_j u_j - \frac{1}{2}\sum_{j=1}^n \sigma_j^2 u_j^2} = \prod_{j=1}^n e^{i\mu_j u_j - \frac{1}{2}\sigma_j^2 u_j^2} = \prod_{j=1}^n \varphi_{X_j}(u_j). \qquad \square$$

2.3 The Borel-Cantelli lemma and 0-1-laws

Recall the definition of the limes superior of a sequence of events $(A_n)_{n\in\mathbb{N}}$

$$\limsup_{n\to\infty} A_n = \bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} A_m = \{\, \omega \in \Omega \mid \#\{\, n \in \mathbb{N} \mid \omega \in A_n \,\} = \infty \,\},$$

i.e., $\omega \in \limsup_{n\to\infty} A_n$ if and only if ω is contained in infinitely many sets of the sequence (A_n).

Theorem 2.26 (Borel-Cantelli)**.** *Let $(A_n)_{n\in\mathbb{N}}$ be a sequence of events from the probability space $(\Omega, \mathcal{F}, P)$.*

(i) If $\sum_{n=1}^{\infty} P(A_n) < \infty$, then $P\left(\limsup_{n\to\infty} A_n\right) = 0$.

(ii) If $(A_n)_{n\in\mathbb{N}}$ is independent and $\sum_{n\in\mathbb{N}} P(A_n) = \infty$, then $P\left(\limsup_{n\to\infty} A_n\right) = 1$.

Proof. For the proof of (i), we note that $A \coloneqq \limsup_{n\to\infty} A_n \subset \bigcup_{m=n}^{\infty} A_m$. By subadditivity, this implies that

$$P(A) \leq \sum_{m=n}^{\infty} P(A_m) \xrightarrow{n\to\infty} 0.$$

For (ii), consider A^c and note that

$$A^c = \bigcup_{n=1}^{\infty} \bigcap_{m=n}^{\infty} A_m^c = \lim_{n\to\infty} \bigcap_{m=n}^{\infty} A_m^c$$

is the limit of an increasing sequence of sets, which implies (continuity from below and independence) that

$$P(A^c) = \lim_{n\to\infty} P\left(\bigcap_{m=n}^{\infty} A_m^c\right) = \lim_{n\to\infty} \prod_{m=n}^{\infty} [1 - P(A_m)].$$

For $0 \leq x \leq 1$ an elementary calculation shows that $\log(1-x) \leq -x$. Now we fix some n and take the logarithm in the above equation and obtain (with the convention that $0 = \exp(\log(0))$)

$$P(A^c) \leq \exp\left(\sum_{m=n}^{\infty} \log(1 - P(A_m))\right) \leq \exp\left(\sum_{m=n}^{\infty} -P(A_m)\right) = 0. \qquad \square$$

Definition 2.27. Given a sequence $\mathcal{F}_n \subset \mathcal{F}$ of σ-algebras, we define the *tail σ-algebra*

$$\mathcal{T} \coloneqq \bigcap_{n=1}^{\infty} \sigma\left(\bigcup_{m=n}^{\infty} \mathcal{F}_m\right).$$

Any event $A \in \mathcal{T}$ is called *tail event*.

Moreover, given a sequence of random variables $(X_n)_{n\in\mathbb{N}}$ we denote by $\mathcal{T}((X_n)_{n\in\mathbb{N}})$ the tail σ-algebra generated by the sequence $\sigma(X_n)$, $n \in \mathbb{N}$.

Example 2.28. Let $A_n \in \mathcal{F}_n$ be a sequence of events and consider $A \coloneqq \limsup_{n\to\infty} A_n$. Then $A \in \mathcal{T}$. Indeed, for any fixed N, we have

$$A = \bigcap_{n\in\mathbb{N}} \bigcup_{m\geq n} A_m = \bigcap_{n\geq N} \underbrace{\bigcup_{m\geq n} A_m}_{\in\sigma\left(\bigcup_{m=n}^{\infty} \mathcal{F}_m\right)} \in \sigma\left(\bigcup_{m=N}^{\infty} \mathcal{F}_m\right).$$

Thus, $A \in \mathcal{T}$.

Example 2.29. Given a sequence of real random variables $(X_n)_{n\in\mathbb{N}}$. Then both $\limsup_{n\to\infty} X_n$ and $\limsup_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} X_k$ are measurable with respect to $\mathcal{T}((X_n))$. The idea in both cases is that the lim sup only depends on the tail-behavior of the sequence. Let us be more precise in the second case. We have

$$\limsup_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} X_k = \limsup_{n\to\infty} \frac{1}{n}\sum_{k=N}^{n} X_k,$$

which is clearly $\sigma\left(\bigcup_{k=N}^{\infty} \sigma(X_k)\right)$-measurable, for any fixed N. Thus, $\limsup_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} X_k$ is $\mathcal{T}$-measurable.

Theorem 2.30 (Kolmogorov's 0-1-law)**.** *Let $(\mathcal{F}_n)_{n\in\mathbb{N}}$ be a sequence of independent σ-algebras. Then the tail σ-algebra $\mathcal{T}$ is P-trivial, i.e.,*

$$\forall A \in \mathcal{T} : \; P(A) \in \{0, 1\}.$$

Proof. Choose $A \in \mathcal{T}$ and let $\mathcal{D}$ be the set of all events $D \in \mathcal{F}$ which are independent of A, i.e., $P(A \cap D) = P(A) \cap P(D)$. We want to prove that $A \in \mathcal{D}$, because then

$$P(A) = P(A \cap A) = P(A)^2,$$

implying that either $P(A) = 0$ or $P(A) = 1$. Arguing as in the proof of Lemma 2.5, we see that $\mathcal{D}$ is a Dynkin system. By definition of $\mathcal{T}$, for any n we have that $\sigma(\mathcal{F}_1 \cup \cdots \cup \mathcal{F}_n)$ is independent of $\mathcal{T}$, and thus

$$\bigcup_{n=1}^{\infty} \sigma(\mathcal{F}_1 \cup \cdots \cup \mathcal{F}_n) \subset \mathcal{D}.$$

Notice that the union on the left side is a union of an increasing sequence of σ-algebras. Therefore, it is closed under intersections. Thus, we have

$$\mathcal{A} := \sigma\left(\bigcup_{n=1}^{\infty} \sigma(\mathcal{F}_1 \cup \cdots \cup \mathcal{F}_n)\right) = \delta\left(\bigcup_{n=1}^{\infty} \sigma(\mathcal{F}_1 \cup \cdots \cup \mathcal{F}_n)\right) \subset \mathcal{D}.$$

Since any $\mathcal{F}_n \subset \mathcal{A} \subset \mathcal{D}$, this implies that $\mathcal{T} \subset \mathcal{A} \subset \mathcal{D}$. □

Kolmogorov's 0-1-law sheds some new light on the second part of the Borel-Cantelli lemma: indeed, when (A_n) are independent, $A = \limsup_n A_n$ is a tail event, and, thus, can only have probability 0 or 1. If we have a sequence of independent random variables (X_n), then by Example 2.29 $\limsup X_n$ and, likewise, $\liminf X_n$ are measurable with respect to the tail σ-algebra. Therefore, those random variables are almost surely constant. In particular, the limit of a sequence of independent random variables either exists with probability 0 or with probability 1, and in the latter case is almost surely constant. The same applies to the Cesàro means.

Remark 2.31. There is a more general theorem called *Hewitt-Savage 0-1-law*, which says that any event depending on a sequence of independent random variables (X_n), which is invariant under finite permutations of the indices of (X_n), has probability 0 or 1. In particular, define the cumulative sums (or the random walk) $S_n := \sum_{k=1}^{n} X_k$, $n \in \mathbb{N}$. Clearly, $\limsup_n S_n$ does not depend on finite permutations of the indices. Thus, $\limsup_n S_n$ and, likewise, $\liminf_n S_n$ are almost surely constant. Notice, however, that they are not measurable with respect to the tail σ-algebra.

Example 2.32. Let (X_n) be an independent sequence of random variables distributed with the Cauchy distribution with $\alpha = 1$. Then, by Example 2.24 we know that the scaled random walk $\frac{1}{n}S_n$ has again a Cauchy distribution with parameter $\alpha = 1$. Consider $\overline{X} := \limsup_{n\to\infty} S_n/n$ and $\underline{X} := \liminf_{n\to\infty} S_n/n$. Then for any $c \in \mathbb{R}$ we have

$$0 < \int_c^{\infty} \frac{1}{\pi} \frac{1}{1+x^2} dx = P\left(\frac{S_n}{n} \geq c\right) \leq P\left(\sup_{k\geq n} \frac{S_k}{k} \geq c\right)$$
$$\xrightarrow[n\to\infty]{} P\left(\limsup_{n\to\infty} \frac{S_n}{n} \geq c\right) = P\left(\overline{X} \geq c\right),$$

where we have used that $\{ \sup_{k \geq n} S_k/k \geq c \}$ is a decreasing sequence of sets and that $\lim_{n\to\infty} \sup_{k\geq n} x_k = \limsup_{n\to\infty} x_n$. By Theorem 2.30, this implies that $P\left(\overline{X} \geq c\right) = 1$ for every real number c, and we may conclude that $\overline{X} = \infty$ almost surely. In the same manner, we get $\underline{X} = -\infty$ almost surely.

Chapter 3

The strong law of large numbers

3.1 The strong law of large numbers

We consider a sequence of random variables $(X_n)_{n\in\mathbb{N}}$ and denote

$$S_n := \sum_{i=1}^{n} X_i, \quad S_n^\star := \frac{1}{n} S_n, \quad n \in \mathbb{N}. \tag{3.1}$$

We want to study the asymptotic properties of the re-scaled random walk $S_n^\star$. We will at least assume that random variables are identically distributed and integrable.

Definition 3.1. The sequence $(X_n)_{n\in\mathbb{N}}$ satisfies the *strong (weak) law of large numbers* if $\lim_{n\to\infty} S_n^\star = E[X_1]$ almost surely (in probability).

In this chapter, we will prove two versions of the theorem, i.e., we will prove that the strong law of large numbers holds under two sets of conditions on the random variables. The first formulation is rather easy to prove, but far from optimal.

Theorem 3.2. *Let the random variables (X_n) be identically distributed, independent and assume that $E[X_1^4] < \infty$. Then the sequence satisfies the strong law of large numbers.*

Proof. By passing to a sequence $Y_n := X_n - E[X_1]$ if necessary, we may assume that $E[X_1] = 0$. Then let

$$A_n := \left\{ \left| S_n^\star \right| \geq n^{-1/8} \right\}, \quad A := \limsup_{n\to\infty} A_n.$$

Notice that $S_n^\star(\omega) \to 0$ for every $\omega \in A^c$.

Markov's inequality (with $u(x) = x^4$) implies that

$$P(A_n) \leq \frac{\sqrt{n}}{n^4} E\left[S_n^4\right]. \tag{3.2}$$

We now compute the fourth moment of S_n:

$$\begin{aligned} E[S_n^4] &= E\left[\left(\sum_{i=1}^n X_i\right)^4\right] = E\left[\sum_{i_1,i_2,i_3,i_4=1}^n X_{i_1}X_{i_2}X_{i_3}X_{i_4}\right] = \sum_{i_1,i_2,i_3,i_4=1}^n E[X_{i_1}X_{i_2}X_{i_3}X_{i_4}] \\ &= \sum_{i=1}^n E[X_i^4] + 3\sum_{i\neq j} E[X_i^2X_j^2] + \sum_{\substack{i_1,i_2,i_3,i_4\in\{1,\dots,n\}\\ i_1\notin\{i_2,i_3,i_4\}}} \underbrace{E[X_{i_1}X_{i_2}X_{i_3}X_{i_4}]}_{=E[X_{i_1}]E[X_{i_2}X_{i_3}X_{i_4}]=0} \\ &= nE[X_1^4] + 3n(n-1)E[X_1^2]^2. \end{aligned}$$

Inserting this term into (3.2) and summing over n, we obtain

$$\sum_{n=1}^\infty P(A_n) \leq E[X_1^4]\sum_{n=1}^\infty \frac{n^{3/2}}{n^4} + 3E[X_1^2]^2\sum_{n=1}^\infty \frac{n^{5/2}-n^{3/2}}{n^4} < \infty.$$

By the Borel-Cantelli lemma (Theorem 2.26) we get that $P(A) = 0$. □

Next, we cite a much more general version of the strong law of large numbers, which we are not going to prove.

Theorem 3.3 (Etemadi)**.** *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of pairwise independent and identically distributed integrable random variables. Then the strong law of large numbers holds.*

The basic idea of the proof of Theorem 3.3 is to cut the random variables, i.e., to consider random variables $Y_n := X_n\mathbf{1}_{[-n,n]}(X_n)$ instead. The sequence (Y_n) satisfies the conditions of Theorem 3.2 (except for only pairwise independence), thus the strong law of large numbers holds for (Y_n). Finally, one shows that $\sum_n P(X_n \neq Y_n) < \infty$, which, by the Borel-Cantelli lemma, implies that $X_n = Y_n$ for almost all indices almost surely. This will then allow to conclude.

With completely different methods we shall give a proof of the following version of the strong law of large numbers.

Theorem 3.4. *Assume that the sequence $(X_n)_{n\in\mathbb{N}}$ is independent and identically distributed and the random variables are integrable. Then the strong law of large numbers holds.*

Remark 3.5. One common feature of all our conditions is that we always assume a sequence of identically distributed random variables. At least in the framework of Theorem 3.4, we know, however, that $\lim_n S_n^\star$ mus be almost surely constant, provided that the limit exists – this follows from Theorem 2.30. Therefore, it makes sense to generalize the strong law of large number to the statement that

$$\frac{1}{n}\sum_{k=1}^n (X_k - E[X_k]) = 0, \text{ almost surely.}$$

Kolmogorov proved that the above version of the strong law of large numbers holds whenever (X_n) are independent, square integrable and

$$\sum_{n=1}^\infty \operatorname{var}[X_n]/n^2 < \infty.$$

Moreover, this condition is optimal in the sense that for any sequence (α_n) of positive numbers satisfying $\sum_n \alpha_n/n^2 = \infty$, there is a sequence of independent, square integrable random variables (X_n) with $\mathrm{var}[X_n] = \alpha_n$ such that the law of large numbers does not hold.

Remark. One might also wonder whether integrability is really necessary. However, given a sequence (X_n) of independent, identically distributed random variables, one can show that the almost sure convergence of $S_n^\star$ to some deterministic number μ already implies that X_1 is integrable and $\mu = E[X_1]$ – see Bauer, Satz 12.3. Indeed, in Example 2.32, we have seen that $S_n^\star$ does not converge to a deterministic number, when $X_1 \sim \gamma_1$. This is in line with the fact that the Cauchy distribution does not have a mean value.

Remark 3.6. Let (X_n) be a sequence of i.i.d. square integrable random variables with $E[X_1] = 0$ and consider, as usual, the random walk $S_n := X_1 + \cdots + X_n$. We are interested in the asymptotic behavior of the paths of the random walk. By the law of large numbers, we know that

$$\frac{1}{n} S_n \to 0 \text{ almost surely,}$$

but is this the best possible result? A sharper bound would be given by the scaling factor $1/\sqrt{n}$, but in that case we only have the central limit theorem,

$$\frac{1}{\sqrt{n}} S_n \xrightarrow[n\to\infty]{d} \mathcal{N}(0, V[X_1]).$$

If we want to have almost sure results, then one can show that, for instance, in the normal case,

$$\liminf_{n\to\infty} \frac{1}{\sqrt{n}} S_n = -\infty, \ \limsup_{n\to\infty} \frac{1}{\sqrt{n}} S_n = \infty, \text{ almost surely,}$$

implying that $\sqrt{n}$ is too small. The precise asymptotics in the almost sure sense is given by the *law of the iterated logarithm*

$$\limsup_{n\to\infty} \frac{S_n}{\sqrt{2n \log(\log(n))}} = \sqrt{\mathrm{var}[X_1]} \text{ a.s.,} \tag{3.3a}$$

$$\liminf_{n\to\infty} \frac{S_n}{\sqrt{2n \log(\log(n))}} = -\sqrt{\mathrm{var}[X_1]} \text{ a.s.} \tag{3.3b}$$

3.2 Large deviations

Let $(X_n)_{n\in\mathbb{N}}$ denote a sequence of integrable i.i.d. (independent, identically distributed) random variables. Set $\mu := E[X_1]$ and consider the random walk S_n and its scaled version $S_n^\star$ as defined in equation (3.1). By Theorem 3.4, the strong law of large number holds,which in particular implies the weak law of large numbers, i.e.,

$$\forall \epsilon > 0 : \lim_{n\to\infty} P\left(\left|S_n^\star - \mu\right| \geq \epsilon\right) = 0. \tag{3.4}$$

The theory of large deviations is concerned with the speed of the convergence in the weak law of large numbers, i.e., it tries to find estimates for $P(\left|S_n^\star - \mu\right| \geq \epsilon)$ for finite n. Consider the (generalized) Laplace transform of P_{X_1} given by

$$\phi(u) := E[e^{uX_1}],$$

which is defined for $u \in \mathbb{R}$ as a function taking values in $]0, \infty]$. (In that sense, we do not need any integrability condition.) We define the *Cramér transformation* as

$$I(x) \coloneqq \sup_{u \in \mathbb{R}} (ux - \log \phi(u)), \quad x \in \mathbb{R}. \tag{3.5}$$

(This is the Legendre transform of $\log \phi$.) Since $\phi(0) = 1$, we see that

$$I(x) = \sup_{u \in \mathbb{R}} (ux - \log \phi(u)) \geq 0 - \log \phi(0) = 0,$$

which implies that $I : \mathbb{R} \to [0, \infty]$. By Jensen's inequality and convexity of $x \mapsto e^{ux}$,

$$e^{u\mu} \leq E[e^{uX_1}] = \phi(u)$$

implying that

$$\forall u \in \mathbb{R} : \ u\mu - \log \phi(u) \leq 0. \tag{3.6}$$

Therefore, $I(\mu) = 0$.

Example 3.7. If $X_1 \equiv 0$, then $\phi(u) \equiv 1$ and $I(x) = \infty$ for every $x \neq 0$. Thus, I can indeed take the value ∞.

Theorem 3.8 (Cramér-Chernov)**.** *Under the above assumptions, we have for $\xi \geq \mu$ and any $n \in \mathbb{N}$*

$$P(S_n^\star \geq \xi) \leq e^{-I(\xi)n}$$

and for $\xi \leq \mu$ and any n

$$P(S_n^\star \leq \xi) \leq e^{-I(\xi)n}.$$

Proof. First of all, note that the second assertion follows from the first assertion by replacing (X_n) by $(-X_n)$ (and noting that the corresponding Cramér transformation is $x \mapsto I(-x)$). Moreover, we may replace (X_n) by $(X_n - \xi)$, which gives the Cramér transform $x \mapsto I(x + \xi)$, which allows to infer then first assertion for general $\xi \geq \mu$ from the corresponding assertion for $\xi = 0$ and $\mu \leq 0$. Thus, in the following we may assume that $\xi = 0$ and $\mu \leq 0$.

By independence, we get for $u \geq 0$

$$\begin{aligned} P(S_n^\star \geq 0) &= P(X_1 + \cdots + X_n \geq 0) \\ &= P\left(e^{u(X_1 + \cdots + X_n)} \geq 1\right) \\ &\leq E\left[e^{u(X_1 + \cdots + X_n)}\right] \\ &= \phi(u)^n, \end{aligned}$$

implying that $\log P(S_n^\star \geq 0) \leq n \log \phi(u)$, for any $u \geq 0$. Therefore,

$$\frac{1}{n} \log P(S_n^\star \geq 0) \leq \inf \{ \log \phi(u) \,\big|\, u \geq 0 \}.$$

On the other hand, from (3.6) we obtain for any $u < 0$

$$-\log \phi(u) \leq u\mu - \log \phi(u) \leq 0.$$

On the other hand, $I(0) \geq 0$. Therefore, we have

$$I(0) = \sup \{ \underbrace{-\log \phi(u)}_{\leq 0 \text{ for } u<0} \mid u \geq 0 \} = -\inf \{ \log \phi(u) \,\big|\, u \geq 0 \}.$$

Therefore, we can conclude that

$$P(S_n^\star \geq 0) \leq \exp\left(n \inf\{ \log \phi(u) \;\middle|\; u \geq 0 \}\right) = e^{-nI(0)}. \qquad \square$$

Exercise 3.9. Convince yourself of the remarks in the first paragraph of the proof of Theorem 3.8.

Thus, the "exceptional" probabilities $P(S_n^\star \geq E[X_1] + \epsilon)$ decrease exponentially fast in n. Note that the right hand side of the inequalities in Theorem 3.8 should be read as "0" when $I(\xi) = \infty$.

3.3 The ergodic theorem

If we interpret the the index n as time, one can describe the strong law of large numbers as "time average = space average". Ergodic theory is concerned with this kind of phenomena in more generality.

Definition 3.10. A sequence of random variables $(X_n)_{n\in\mathbb{N}}$ is called *stationary* if the law of $(X_n)_{n\in\mathbb{N}}$ is equal to the law of $(X_{n+1})_{n\in\mathbb{N}}$, i.e., if $P_{(X_n)} = P_{(X_{n+1})}$.

Of course, this also means that the law of $(X_n)_{n\in\mathbb{N}}$ is equal to the law of $(X_{n+k})_{n\in\mathbb{N}}$ for any $k \in \mathbb{N}$. We start with two fundamental definitions for ergodic theory. In the following, $\tau : \Omega \to \Omega$ is an $\mathcal{F}$-$\mathcal{F}$-measurable map.

Definition 3.11. An event $A \in \mathcal{F}$ is called *invariant* if $\tau^{-1}(A) = A$. The σ-algebra of invariant events is denoted by $\mathcal{I} = \left\{ A \in \mathcal{F} \;\middle|\; \tau^{-1}(A) = A \right\}$.

At the level of random variables, X is $\mathcal{I}$-measurable if and only if $X \circ \tau = X$.

Definition 3.12. τ is called *measure preserving* if for every $A \in \mathcal{F}$ we have $P(\tau^{-1}(A)) = P(A)$. If, in addition, $\mathcal{I}$ is P-trivial, i.e., $P(A) \in \{0, 1\}$ for every $A \in \mathcal{I}$, then the system $(\Omega, \mathcal{F}, P, \tau)$ is called *ergodic*.

Obviously, ergodicity is equivalent to the property that every $\mathcal{I}$-measurable random variable is almost surely constant.

Theorem 3.13 (Birkhoff's individual ergodic theorem)**.** *Let τ be a measure preserving transformation and X be an integrable random variable. Define $X_1 := X$ and $X_n := X \circ \tau^{n-1}$, $n \geq 2$. Then there is an $\mathcal{I}$-measurable random variable Y with $E[Y] = E[X]$ such that*

$$S_n^\star := \frac{1}{n}\sum_{k=1}^{n} X_k \xrightarrow{n\to\infty} Y \text{ } P\text{-almost surely.}$$

If τ is ergodic, then $S_n^\star \to E[X]$ a.s. for $n \to \infty$.

Proof. If τ is ergodic, then every $\mathcal{I}$-measurable random variable is almost surely constant, which shows the last part of the theorem.

We assume, without less of generality, that $X \geq 0$. Let $\overline{X} := \limsup S_n^\star$ and $\underline{X} := \liminf S_n^\star$. Clearly, $\overline{X} \circ \tau = \overline{X}$ and $\underline{X} \circ \tau = \underline{X}$, which implies that both random variables are $\mathcal{I}$-measurable. We want to show that

$$E\left[\overline{X}\right] \leq E[X] \leq E\left[\underline{X}\right]. \tag{3.7}$$

Indeed, equation (3.7) together with $\underline{X} \leq \overline{X}$ would show that $\underline{X} = \overline{X} = \lim S_n^\star$ almost surely, and $E[\overline{X}] = E[X]$ by monotone convergence – note that $E[S_n^\star] = E[X]$ for any n. This will finish the proof.

In fact, we only prove the first inequality – the second one works along the same lines. First we need to make sure that $\overline{X} < \infty$ by cutting it off: for $M > 0$ set $\overline{X}_M \coloneqq \min(\overline{X}, M)$. For given $\epsilon > 0$ define the random variable N_ϵ as

$$N_\epsilon \coloneqq \inf\left\{ n \in \mathbb{N} \;\middle|\; S_n^\star \geq \overline{X}_M - \epsilon \right\} < \infty. \tag{3.8}$$

Indeed, N_ϵ is measurable, because

$$\{ N_\epsilon = n \} = \bigcap_{k=1}^{n-1} \{ S_k^\star < \overline{X}_M - \epsilon \} \cap \{ S_n^\star \geq \overline{X}_M - \epsilon \} \in \mathcal{F}, \ \forall n \in \mathbb{N}.$$

While N_ϵ can be unbounded, we certainly have $\bigcap_{K\in\mathbb{N}} \{ N_\epsilon > K \} = \emptyset$, which implies that $\lim_K P(\{ N_\epsilon > K \}) = 0$, because the sequence $\{ N_\epsilon > K \}$ is decreasing in K. Thus, there is a $K_\epsilon \in \mathbb{N}$ with $P(\{ N_\epsilon > K_\epsilon \}) \leq \epsilon$. Define random variables

$$\widetilde{X} \coloneqq \begin{cases} X, & N_\epsilon \leq K_\epsilon, \\ M, & \text{else}, \end{cases} \qquad \widetilde{N}_\epsilon \coloneqq \begin{cases} N_\epsilon, & N_\epsilon \leq K_\epsilon, \\ 1, & \text{else}. \end{cases} \tag{3.9}$$

Note that $X(\omega) > M$ implies that $S_1^\star(\omega) = X(\omega) > \overline{X}_M(\omega) - \epsilon$, giving $N_\epsilon(\omega) = 1$. Therefore,

$$X \leq \widetilde{X}.$$

We claim that

$$\frac{1}{\widetilde{N}_\epsilon} \sum_{n=1}^{\widetilde{N}_\epsilon} \widetilde{X} \circ \tau^{n-1} \geq \overline{X}_M - \epsilon. \tag{3.10}$$

The inequality holds in the case of $N_\epsilon(\omega) > K_\epsilon$ by construction of $\widetilde{X}(\omega)$ and $\widetilde{N}_\epsilon(\omega)$. In the other case, note that $\widetilde{N}_\epsilon(\omega) = N_\epsilon(\omega)$ and $\widetilde{X}(\omega) = X(\omega)$, and so (3.10) holds by definition of N_ϵ.

On the other hand, we have by definition of K_ϵ that

$$E\left[\widetilde{X}\right] = E\left[\widetilde{X}\mathbf{1}_{N_\epsilon > K_\epsilon}\right] + E\left[\widetilde{X}\mathbf{1}_{N_\epsilon \leq K_\epsilon}\right] \leq M\epsilon + E[X]. \tag{3.11}$$

Inductively define $n_0(\omega) \coloneqq 0$, $n_1(\omega) \coloneqq \widetilde{N}_\epsilon(\omega)$ and

$$n_{k+1}(\omega) \coloneqq n_k(\omega) + \widetilde{N}_\epsilon(\tau^{n_k(\omega)}(\omega)), \quad k \in \mathbb{N}.$$

Moreover, for $l \in \mathbb{N}$ let $M_l(\omega) \coloneqq \sup\{ k \in \mathbb{N} \mid n_k(\omega) \leq l \}$. Since $\widetilde{N}_\epsilon \leq K_\epsilon$,we have $l - n_{M_l} \leq K_\epsilon$. Note that

$$\sum_{k=1}^{l} \widetilde{X} \circ \tau^{k-1} \geq \sum_{k=1}^{n_{M_l}} \widetilde{X} \circ \tau^{k-1} = \sum_{k=1}^{n_1} \widetilde{X} \circ \tau^{k-1} + \sum_{k=n_1+1}^{n_2} \widetilde{X} \circ \tau^{k-1} + \cdots + \sum_{k=n_{M_l-1}+1}^{n_{M_l}} \widetilde{X} \circ \tau^{k-1}.$$

Applying the inequality (3.10) to each of the terms in the above sum, we obtain

$$\sum_{k=1}^{l} \widetilde{X} \circ \tau^{k-1} \geq n_1(\overline{X}_M - \epsilon) + (n_2 - n_1)(\overline{X}_M \circ \tau^{n_1} - \epsilon) + \cdots + (n_{M_l} - n_{M_l-1})(\overline{X}_M \circ \tau^{n_{M_l}-1} - \epsilon),$$

noting that $n_k - n_{k-1} = \widetilde{N}_\epsilon \circ \tau^{n_{k-1}}$. But $\overline{X}_M$ is $\mathcal{I}$-measurable, implying that $\overline{X}_m \circ \tau^n = \overline{X}_M$, even when n is itself random. Thus, we have a telescoping sum and further obtain

$$\sum_{k=1}^{l} \widetilde{X} \circ \tau^{k-1} \geq n_{M_l}\overline{X}_M - n_{M_l}\epsilon \geq l\overline{X}_M + (n_{M_l} - l)\overline{X}_M - l\epsilon \geq l\overline{X}_M - K_\epsilon M - l\epsilon.$$

Since τ is measure preserving, we have $E[\widetilde{X} \circ \tau^k] = E[\widetilde{X}]$. Thus, by dividing the previous inequality by l, we get

$$E\left[\widetilde{X}\right] \geq E\left[\overline{X}_M\right] - \frac{K_\epsilon M}{l} - \epsilon$$

and (3.11) implies that

$$E[X] \geq E\left[\overline{X}_M\right] - \frac{K_\epsilon M}{l} - \epsilon - M\epsilon,$$

for all choices of M, ϵ and l. Therefore, we have $E[X] \geq E\left[\overline{X}_M\right]$ for any M, which implies the first inequality (3.7) by monotone convergence. □

Remark. In fact, one can show that the random variable Y in Theorem 3.13 can be chosen to be $E[X|\mathcal{I}]$.

Remark. Von Neumann's *statistical ergodic theorem* shows that the convergence in Theorem 3.13 also holds in $L^p(\Omega)$, if $X \in L^p$.

Next we want to show that Theorem 3.4 can be derived from Theorem 3.13. Let us first describe the setting.

Lemma 3.14. *Given a sequence of independent, identically distributed real random variables $(X_n)_{n\in\mathbb{N}}$ considered as one random variable taking values in $(\mathbb{R}, \mathcal{B}(\mathbb{R}))^{\otimes\mathbb{N}}$. Define a map $\tau : \mathbb{R}^{\mathbb{N}} \to \mathbb{R}^{\mathbb{N}}$ by $\tau((x_n)_{n\in\mathbb{N}}) = (x_{n+1})_{n\in\mathbb{N}}$. Then $\left(\mathbb{R}^{\mathbb{N}}, \mathcal{B}(\mathbb{R})^{\otimes\mathbb{N}}, P_{(X_n)_{n\in\mathbb{N}}}, \tau\right)$ is ergodic.*

Proof. Since the pre-image of a cylinder set under τ is a cylinder set, τ is measurable. By stationarity of $(X_n)_{n\in\mathbb{N}}$, τ is measure preserving. Thus, we are left with showing that invariant sets are trivial. For any invariant set A and any n we clearly have $A = \tau^{-n}(A)$. Therefore, A only depends on the coordinates $(x_{n+1}, x_{n+2}, \ldots)$. Thus, $B := (X_n)^{-1}(A) = \{\, \omega \in \Omega \mid (X_n(\omega))_{n\in\mathbb{N}} \in A \,\}$ is measurable with respect to the σ-algebra $\mathcal{F}_n := \sigma(X_n, X_{n+1}, \ldots)$ for any n. Therefore, $B \in \mathcal{T}((X_n)_{n\in\mathbb{N}})$. By Theorem 2.30, $P(B) \in \{\, 0, 1 \,\}$, which implies that $P_{(X_n)}(A) = P(B) \in \{\, 0, 1 \,\}$. □

Proof of Theorem 3.4. We are given an i.i.d. sequence of integrable real random variables $(X_n)_{n\in\mathbb{N}}$. As in Lemma 3.14, we consider the probability space $(\Omega', \mathcal{F}', P') := (\mathbb{R}^{\mathbb{N}}, \mathcal{B}(\mathbb{R})^{\otimes\mathbb{N}}, P_{(X_n)_{n\in\mathbb{N}}})$. Consider the random variable $p_1 : \Omega' \to \mathbb{R}$ defined by $p_1((x_n)_{n\in\mathbb{N}}) = x_1$. By Lemma 3.14, the shift operator τ is ergodic on $(\Omega', \mathcal{F}', P')$. Therefore, Theorem 3.13 implies that

$$\frac{1}{n}\sum_{k=1}^{n} p_1 \circ \tau^{k-1} \xrightarrow{n\to\infty} \int p_1 dP', \quad P'\text{-almost surely.}$$

But P'-almost sure convergence of a sequence of random variables π_n on Ω' is equivalent to P-almost sure convergence of the related sequence of random variables $\pi_n \circ$

$(X_k)_{k\in\mathbb{N}}$ on Ω – by definition of the image measure. Therefore, we have shown that

$$\frac{1}{n}\sum_{k=1}^{n} p_1 \circ \tau^{k-1} \circ (X_j)_{j\in\mathbb{N}} = \frac{1}{n}\sum_{k=1}^{n} p_1 \circ (X_{k-1+j})_{j\in\mathbb{N}} = \frac{1}{n}\sum_{k=1}^{n} X_k$$

converges P-almost surely to $\int p_1 dP' = E[X_1]$. □

3.4 Applications and examples

3.4.1 An example of ergodicity

Let $\Omega = [0,1[$, $\mathcal{F} = \mathcal{B}([0,1[)$ and $P = \lambda|_{[0,1[}$, the Lebesgue measure restricted to the interval $[0,1[$. For $r \in]0,1[$ we set $\tau_r(x) \coloneqq x + r \mod 1$.

Theorem 3.15. *The system $(\Omega, \mathcal{F}, P, \tau_r)$ is ergodic if and only if r is irrational.*

Before we give a proof of this statement, recall some fact about Fourier series.

Lemma 3.16. *Given a square integrable measurable function $f : [0,1[\to \mathbb{R}$. Then there is a unique representation*

$$f(x) = \sum_{n\in\mathbb{Z}} c_n e^{2\pi i n x}$$

in the sense that the above series converges in $L^2([0,1[, dx)$ and the sequence of coefficients $(c_n)_{n\in\mathbb{Z}}$ is unique.

Proof of Theorem 3.15. Take any $\mathcal{I}$-measurable square integrable function $f : [0,1[\to \mathbb{R}$ and consider its Fourier series expansion

$$f(x) = \sum_{n\in\mathbb{Z}} c_n e^{2\pi i n x}. \tag{3.12}$$

Then $f = f \circ \tau_r$, and if we insert this equation into the Fourier series (3.12), we obtain

$$\sum_{n\in\mathbb{Z}} c_n e^{2\pi i n x} = \sum_{n\in\mathbb{Z}} c_n e^{2\pi i n r} e^{2\pi i n x}.$$

By comparison of coefficients, we have $\forall n \in \mathbb{Z} : \; c_n = c_n e^{2\pi i n r}$. If r is irrational, then this relation can only be satisfied if $c_n = 0$ for every $n \neq 0$. Thus, $f(x) = c_0$ almost surely, implying that $\mathcal{I}$ is P-trivial, and the system is, hence, ergodic.

On the other hand, if r is rational, there is some $n \in \mathbb{Z} \setminus \{0\}$ with $rn \in \mathbb{Z}$, implying that $e^{2\pi i n r} = e^{-2\pi i n r} = 1$. Therefore, we can freely choose c_n, c_{-n} (and c_{2n}, c_{-2n}, …) and obtain some non-trivial $\mathcal{I}$-measurable function f. □

What does the ergodic theorem imply for this example? Take any integrable, measurable function $f : [0,1[\to \mathbb{R}$ – for simplicity we extend the function to a periodic function $\mathbb{R} \to \mathbb{R}$. If r is irrational, for almost every starting point $x \in [0,1[$, we have the convergence

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} f(x + (k-1)r) = \int_0^1 f(y)dy. \tag{3.13}$$

3.4.2 Monte Carlo simulation

The second application we have in mind is the *Monte Carlo method*. The Monte Carlo method belongs to the most important numerical methods. It was developed by giants of mathematics and physics like J. von Neumann, E. Teller and S. Ulam and N. Metropolis during the development of the H-bomb. Today, it is widely used in fields like statistical mechanics, particle physics, computational chemistry, molecular dynamics, computational biology and computational finance.

Assume that we want to compute the quantity

$$I[f;X] \coloneqq E\left[f(X)\right], \tag{3.14}$$

assuming only that we can actually sample from the distribution of X and that $E\left[|f(X)|\right] < \infty$. Taking a sequence $X_1, X_2, \dots$ of independent realizations of X, the strong law of large numbers implies that

$$I[f;X] = \lim_{M\to\infty} \frac{1}{M}\sum_{i=1}^{M} f(X_i), \quad P-\text{a.s.} \tag{3.15}$$

However, in numerics we are usually not quite satisfied with a mere convergence statement like in (3.15). Indeed, we would like to be able to control the error, i.e., we would like to have an error estimate or bound and we would like to know how fast the error goes to 0 if we increase M. Before continuing the discussion, let us formally introduce the Monte Carlo integration error by

$$\epsilon_M = \epsilon_M(f;X) \coloneqq I[f;X] - I_M[f;X], \quad \text{where } I_M[f;X] \coloneqq \frac{1}{M}\sum_{i=1}^{M} f(X_i) \tag{3.16}$$

is the estimate based on the first M samples. Note that $I_M[f;X]$ is an *unbiased* estimate for $I[f;X]$ in the statistical sense, i.e., $E\left[I_M[f;X]\right] = I[f;X]$, implying $E\left[\epsilon_M(f)\right] = 0$. We also introduce the *mean square error* $E\left[\epsilon_M(f;X)^2\right]$ and its square root, the error in L^2. The *central limit theorem* immediately implies both error bounds and convergence rate provided that $f(X)$ is square integrable.

Theorem 3.17. *Let $\sigma = \sigma(f;X) < \infty$ denote the standard deviation of the random variable $f(X)$. Then the root mean square error satisfies*

$$E\left[\epsilon_M(f;X)^2\right]^{1/2} = \frac{\sigma}{\sqrt{M}}.$$

Moreover, $\sqrt{M}\epsilon_M(f;X)$ is asymptotically normal (with standard deviation $\sigma(f;X)$). i.e., for any constants $a < b \in \mathbb{R}$ we have

$$\lim_{M\to\infty} P\left(\frac{\sigma a}{\sqrt{M}} < \epsilon_M < \frac{\sigma b}{\sqrt{M}}\right) = \Phi(b) - \Phi(a),$$

where Φ denotes the distribution function of a standard normal random variable.

Proof. Using independence of the X_i and the fact that $I_M[f;X]$ is unbiased,

$$E\left[\epsilon_M^2\right] = \operatorname{var}\left[\frac{1}{M}\sum_{i=1}^{M} f(X_i)\right] = \frac{1}{M^2}\sum_{i=1}^{M} \operatorname{var}[f(X_i)] = \frac{M\operatorname{var}[f(X_1)]}{M^2} = \frac{\sigma^2}{M}.$$

Asymptotic normality is an immediate consequence of the central limit theorem. □

Theorem 3.17 has two important implications.

1. The error is probabilistic: there is no deterministic error bound. For a particular simulation, and a given sample size M, the error of the simulation can be as large as you want. However, large errors only occur with probabilities decreasing in M.

2. The "typical" error (e.g., the root mean square error $\sqrt{E\left[\epsilon_M^2\right]}$) decreases to zero like $1/\sqrt{M}$. In other words, if we want to increase the accuracy of the result tenfold (i.e., if we want to obtain one more significant digit), then we have to increase the sample size M by a factor $10^2 = 100$. We say that the Monte Carlo method *converges with rate* $1/2$.

Before continuing the discussion of the convergence rate, let us explain how to control the error of the Monte Carlo method taking its random nature into account. The question here is, how do we have to choose M (the only parameter available) such that the probability of an error larger than a given tolerance level $\varepsilon > 0$ is smaller than a given $\delta > 0$, symbolically

$$P\left(|\epsilon_M(f;X)| > \varepsilon\right) < \delta.$$

Fortunately, this question is already almost answered in Theorem 3.17. Indeed, it implies that

$$P\left(|\epsilon_M| > \varepsilon\right) = 1 - P\left(-\frac{\sigma\tilde{\varepsilon}}{\sqrt{M}} < \epsilon_M < \frac{\sigma\tilde{\varepsilon}}{\sqrt{M}}\right) \sim 1 - \Phi(\tilde{\varepsilon}) + \Phi(-\tilde{\varepsilon}) = 2 - 2\Phi(\tilde{\varepsilon}),$$

where $\tilde{\varepsilon} = \sqrt{M}\varepsilon/\sigma$. Of course, the normalized Monte Carlo error is only asymptotically normal, which means the equality between the left and the right hand side of the above equation only holds for $M \to \infty$, which is signified by the "$\sim$"-symbol. Equating the right hand side with δ and solving for M yields

$$M = \left(\Phi^{-1}\left(\frac{2-\delta}{2}\right)\right)^2 \sigma^2\varepsilon^{-2}. \tag{3.17}$$

Thus, as we have already observed before, the number of samples depends on the tolerance like $1/\varepsilon^2$.

Remark. This analysis tacitly assumed that we know $\sigma = \sigma(f;X)$. Since we started the whole endeavor in order to compute $I[f;X]$, it is, however, very unlikely that we already know the variance of $f(X)$. Therefore, in practice we will have to replace $\sigma(f;X)$ by a sample estimate. (This is not unproblematic: what about the Monte Carlo error for the approximation of $\sigma(f;X)$?)

Remark 3.18. Let us come back to the merits of Monte Carlo simulation. For simplicity, let us assume that X is a d-dimensional uniform random variable, i.e.,

$$I[f] \coloneqq I[f;U] = \int_{[0,1]^d} f(x)dx.$$

Note that the dimension of the space did not enter into our discussion of the convergence rate and of error bounds at all. This is remarkable if we compare the Monte Carlo method to traditional methods for numerical integration. Those methods are

usually based on a grid $0 \leq x_1 < x_2 < \cdots < x_N \leq 1$ of arbitrary length N. The corresponding d-dimensional grid is simply given by $\{ x_1, \ldots, x_N \}^d$, a set of size N^d. The function f is evaluated on the grid points and an approximation of the integral is computed based on interpolation of the function between grid-points by suitable functions (e.g., piecewise polynomials), whose integral can be explicitly computed. Given a numerical integration method of order k, the error is the proportional to $\left(\frac{1}{N}\right)^k$. However, the we had to evaluate the function on N^d points. Therefore, the accuracy in terms of points merely is like $n^{-k/d}$, where n denotes the total number of points involved, which is proportional to the computational cost. This is known as the *curse of dimensionality*: even methods, which are very well suited in low dimensions, deteriorate very fast in higher dimensions.

The curse of dimensionality is the main reason for the popularity of the Monte Carlo method. As we will see later, in financial applications the dimension of the state space can easily be in the order of 100 (or much higher), which already makes traditional numerical integration methods completely unfeasible. In other applications, like molecular dynamics, the dimension of the state space might be in the magnitude of 10^{12}!

Chapter 4

Conditional expectations

Let $(\Omega, \mathcal{F}, P)$ be a discrete probability space, and assume that Ω is finite and each elementary event $\{\omega\}$ has positive probability. Recall that for an event $A \neq \emptyset$ the conditional probability of some other event B given A is defined by

$$P(B|A) \coloneqq \frac{P(A \cap B)}{P(A)}. \tag{4.1}$$

Note that $B \mapsto P(B|A)$ defines a probability measure on Ω, and we can therefore define the *conditional expectation* of some random variable $Y : \Omega \to \mathbb{R}$ given the event A by the expectation of Y under $P(\cdot|A)$, i.e.,

$$E[Y|A] \coloneqq \sum_{\omega \in \Omega} Y(\omega) P(\{\omega\}|A).$$

Example 4.1. Let $\Omega = \{1, \ldots, 6\}$ with the uniform probability measure, let $A = \{2, 4, 6\}$ and $Y(\omega) = \omega$. The interpretation is that Y is the outcome of throwing a die and A is the event that the die shows an even number of eyes. Then, the conditional expectation of Y given A is (mathematically and intuitively) $E[Y|A] = 4$.

As usual, probability theory is mostly concerned with properties that only depend on the involved distributions, not on the probability space. Thus, the notion of the conditional expectation of one random variable Y with respect to another random variable X is even more interesting. We define

$$E[Y|X](\omega) \coloneqq E[Y|A_\omega], \text{ where } A_\omega \coloneqq \{\, \omega' \in \Omega \;\big|\; X(\omega') = X(\omega) \,\}. \tag{4.2}$$

Again, the definition is best illustrated by an example.

Example 4.2. In continuation of Example 4.1 above, set

$$X(\omega) \coloneqq \mathbf{1}_A(\omega) = \begin{cases} 1, & \omega \in A, \\ 0, & \omega \notin A. \end{cases}$$

Then we obtain $E[Y|X](\omega) = 3\mathbf{1}_{A^c}(\omega) + 4\mathbf{1}_A(\omega)$.

In other words, we see that (4.2) is really an adequate generalization of the conditional expectation with respect to sets. We also see that it is natural to define $E[Y|X]$ as a random variable, whose value depends on the value of X, i.e., which is $\sigma(X)$-measurable. On the other hand, we also see that $E[Y|X]$ depends only on the subsets of Ω, on which X is constant, but not on the values of X itself – $E[Y|X] = E[Y|2X]$. This indicates that $E[Y|X]$ should actually only depend on X via $\sigma(X)$.

Remark. The same constructions also make sense in the case of a general probability space, when X only takes finitely many values. By density, this could be used to extend the definitions to general random variables.

4.1 Conditional expectations

As usual, we assume that we are given a probability space $(\Omega, \mathcal{F}, P)$. *Moreover, in this section we will often omit the qualification "almost surely" for (in)equalities of random variables.* $\mathcal{A}$ will generally denote a sub-σ-algebra of $\mathcal{F}$.

Definition 4.3. Given a sub-σ-algebra $\mathcal{A} \subset \mathcal{F}$ and an integrable real random variable X. A random variable Y is called *conditional expectation* of X given $\mathcal{A}$ if and only if

(i) Y is $\mathcal{A}$-measurable and

(ii) for every $A \in \mathcal{A}$ we have $E[Y\mathbf{1}_A] = E[X\mathbf{1}_A]$.

We write $Y = E[X|\mathcal{A}]$. The *conditional probability* of $A \in \mathcal{F}$ is defined by by $P(A|\mathcal{A}) \coloneqq E[\mathbf{1}_A|\mathcal{A}]$.
Moreover, for a fixed random variable Z, we define $E[X|Z] \coloneqq E[X|\sigma(Z)]$, and, likewise, $P(A|Z) \coloneqq P(A|\sigma(Z))$.

Lemma 4.4. *The conditional expectation $E[X|\mathcal{A}]$ exists and is unique up to almost sure equality.*

Proof. It suffices to proof existence of $E[X|\mathcal{A}]$ for non-negative integrable random variables X. Indeed, if $E[X^+|\mathcal{A}]$ and $E[X^-|\mathcal{A}]$ both exist, then a version of $E[X|\mathcal{A}]$ is certainly given by $E[X^+|\mathcal{A}] - E[X^-|\mathcal{A}]$.

So assume that $X \geq 0$ and define a finite measure μ on $(\Omega, \mathcal{A})$ by setting

$$\mu(C) \coloneqq E[X\mathbf{1}_C], \quad C \in \mathcal{A}.$$

Clearly, $\mu \ll P|_{\mathcal{A}}$. Therefore, the Radon-Nikodym theorem implies that there is an $\mathcal{A}$-measurable density for μ with respect to $P|_{\mathcal{A}}$ denoted by Y, i.e.,

$$\mu(C) = E[X\mathbf{1}_C] = \int_C Y dP|_{\mathcal{A}} = E[Y\mathbf{1}_C], \quad C \in \mathcal{A}.$$

Thus, $Y = E[X|\mathcal{A}]$. Moreover, uniqueness follows by the uniqueness in the Radon-Nikodym theorem. □

By the construction of the integral, the following characterization holds: let $X \in L^p(\Omega, \mathcal{F}, P)$, then $Y = E[X|\mathcal{A}]$ if and only if

$$\forall Z \in L^q(\Omega, \mathcal{A}, P|_{\mathcal{A}}) : \; E[ZY] = E[ZX], \text{ with } 1/p + 1/q = 1, \; p \in [1, \infty]. \tag{4.3}$$

Note that the ordinary expectation is a special case of the conditional expectation: $E[X] = E[X|\{\emptyset, \Omega\}]$. We collect some simple properties of the conditional expectation.

Lemma 4.5. *Let $X, Y \in L^1(\Omega, \mathcal{F}, P)$ and let $\mathcal{H} \subset \mathcal{G} \subset \mathcal{F}$ be σ-algebras.*

*(i) For $\lambda, \mu \in \mathbb{R}$ we have $E[\lambda X + \mu Y|\mathcal{G}] = \lambda E[X|\mathcal{G}] + \mu E[Y|\mathcal{G}]$ a.s. (*linearity*).*

(ii) *If $X \geq Y$ a.s., then $E[X|\mathcal{G}] \geq E[Y|\mathcal{G}]$ a.s. (*monotonicity*).*[1]

(iii) *If X is $\mathcal{G}$-measurable, then $E[X|\mathcal{G}] = X$ a.s. Moreover, if $E[|XY|] < \infty$, then $E[XY|\mathcal{G}] = XE[Y|\mathcal{G}]$ a.s.*

(iv) *$E[E[X|\mathcal{G}]|\mathcal{H}] = E[E[X|\mathcal{H}]|\mathcal{G}] = E[X|\mathcal{H}]$ a.s. (*tower property*).*

(v) *$|E[X|\mathcal{G}]| \leq E[|X|\,|\mathcal{G}]$ a.s. (*triangle inequality*).*

(vi) *If X is independent of $\mathcal{G}$, then $E[X|\mathcal{G}] = E[X]$ a.s.*

(vii) *Assume that $X_n \to X$ in L^1. Then $\lim_{n\to\infty} E[X_n|\mathcal{G}] = E[X|\mathcal{G}]$ in L^1.*

Proof. (i): The right hand side is $\mathcal{G}$-measurable and for every $C \in \mathcal{G}$ we have

$$E\left[\mathbf{1}_C(\lambda E[X|\mathcal{G}] + \mu E[Y|\mathcal{G}])\right] = \lambda E\left[\mathbf{1}_C E[X|\mathcal{G}]\right] + \mu E\left[\mathbf{1}_C E[Y|\mathcal{G}]\right]$$
$$= \lambda E[\mathbf{1}_C X] + \mu E[\mathbf{1}_C Y] = E[\mathbf{1}_C(\lambda X + \mu Y)].$$

(ii): Let $C = \{\, E[X \mid \mathcal{G}] < E[Y|\mathcal{G}] \,\} \in \mathcal{G}$. Then

$$0 \geq E\left[\mathbf{1}_C(E[X|\mathcal{G}] - E[Y|\mathcal{G}]\right] = E[\mathbf{1}_C(X - Y)] \geq 0,$$

implying that $P(C) = 0$.

(iii): The first part of the assertion follows directly from the definition. For the second part, let us define $X_n := \max(\min(X, n), -n)$, so that $|X_n| \leq n$, $X_n \to X$ a.s. Then $|X_n Y| \leq |XY|$ and $E[X_n Y|\mathcal{G}] \to E[XY|\mathcal{G}]$ by assertion (vii), whose proof will not depend on assertion (iii). So we are left with proving the second part of assertion (iii) for bounded, $\mathcal{G}$-measurable random variables X. Note that for every $Z \in L^\infty(\Omega, \mathcal{G}, P|_{\mathcal{G}})$, the product $ZX \in L^\infty(\Omega, \mathcal{G}, P|_{\mathcal{G}})$, and, by property (4.3) for $E[Y|\mathcal{G}]$,

$$E\left[ZXE[Y|\mathcal{G}]\right] = E[ZXY].$$

Therefore, $XE[Y|\mathcal{G}] = E[XY|\mathcal{G}]$ again by the characterization (4.3).

(iv): The second equality follows from the first part of (iii), and we only need to show that $E[E[X|\mathcal{G}]|\mathcal{H}] = E[X|\mathcal{H}]$. For $C \in \mathcal{H} \subset \mathcal{G}$ we have

$$E[\mathbf{1}_C E[E[X|\mathcal{G}]|\mathcal{H}]] = E[\mathbf{1}_C E[X|\mathcal{G}]] = E[\mathbf{1}_C X].$$

(v): By monotonicity, we have $E[X|\mathcal{G}] \leq E[X^+|\mathcal{G}]$. Since the latter is non-negative, we obtain $(E[X|\mathcal{G}])^+ \leq E[X^+|\mathcal{G}]$ and, similarly, $(E[X|\mathcal{G}])^- \leq E[X^-|\mathcal{G}]$. Thus, by linearity,

$$|E[X|\mathcal{G}]| = (E[X|\mathcal{G}])^+ + (E[X|\mathcal{G}])^- \leq E[X^+|\mathcal{G}] + E[X^-|\mathcal{G}] = E[|X|\,|\mathcal{G}].$$

(vi): For any bounded, $\mathcal{G}$-measurable random variable Z we have

$$E[ZX] = E[Z]E[X] = E[ZE[X]],$$

implying $E[X|\mathcal{G}] = E[X]$ by (4.3).

(vii): By the triangle inequality and linearity, we have

$$E[|E[X|\mathcal{G}] - E[X_n|\mathcal{G}]|] \leq E[E[|X - X_n|\,|\mathcal{G}]] = E[|X - X_n|] \to 0,$$

which shows convergence in L^1. □

[1] In particular, (ii) gives a direct proof of a.s. uniqueness of the conditional expectation.

Property (vii) says that $E[\cdot|\mathcal{A}] : L^1(\Omega, \mathcal{F}, P) \to L^1(\Omega, \mathcal{A}, P|_{\mathcal{A}})$ is continuous. Moreover, we will see shortly that the same property holds true for any $p \geq 1$, i.e., for the map $E[\cdot|\mathcal{A}] : L^p(\Omega, \mathcal{F}, P) \to L^p(\Omega, \mathcal{A}, P|_{\mathcal{A}})$.

Lemma 4.6. *A convex function $\varphi : \mathbb{R} \to \mathbb{R}$ is everywhere right-differentiable and the right-derivative $D^+\varphi$ is increasing and for all $x \in \mathbb{R}$*

$$\varphi(x) = \sup_{y \in \mathbb{R}} \left(\varphi(y) + D^+\varphi(x)(y - x)\right) = \sup_{y \in \mathbb{Q}} \left(\varphi(y) + D^+\varphi(x)(y - x)\right).$$

Proof. By convexity, φ is continuous and for fixed $x \in \mathbb{R}$ the map

$$y \mapsto \frac{\varphi(x) - \varphi(y)}{x - y} =: S(x, y)$$

is increasing on $]x, \infty[$.[2] Thus,

$$D^+\varphi(x) = \lim_{y \searrow x} \frac{\varphi(x) - \varphi(y)}{x - y}$$

exists. Moreover, the inequalities $S(x, t) \leq S(x, y) \leq S(y, u)$ for $x < t < y < u$ imply with $t \searrow x$ and $u \searrow y$ that $D^+\varphi(x) \leq D^+\varphi(y)$. Moreover, we obtain

$$\varphi(y) \leq \varphi(x) + \frac{\varphi(y) - \varphi(u)}{y - u}(y - x),$$

implying for $u \searrow y$ that

$$\varphi(x) \geq \varphi(y) + (x - y)D^+\varphi(y)$$

for $y > x$. Similarly, we can obtain the same inequality for $y < x$. Noting that we have equality for $y = x$, we obtain the first formula from the statement of the lemma. Now take a sequence $x_n \to x$ of rational numbers. Since $D^+\varphi$ is increasing, it is bounded in a neighborhood of x. Therefore, continuity of φ implies that

$$\varphi(x) = \lim_{n \to \infty} \left[\varphi(x_n) + D^+\varphi(x_n)(x_n - x)\right]. \qquad \square$$

Theorem 4.7 (Jensen's inequality)**.** *Let $\varphi : \mathbb{R} \to \mathbb{R}$ be convex and let X be an integrable random variable such that $E[\varphi(X)]$ exists. Then*

$$\varphi\left(E\left[X|\mathcal{A}\right]\right) \leq E\left[\varphi(X)|\mathcal{A}\right] \leq \infty.$$

Proof. For any fixed $x \in \mathbb{Q}$ and any $\omega \in \Omega$ we have

$$\varphi(X(\omega)) \geq \varphi(x) + (X(\omega) - x)D^+\varphi(x),$$

implying the almost sure inequality

$$E[\varphi(X)|\mathcal{A}] \geq \varphi(x) + (E[X|\mathcal{A}] - x)D^+\varphi(x). \tag{4.4}$$

Let N_x denote the exceptional set of (4.4) and set $N := \bigcup_{x \in \mathbb{Q}} N_x$. On N^c, (4.4) holds uniformly for all rational x. Choosing $\omega \in N^c$, we obtain

$$E[\varphi(X)|\mathcal{A}](\omega) \geq \sup_{x \in \mathbb{Q}} \left[\varphi(x) + (E[X|\mathcal{A}](\omega) - x)D^+\varphi(x)\right] = \varphi\left(E[X|\mathcal{A}](\omega)\right)$$

by Lemma 4.6, which gives the result since $P(N) = 0$. $\square$

[2] Indeed, for $x < t < y$ we have $S(x, t) \leq S(x, y) \leq S(t, y)$, which is left as an exercise.

In particular, now we know that $X \in L^p(\Omega, \mathcal{F}, P)$ implies that $E[X|\mathcal{A}] \in L^p(\Omega, \mathcal{A}, P)$ for every $p > 1$ and that the map is continuous. This leads to the following, geometrical interpretation of the conditional expectation for square integrable random variables.

Theorem 4.8. *The map $E[\cdot|\mathcal{A}] : L^2(\Omega, \mathcal{F}, P) \to L^2(\Omega, \mathcal{A}, P|_{\mathcal{A}})$ is the orthogonal projection onto the closed sub-Hilbert-space $L^2(\Omega, \mathcal{A}, P|_{\mathcal{A}})$, i.e., for square-integrable random variables X we have*

$$E\left[(X - E[X|\mathcal{A}])^2\right] = \min_{Y \in L^2(\Omega,\mathcal{A},P)} E\left[(X - Y)^2\right].^3$$

Proof. Let $Y \in L^2(\Omega, \mathcal{A}, P)$ and set $Z := E[X|\mathcal{A}]$. Then, by (4.3), we have $E[XY] = E[ZY]$, and, for $Y = Z$, $E[XZ] = E[Z^2]$. Combining the two equalities, we get

$$E\left[(X - Y)^2\right] - E\left[(X - Z)^2\right] = E\left[(Y - Z)^2\right] \geq 0.$$

Thus, $Y = Z = E[X|\mathcal{A}]$ is a minimizer for $E\left[(X - Y)^2\right]$ and for every other minimizer Y we have $E\left[(Y - Z)^2\right] = 0$, implying that the minimizer is unique in L^2. □

4.2 Markov kernels and regular conditional distributions

Intuitively, we might understand the conditional distribution of some real random variable X given a σ-algebra $\mathcal{A}$ as the "random measure"

$$\mu_{\mathcal{A}}(C)(\omega) = P(X \in C|\mathcal{A})(\omega) = E[\mathbf{1}_C(X)|\mathcal{A}](\omega), \quad C \in \mathcal{B}(\mathbb{R}).$$

However, it is not so clear, what we understand by a random measure. For instance, by various assertions in Lemma 4.5, we have the following (in)equalities almost surely, for $C \in \mathcal{B}(\mathbb{R})$ and disjoint sets $C_n \in \mathcal{B}(\mathbb{R})$:

$$\begin{aligned} \mu_{\mathcal{A}}(\mathbb{R}) &= 1, \\ 0 \leq \mu_{\mathcal{A}}(C) &\leq 1, \\ \mu_{\mathcal{A}}\left(\bigcup_{n=1}^{\infty} C_n\right) &= \sum_{n=1}^{\infty} \mu_{\mathcal{A}}(C_n). \end{aligned}$$

However, it is, in general, *not* true that there is a common, fixed null-set N such that the set-function $C \in \mathcal{B}(\mathbb{R}) \mapsto P(X \in C|\mathcal{A})(\omega)$ is a probability measure for every $\omega \in N^c$: usually, the null-sets will depend on the sets C in the above properties. If we can indeed find such a common null-set, then we will speak of a *regular conditional distribution*.

Before coming back to this question, let us first take a closer look to one important probabilistic context of conditional expectations, namely the case of a conditional expectation of one random variable Y given another random variable X, which was defined by $E[Y|\sigma(X)]$. Intuitively, we would like to be able to condition Y on X taking certain values, i.e., we are interested in expressions like $P(Y \in C|X = x)$. If $P(X = x) > 0$, this can already be done in elementary probability theory. However, in many interesting cases, this is not the case, but one can still make sense of $P(Y \in C|X = x)$.

[3] From now, we simply write P instead of $P|_{\mathcal{A}}$.

Lemma 4.9 (Factorization lemma)**.** *Given a set Ω, a measurable space $(E, \mathcal{E})$, and functions $X : \Omega \to E$ and $Y : \Omega \to \mathbb{R}$. Then Y is $\sigma(X)$-$\mathcal{B}(\mathbb{R})$-measurable, if and only if there is an $\mathcal{E}$-$\mathcal{B}(\mathbb{R})$-measurable function $\varphi : E \to \mathbb{R}$ such that $Y = \varphi \circ X$.*

Proof. By the usual arguments, we may, for simplicity, assume that $Y \geq 0$. If a map φ as in the statement exists, then Y clearly is $\sigma(X)$-$\mathcal{B}(\mathbb{R})$-measurable, so we only need to prove the converse direction.

We know that there is a sequence of non-negative simple functions Y_n converging pointwise and monotone to Y and one can then show rather easily that there exists a representation

$$Y = \sum_{n \in \mathbb{N}} \alpha_n \mathbf{1}_{A_n}, \quad A_n \in \sigma(X).^4$$

By definition of $\sigma(X)$, one can find sets $B_n \in \mathcal{E}$ such that $X^{-1}(B_n) = A_n$. Define

$$\varphi := \sum_{n \in \mathbb{N}} \alpha_n \mathbf{1}_{B_n}.$$

Then we have

$$\varphi \circ X = \sum_{n \in \mathbb{N}} \alpha_n \mathbf{1}_{B_n} \circ X = \sum_{n \in \mathbb{N}} \alpha_n \mathbf{1}_{A_n} = Y. \qquad \square$$

Remark. In fact, the lemma remains valid if $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is replaced by some polish space with its Borel-σ-algebra.

Definition 4.10. Given an integrable real random variable Y and a random variable X taking values in some measurable space $(E, \mathcal{E})$. Then for every $x \in E$ we define the *factorized conditional expectation* $E[Y|X = x]$ by

$$E[Y|X = x] := \varphi(x), \text{ where } \varphi(X) = E[Y|X]$$

as guaranteed by the factorization lemma. In particular, we set $P(Y \in A|X = x) = E[\mathbf{1}_A(Y)|X = x]$, $A \in \mathcal{F}$.

We note that for every set $A \in \mathcal{F}$, the conditional probability $P(Y \in A|X = x)$ was only defined for almost all $x \in E$ and the exceptional null-set will depend on A. Therefore, in general, we cannot hope that one can find a universal exceptional set for all sets A with probability 0. It turns out, however, that this is true when the σ-algebra $\mathcal{E}$ is countably-generated, e.g., in the case of the Borel-σ-algebra of a polish space.

Definition 4.11. Given two measurable spaces $(E_1, \mathcal{E}_1)$ and $(E_2, \mathcal{E}_2)$. A map $\kappa : E_1 \times \mathcal{E}_2 \to [0, \infty]$ is called *transition kernel* if

(i) for every $A \in \mathcal{E}_2$, the map $x \mapsto \kappa(x, A)$ is $\mathcal{E}_1$-$\mathcal{B}([0, \infty])$-measurable and

(ii) for every $x \in E_1$, the map $A \mapsto \kappa(x, A)$ is a σ-finite measure on $\mathcal{E}_2$.

If for every $x \in E_1$ we have $\kappa(x, E_2) = 1$, then we call κ a *stochastic* or *Markov kernel*.

[4] Choose Y_n such that it takes values in $2^{-n}\{0, \ldots, n2^n\}$ only. Then the difference $Y_n - Y_{n-1}$ can only take the values $2^{-n}\{0, \ldots, n2^n\}$, too. Therefore, it can be written as a linear combination of $n2^n$ indicator functions. Now write $Y = Y_0 + \sum_{n=1}^{\infty}(Y_n - Y_{n-1})$, which is a countable linear combination of indicator functions.

Example 4.12. Given a real random variable X and a measurable map $f : \mathbb{R}^2 \to \mathbb{R}$. Consider $\kappa(x, A) := P(f(x, X) \in A)$ for $x \in \mathbb{R}$ and $A \in \mathcal{B}(\mathbb{R})$. Then for every fixed x, $\kappa(x, \cdot)$ clearly is a probability measure. On the other hand, for fixed $A \in \mathcal{B}(\mathbb{R})$, $(x, y) \mapsto \mathbf{1}_A \circ f(x, y)$ is $\mathcal{B}(\mathbb{R})$-$\mathcal{B}(\mathbb{R}^2)$-measurable, and Fubini's theorem implies measurability of

$$x \mapsto P(f(x, X) \in A) = \int_{\mathbb{R}} \mathbf{1}_A \circ f(x, y) P_X(dy).$$

Thus, κ is a Markov kernel. Note that this is the usual construction of *Markov chains* (or random dynamical systems): Given a state $X_n = x$, the next state X_{n+1} is constructed by $X_{n+1} = f(X_n, Y)$ for some independent random variable Y.

Definition 4.13. Given a random variable Y with values in $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ and a random variable X taking values in some measurable space $(E, \mathcal{E})$, both defined on the probability space $(\Omega, \mathcal{F}, P)$. A Markov kernel $\kappa : E \times \mathcal{B}(\mathbb{R}) \to [0, 1]$ is called *regular version of the conditional distribution of Y given X* if $\kappa(x, A) = P(Y \in A | X = x)$ for every $A \in \mathcal{B}(\mathbb{R})$ and $x \in E$.

More precisely, inserting in the Definition 4.10, this means that for every Borel-set A, we have $\kappa(X, A) = E[\mathbf{1}_A(Y)|X]$ almost surely. By the definition of a Markov kernel, this implies that $A \mapsto \kappa(X(\omega), A)$ is, indeed, a probability measure for every ω.

Remark. We have only formulated the definition above for factorized conditional distributions, not for general conditional distributions $P(Y \in A|\mathcal{G})$ for some $\mathcal{G} \subset \mathcal{F}$. In that case, a regular version of the conditional probability is a map $\kappa : \Omega \times \mathcal{B}(\mathbb{R}) \to \mathbb{R}$ such that

- for $A \in \mathcal{B}(\mathbb{R})$ fixed, the random variable $\kappa(A) = P(Y \in A|\mathcal{G})$ a.s.,
- for any $\omega \in \Omega$, $A \mapsto \kappa(A)(\omega)$ is a probability measure on $\mathcal{B}(\mathbb{R})$.

Of course, we could also replace the real valued random variable Y by a random variable taking values in a polish space (with its Borel-σ-algebra). Without proof, we state the fact that in the case of real random variables, regular versions of the conditional expectations always exist.

Lemma 4.14. *Given a real random variable Y and a random variable X as in Definition 4.13. Then a regular version of the conditional expectation of Y given X exists.*

Remark 4.15. The statement remains true when Y is a random variable taking values in some polish space (equipped with the Borel-σ-algebra). In fact, it holds for *Borel spaces*, i.e., measurable spaces $(E, \mathcal{E})$ which are *isomorphic* to $(A, \mathcal{B}(A))$ for some $A \in \mathcal{B}(\mathbb{R})$ in the sense that there is a bijective map $\psi : E \to A$ such that ψ is $\mathcal{E}$-$\mathcal{B}(A)$-measurable and ψ^{-1} is $\mathcal{B}(A)$-$\mathcal{E}$-measurable. It turns out that any polish space equipped with its Borel-σ-algebra is a Borel space.

The usability of the concept of regular conditional distributions, comes from the fact that many intuitive expressions hold provided that we are given a regular version of the conditional distribution, but do not hold in general. As examples of such formulas, we present the following lemma.

Lemma 4.16. *Given a real random variable Y and a random variable X with values in some measurable space $(E, \mathcal{E})$, both defined on $(\Omega, \mathcal{F}, P)$, and assume that $P(\cdot|X = \cdot)$ is a regular version of the conditional distribution of Y given X. Then the following formulas hold:*

(i) for any Borel-function $f : \mathbb{R} \to \mathbb{R}$ such that $E[|f(Y)|] < \infty$ and any $x \in E$ we have

$$E[f(Y)|X = x] = \int_{\mathbb{R}} f(y)P(dy|X = x);$$

(ii) more generally, for any $g : (E, \mathcal{E})\otimes(\mathbb{R}, \mathcal{B}(\mathbb{R})) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ such that $E[|g(X, Y)|] < \infty$ and any $x \in E$ we have

$$E[g(X, Y)|X = x] = \int_{\mathbb{R}} g(x, y)P(dy|X = x).$$

Proof. We omit a formal proof. The idea for both statements is very straightforward. In the first case, first assume that $f(Y) = \mathbf{1}_A(Y)$ for some measurable set. Then the equality holds by definition. The general case is then obtained by approximating f_+ and f_- by elementary functions. In the second case, we start by assuming $g(x, y) = \mathbf{1}_A(x)\mathbf{1}_B(y)$. Once again, the equality follows easily in this case. More generally, we approximate g_+ and g_- by linear combinations of products of indicator functions. □

After all the definitions, we would like to comment on how conditional expectations can actually be computed. In the discrete case, this is quite obvious. Indeed, assume that X only takes the values $(x_n)_{n\in\mathbb{N}}$, $P(X = x_n) > 0$ for every n. Then we can define $E[Y|X = x]$ as an elementary conditional expectation.

Lemma 4.17. *Assume that X and Y are real random variables with a joint density $f_{X,Y}$ such that Y is integrable and the marginal density of X is positive, i.e.,*

$$f_X(x) \coloneqq \int_{\mathbb{R}} f_{X,Y}(x, y)dy > 0, \quad \forall x \in \mathbb{R}.$$

Then we can define the conditional density *of Y given X by*

$$f_{Y|X}(y|x) \coloneqq \frac{1}{f_X(x)} f_{X,Y}(x, y),$$

and we obtain that

$$E[Y|X = x] = \int_{\mathbb{R}} y f_{Y|X}(y|x)dy. \tag{4.5}$$

Proof. Denoting the right hand side of (4.5) by $g(x)$ and rolling up all the definitions in this chapter, we see that we have to prove that

$$\int_C g dP_X = \int_{X^{-1}(C)} Y dP \tag{4.6}$$

for every Borel-set C. We start with the right hand side and see that

$$\int_{X^{-1}(C)} Y dP = \int_\Omega Y\mathbf{1}_C(X)dP = \int_\Omega Y\mathbf{1}_{C\times\mathbb{R}}(X, Y)dP = \int_{C\times\mathbb{R}} y f_{X,Y}(x, y)dxdy. \tag{4.7}$$

Note that

$$E[|Y|] = \int_{\mathbb{R}^2} |y|\, f_{X,Y}(x, y)dxdy < \infty,$$

and Fubini's theorem implies that $x \mapsto \int y f_{X,Y}(x,y)dy$ is λ-almost everywhere defined and integrable, and, by a similar argument, $f_X(x)$ is a.e. defined and finite. Thus,

$$g(x) = \frac{1}{f_X(x)} \int_{\mathbb{R}} y f_{X,Y}(x,y)dy$$

is well-defined, except for a Lebesgue-null-set N, on which either $f_X = \infty$ or $x \mapsto \int y f_{X,Y}(x,y)dy$ is not integrable. Since $P_X \ll \lambda$, $P_X(N) = 0$ as well. Moreover, g is P_X-integrable, since

$$\int |g| \, dP_X = \int |g| \, f_X dx = \int \left| \int y f_{X,Y}(x,y)dy \right| dx \leq \int_{\mathbb{R}^2} |y| \, f_{X,Y}(x,y)dydx < \infty.$$

Therefore, we get for the left hand side of (4.6)

$$\int_C g dP_X = \int_C g(x) f_X(x)dx = \int_{C\times\mathbb{R}} y f_{X,Y}(x,y)dxdy,$$

which, together with (4.7), shows (4.6) and we may conclude that $g(x) = E[Y|X = x]$. □

4.3 Martingales

In this section we take a first look at one particular class of *stochastic processes*. A stochastic process is nothing but a family $(X_i)_{i\in I}$ of random variables on some probability space $(\Omega, \mathcal{F}, P)$. Since, at least intuitively, the index set I plays the role of time, we usually take $I = \mathbb{N}$ or $I = [0, \infty[$. In the following, we shall only consider the former case.

If the random variables X_n are not independent, then one can accumulate information by observing the random variables. At some time n, we have already observed $X_1, \ldots, X_n$, and this enables us to make a better guess about X_{n+1}, i.e.,

$$E[X_{n+1}|X_1, \ldots, X_n] \coloneqq E[X_{n+1}|\sigma(X_1, \ldots, X_n)] \neq E[X_{n+1}].$$

On the other hand, it is also conceivable that we have other sources of information about X_{n+1}, not just the observations of previous instances. For example, X_n might be the closing price of a particular stock at day n. Certainly, the prices of consecutive days are not independent. On the other hand, there are also external sources of information, for instance general economic quantities (like unemployment statistics) or particular information relevant to the economic sector (news of instability in the Arabic world influencing prices of oil-related stocks) or the individual company (like criminal charges against its CEO). Therefore, it makes sense to consider a more general model for the flow of information.

Definition 4.18. Given a probability space $(\Omega, \mathcal{F}, P)$, a *filtration* is an increasing family $(\mathcal{F}_n)_{n\in\mathbb{N}}$ of sub-σ-algebras of $\mathcal{F}$, i.e., for $n < m$ we have $\mathcal{F}_n \subset \mathcal{F}_m$. $(\Omega, \mathcal{F}, (\mathcal{F}_n)_{n\in\mathbb{N}}, P)$ is called *filtered probability space*.

A stochastic process $(X_n)_{n\in\mathbb{N}}$ is *adapted* to a filtration $(\mathcal{F}_n)_{n\in\mathbb{N}}$ if for every $n \in \mathbb{N}$ the random variable X_n is $\mathcal{F}_n$-measurable.

For instance, in the example of a stock price process, we could choose $\mathcal{F}_n \coloneqq \sigma(X_1, \ldots, X_n)$ – the *natural filtration* – if we do not have any outside sources of information. Clearly, a process is adapted to its natural filtration. On the other hand, in case of outside information, we can have $\mathcal{F}_n \supset \sigma(X_1, \ldots, X_n)$.

Definition 4.19. Given a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_n), P)$ and an adapted stochastic process $(X_n)_{n\in\mathbb{N}}$ such that for every n the random variable X_n is integrable. The process is called

- a *martingale* iff for every $n < m$ we have

$$E[X_m|\mathcal{F}_n] = X_n \quad P\text{-a.s.}, \tag{4.8}$$

- a *supermartingale* iff for every $n < m$ we have $E[X_m|\mathcal{F}_n] \leq X_n$,
- a *submartingale* iff for every $n < m$ we have $E[X_m|\mathcal{F}_n] \geq X_n$.

The same definitions hold, mutatis mutandis, in case of a stochastic process $(X_n)_{n=1}^N$ indexed by finite time.

A martingale is a model of a *fair game*. Indeed, the condition (4.8) means in particular that the *best guess* for X_{n+1} given all the information available at time n is precisely X_n, cf. Theorem 4.8. Now assume we participate in a game of chance, which is offered infinitely many times and let X_n denote our accumulated profit or loss after n rounds of the game. If the game is fair, then our expectation of the individual profit from the next round of the game should be zero, i.e., $E[X_{n+1} - X_n|\mathcal{F}_n] = 0$, or $E[X_{n+1}|\mathcal{F}_n] = X_n$, and by induction this implies that (X_n) is a martingale, where $\mathcal{F}_n$ is the filtration generated by the game. This is, indeed, the classical motivation for the notion of a martingale, and there are interesting consequences of the theory of martingales to games of chance. Let us now give some more concrete examples.

Example 4.20. Let $(Y_n)_{n\in\mathbb{N}}$ be a sequence of integrable, independent, identically distributed random variables and assume that $E[Y_n] = 0$. Then the random walk $S_n := \sum_{k=1}^n Y_k$ is a martingale with respect to the filtration $\mathcal{F}_n := \sigma(Y_1, \ldots, Y_n)$. However, if we change the filtration, S_n will generally loose the martingale property. E.g., if we take the augmented filtration $\mathcal{G}_n := \sigma(\mathcal{F}_n \cup \sigma(S_N))$, then S_n certainly is $\mathcal{G}_n$-adapted, but generally fails to be a martingale. (Show this as an exercise!)

Example 4.21. Let (X_n) be an i.i.d. sequence of integrable random variables, modeling the outcome of the n'th round of a game. In round n, the player bets $e_{n-1}(X_1, \ldots, X_{n-1})$. Let $S_1 := X_1$ and $S_{n+1} := S_n + e_n(X_1, \ldots, X_n)X_{n+1}$ be the total profit of the player after $n + 1$ rounds. Assume that $e_n \geq 0$ is uniformly bounded (in n and ω). Then S_n is integrable for every n, adapted to the filtration $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$, and a martingale, iff and only iff $E[X_1] = 0$. It is a submartingale iff $E[X_1] \geq 0$ and a supermartingale iff $E[X_1] \leq 0$. This setting can be used as a model for many games of chance.

Example 4.22. Given an integrable random variable X and some filtration $(\mathcal{F}_n)$. Then a martingale is given by $X_n := E[X|\mathcal{F}_n]$, $n \in \mathbb{N}$. Since $\mathcal{F}_n$ is increasing, more and more information about X is revealed over time. Therefore, it is an obvious question to ask whether one can finally get all of it back, i.e., whether $X = \lim_{n\to\infty} X_n$. In general, this cannot be true, since $\lim_{n\to\infty} X_n$, if it exists, is $\mathcal{F}_\infty := \sigma(\bigcup_{n\in\mathbb{N}} \mathcal{F}_n)$-measurable, so cannot be equal to X if X is not $\mathcal{F}_\infty$-measurable. On the other hand, if X is indeed $\mathcal{F}_\infty$-measurable, then we really have $X = \lim_{n\to\infty} X_n$, and convergence holds both a.s. and in L^1. This is, indeed, a rather general situation: for any martingale $(X_n)_{n\in\mathbb{N}}$ satisfying some stronger integrability assumption ($E[\sup_n |X_n|] < \infty$ is enough), there is an $\mathcal{F}_\infty$-measurable, integrable random variable X_∞ "closing" the martingale, i.e., with $X_n \to X_\infty$ in L^1 and a.s., implying that $X_n = E[X_\infty|\mathcal{F}_n]$, $n \in \mathbb{N}$.

Notice that for a martingale X_n the sequence $E[X_n]$ of expected values is constant, whereas it is increasing for a submartingale and decreasing for a supermartingale. A stochastic process (X_n) is a martingale if and only if it is both a sub- and a supermartingale, and a process is a supermartingale if and only if $(-X_n)$ is a submartingale. We end with some rather easy properties of martingales.

Lemma 4.23. *Given a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_n), P)$.*

(i) Let (X_n), (Y_n) be martingales, a, b real numbers, then $aX_n + bX_n$ is a martingale.

(ii) Let (X_n), (Y_n) be sub- or supermartingales, a, b non-negative real numbers, then $aX_n + bY_n$ is a sub- or supermartingale, respectively.

(iii) Let (X_n) and (Y_n) be supermartingales. Then $Z_n \coloneqq \min(X_n, Y_n)$ is a supermartingale.

(iv) Let (X_n) be a martingale and $\phi : \mathbb{R} \to \mathbb{R}$ a convex function such that $\phi(X_n)$ is integrable for any n. Then $(\phi(X_n))_{n\in\mathbb{N}}$ is a submartingale.

(v) Given an adapted integrable sequence of random variables (X_n). If we have $E[X_{n+1}|\mathcal{F}_n] = X_n$ for every $n \in \mathbb{N}$, then (X_n) is already a martingale.

4.4 Optional sampling

Definition 4.24. Given a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_n), P)$. A random variable τ with values in $\mathbb{N}_\infty \coloneqq \mathbb{N} \cup \{\infty\}$ is called *stopping time* iff for every $n \in \mathbb{N}$ we have $\{\tau \le n\} \in \mathcal{F}_n$.

Stopping times are essential in modeling strategies in games of chance or in trading of stocks.

Example 4.25. Given an adapted process $(X_n)_{n\in\mathbb{N}}$ and a Borel-set A. Then

$$\tau_A(\omega) \coloneqq \inf\{ n \in \mathbb{N} \mid X_n(\omega) \in A \},$$

the first hitting time of A, is a stopping time, where we use the convention that $\inf \emptyset = \infty$. Indeed,

$$\{\tau_A \le n\} = \bigcup_{k=1}^{n} \underbrace{\{X_k \in A\}}_{\in \mathcal{F}_k} \in \mathcal{F}_n.$$

We note that for filtrations indexed by $\mathbb{N}$, a random time τ is a stopping time iff

$$\{\tau = n\} \in \mathcal{F}_n, \ \forall n \in \mathbb{N}.$$

Definition 4.26. Given a stopping time τ, we define the σ-algebra

$$\mathcal{F}_\tau \coloneqq \{ A \subset \Omega \mid \forall n \in \mathbb{N}_\infty : \ A \cap \{\tau \le n\} \in \mathcal{F}_n \},$$

where $\mathcal{F}_\infty \coloneqq \sigma\left(\bigcup_{n\in\mathbb{N}} \mathcal{F}_n\right)$.

Lemma 4.27. *Let $\sigma \le \tau$ be stopping times with respect to the same filtration $(\mathcal{F}_n)$. Then $\mathcal{F}_\sigma \subset \mathcal{F}_\tau \subset \mathcal{F}_\infty$.*

Proof. First of all, for $A \in \mathcal{F}_\tau$ we have $A = A \cap \{\tau \le \infty\} \in \mathcal{F}_\infty$. Since $\{\tau \le n\} \subset \{\sigma \le n\}$, for $A \in \mathcal{F}_\sigma$ we have

$$A \cap \{\tau \le n\} = \underbrace{A \cap \{\sigma \le n\}}_{\in \mathcal{F}_n} \cap \underbrace{\{\tau \le n\}}_{\in \mathcal{F}_n} \in \mathcal{F}_n,$$

implying that $A \in \mathcal{F}_\tau$. □

Given an adapted process (X_n) with values in $(E, \mathcal{E})$ and a stopping time τ with $P(\tau < \infty) = 1$, we define

$$X_\tau(\omega) := \begin{cases} X_{\tau(\omega)}(\omega), & \omega \in \{\tau < \infty\}, \\ x, & \omega \in \{\tau < \infty\}^c. \end{cases} \tag{4.9}$$

for some fixed $x \in E$. The justification of the definition of the τ-past is given in

Lemma 4.28. *The random variable X_τ is P-a.s. uniquely determined and $\mathcal{F}_\tau$-measurable.*

Proof. We have to prove measurability. For some $A' \in \mathcal{E}$, we set $A := X_\tau^{-1}(A')$ and have to prove that $A \in \mathcal{F}_\tau$. For $n \in \mathbb{N}$ we have

$$A \cap \{\tau \le n\} = A \cap \bigcup_{k=1}^{n} \{\tau = k\} = \bigcup_{k=1}^{n} \left(\underbrace{X_k^{-1}(A')}_{\in \mathcal{F}_k} \cap \{\tau = k\} \right) \in \mathcal{F}_n.$$

Now consider $n = \infty$. Clearly, $\{\tau \le \infty\} = \{\tau < \infty\} \cup \{\tau = \infty\}$, and $A \cap \{\tau < \infty\} = \bigcup_{n \in \mathbb{N}} A \cap \{\tau = n\} \in \mathcal{F}_\infty$. On the other hand,

$$A \cap \{\tau = \infty\} = \begin{cases} \{\tau = \infty\}, & x \in A', \\ \emptyset, & x \notin A', \end{cases} \in \mathcal{F}_\infty.$$ □

Lemma 4.29. *Let $(X_n)_{n \in \mathbb{N}}$ be a martingale and τ a stopping time with respect to the same filtration $(\mathcal{F}_n)_{n \in \mathbb{N}}$. If τ is bounded by a constant K, then $E[X_\tau] = E[X_1]$.*

Proof. By construction of X_τ, we have $X_\tau(\omega) = \sum_{n=1}^{K} X_n(\omega) \mathbf{1}_{\{n\}}(\tau(\omega))$, implying that X_τ is integrable since $|X_\tau| \le \sum_{n=1}^{K} |X_n|$ and that

$$E[X_\tau] = E\left[\sum_{n=1}^{K} X_n \mathbf{1}_{\{\tau = n\}} \right] = \sum_{n=1}^{K} E[X_n \mathbf{1}_{\{\tau = n\}}].$$

Since $\{\tau = n\} \in \mathcal{F}_n$ and $X_n = E[X_K | \mathcal{F}_n]$ a.s., we further get

$$\begin{aligned} E[X_\tau] = \sum_{n=1}^{K} E[E[X_K | \mathcal{F}_n] \mathbf{1}_{\{\tau = n\}}] &= \sum_{n=1}^{K} E[X_K \mathbf{1}_{\{\tau = n\}}] \\ &= E\left[X_K \sum_{n=1}^{K} \mathbf{1}_{\{\tau = n\}} \right] = E[X_K] = E[X_0]. \end{aligned}$$ □

Theorem 4.30 (Optional sampling theorem)**.** *Let (X_n) be an $(\mathcal{F}_n)$-martingale and $\tau_1 \le \cdots \le \tau_N$ be a finite sequence of bounded stopping times. Then $(X_{\tau_n})_{n=1}^N$ is a martingale with respect to $(\mathcal{F}_{\tau_n})_{n=1}^N$.*

Proof. First we note that it suffices to consider the case of two stopping times $\sigma \leq \tau \leq K$. Note that both stopped random variables are integrable, since $|X_\tau| \leq \sum_{n=1}^{K} |X_n|$ and likewise for X_σ. For any fixed $A \in \mathcal{F}_\sigma$ we need to show that $E[X_\tau \mathbf{1}_A] = E[X_\sigma \mathbf{1}_A]$. Define a random time ρ by

$$\rho(\omega) \coloneqq \sigma(\omega)\mathbf{1}_A(\omega) + \tau(\omega)\mathbf{1}_{A^c}(\omega).$$

Since $A \in \mathcal{F}_\sigma$, we also have $A^c \in \mathcal{F}_\tau$, implying that

$$\{\rho \leq n\} = \underbrace{A \cap \{\sigma \leq n\}}_{\in \mathcal{F}_n} \cup \underbrace{A^c \cap \{\tau \leq n\}}_{\in \mathcal{F}_n} \in \mathcal{F}_n.$$

Thus, ρ is a stopping time. By Lemma 4.29, we have

$$\begin{aligned} E[X_\rho] &= E\left[X_\sigma \mathbf{1}_A + X_\tau \mathbf{1}_{A^c}\right] = E[X_1], \\ E[X_\tau] &= E\left[X_\tau \mathbf{1}_A + X_\tau \mathbf{1}_{A^c}\right] = E[X_1]. \end{aligned}$$

Subtracting both equalities then gives

$$E[X_\sigma \mathbf{1}_A] = E[X_\tau \mathbf{1}_A]. \qquad \square$$

If one starts with a supermartingale instead of a martingale, then the sampled process is again a supermartingale, and likewise for submartingales.

Corollary 4.31 (Optional stopping theorem)**.** *Let (X_n) be an $(\mathcal{F}_n)$-(super-)martingale and τ a stopping time with respect to $(\mathcal{F}_n)$. Define the stopped process $X_n^\tau \coloneqq X_{\min(n,\tau)}$. Then (X_n^τ) is again an $(\mathcal{F}_{\min(n,\tau)})$-(super-)martingale.*

Remark 4.32. Various generalizations of the optional sampling theorem removing the stringent assumption of uniformly bounded stopping times are possible. A quite general version is the following: given a martingale $(X_n)_{n\in\mathbb{N}}$ and an increasing sequence $(\tau_n)_{n\in\mathbb{N}}$ of stopping times such that $E\left[\left|X_{\tau_n}\right|\right] < \infty$ and

$$\liminf_{N\to\infty} \int_{\{\tau_n > N\}} |X_N| \, dP = 0$$

for any $n \in \mathbb{N}$. Then the sequence $(X_{\tau_n})_{n\in\mathbb{N}}$ is a martingale.

Essentially, Theorem 4.30 and Corollary 4.31 say that there are no *feasible* winning strategies in fair games. Indeed, we have before argued that fair games of chance can be modeled as martingales. There are many ways how to model strategies in games, but one way is outlined in Example 4.21, where we allowed to choose the bets as a function of the previous history of outcomes. Another way would be to apply optimal stopping rules. In both cases, we have to require some kind of boundedness condition. Either we require that the total loss accrued must be uniformly bounded, or we require that we have to stop before some deterministic time K. Both requirements are natural from the perspective of true gaming situations, and using Example 4.21 and Corollary 4.31, respectively, we see that no such strategy can lead to a gain on average. However, the following example shows that the boundedness requirement is critical.

Example 4.33. Let us toss a fair coin over an over again (independent of each other), and let X_n be 1 if the outcome of the n'th toss is head, and -1 otherwise. Then $(X_n)_{n\in\mathbb{N}}$ is a martingale. (Which filtration should one use?) Assume that we have unbounded

credit and we start with $Y_0 = 0$ on our gaming account. We bet on the subsequent outcome of the coin tosses, and we get back twice the stake if we guess right an lose the stake otherwise. Let us follow the following strategy: we always bet twice the amount of the previous time, but stop when we win for the first time. We start with an initial stake of 1. Assuming that we always bet on head, this means that

$$Y_n = \sum_{k=1}^{n} 2^{k-1} X_k$$

and $\tau = \inf \{ n \mid X_n = 1 \}$. Note that

$$Y_\tau = \sum_{n=1}^{\infty} \left(2^{n-1} - \sum_{k=1}^{n-1} 2^{k-1} \right) \mathbf{1}_{\{\tau = n\}} = \sum_{n=1}^{\infty} \left(2^{n-1} - (2^{n-1} - 1) \right) \mathbf{1}_{\{\tau = n\}} = 1,$$

implying that this *doubling strategy* indeed almost surely ends up profitable. Note that we even have $P(\tau < \infty) = 1$. However, the stopping time τ is unbounded, meaning we have to potentially wait an infinite time, and the accrued loss in the mean time is unbounded as well, since $Y_{n-1} = 1 - 2^{n-1}$ if $\tau = n$. Thus, we indeed need unbounded credit in order to follow this strategy.

Chapter 5

Stochastic processes

In this chapter, we will discuss several classes of stochastic processes, give examples and approach the delicate problem of construction of stochastic processes, i.e., of probability spaces, on which a certain stochastic process can be defined. We will usually work in the following setting. We start with a probability space $(\Omega, \mathcal{F}, P)$. A stochastic process is then a family $(X_i)_{i\in I}$ of real random variables X_i, $i \in I$. Here, the index set I represents time and we will either have discrete time stochastic processes – with $I = \mathbb{N}$ or $I = \{1, \dots, N\}$ – or continuous time stochastic processes – with $I = \mathbb{R}_+ = [0, \infty[$ or $I = [0, T]$.

5.1 Examples

Already in Chapter 4 we have encountered some important classes of stochastic processes, namely martingales, super- and sub-martingales. While we have only defined those notions in discrete time, the generalization to continuous time is obvious. In this section, we are concerned with more "constructive" possible properties of stochastic processes, i.e., with properties which allow or help with constructing the process in the sense indicated above.

A notion of similar importance as the notion of a martingale is the notion of a *Markov process*. Intuitively, a stochastic process has the Markov property if at any time $i \in I$ the whole information available about the future development of the process is already contained in the current value X_i. This is an analogue to the notion of a dynamical system, where again the future dynamics only depends on the current value, but not on the past, i.e., not on the prior development of the process leading to the current position. More formally, we have

Definition 5.1. Given a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_i)_{i\in I}, P)$ and an adapted stochastic process $(X_i)_{i\in I}$. The process satisfies the *Markov property* (and is then called a *Markov process*) w.r.t. the filtration $(\mathcal{F}_i)_{i\in I}$ iff for ever $i < j \in I$ and any $A \in \mathcal{B}(\mathbb{R})$ we have

$$P\left(X_j \in A|\mathcal{F}_i\right) = P\left(X_j \in A|X_i\right) a.s.$$

We will simply call it a Markov process, if the filtration is generated by the process, i.e., if $\mathcal{F}_i = \sigma\left(X_j : j \in I,\ j \le i\right)$.

As already in the case of a martingale, in discrete time it suffices to check the Markov property for indices $j = i + 1$. When $I = \mathbb{N}$, the distribution of a Markov

process $(X_n)_{n\in\mathbb{N}}$ is determined by the initial distribution, i.e., the distribution of X_1, and the *transition distributions* $P(X_{n+1} \in A|X_n)$, $n \in \mathbb{N}$.

Example 5.2. Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent, identically distributed real random variables, Y a real random variable independent of (X_n) and define a stochastic process $(S_n)_{n\in\mathbb{N}_0}$ by $S_0 := Y$ and $S_n := S_{n-1} + X_n$, $n \in \mathbb{N}$. (S_n is called a *random walk*.) Then (S_n) is a Markov process (with respect to the natural filtration). Indeed, X_n is independent of $\sigma(S_1, \dots, S_{n-1}) = \sigma(Y, X_1, \dots, X_{n-1})$, implying that
(5.1)
$$P(S_n \in A|\mathcal{F}_{n-1}) = E[\mathbf{1}_A(S_{n-1} + X_n)|\mathcal{F}_{n-1}] = E[\mathbf{1}_A(s + X_n)]|_{s=S_{n-1}} = P(S_n \in A|S_{n-1}),$$

where we have used Lemma 4.16 part (ii) together with Lemma 4.5 part (vi).[1]

If $E[X_n] = 0$ for every n, then $(X_n)_{n\in\mathbb{N}}$ is also a martingale

Example 5.3. Assume we are given a Markov process $(X_n)_{n\in\mathbb{N}}$ in discrete time with a finite state space, w.l.o.g., the state space can be chosen to be $\{1, \dots, M\}$ for some $M \in \mathbb{N}$. In this case, (X_n) is called a (discrete time) *Markov chain*. We may represent the transition distributions by matrices $P_n := (p_n^{i,j})_{i,j=1}^M$, $n \in \mathbb{N}$ defined by

$$p_n^{i,j} := P(X_{n+1} = j|X_n = i).$$

P_n has the property that all entries are non-negative and $\forall i : \sum_{j=1}^M p_n^{i,j} = 1$. On the other hand, every such matrix P_n can be interpreted as the transition matrix of a Markov chain. In most cases, one considers time-homogeneous Markov chains, which means that $P_n \equiv P$ for any n. Note that the transition probabilities over more than one step of the Markov chain are obtained via matrix multiplication. Indeed, by the *Chapman-Kolmogorov* equation, we have

$$P(X_{n+2} = j|X_n = i) = \sum_{k=1}^{n} p^{i,k}p^{k,j} = (P \cdot P)^{i,j} = (P^2)^{i,j}.$$

For the actual distribution of X_n, we also need to fix the initial distribution, i.e., the distribution of X_1. We are free to choose whatever distribution, say $\pi^i = P(X_1 = i)$, $i = 1, \dots, M$. Then we have for any n and any i

$$\begin{pmatrix} P(X_{n+1} = 1) \\ \vdots \\ P(X_{n+1} = M) \end{pmatrix} = (P^n)^T \begin{pmatrix} \pi^1 \\ \vdots \\ \pi^M \end{pmatrix}.$$

This means that we, indeed, can construct an underlying probability space for any discrete time Markov chain.

[1] More precisely, we argue as follows for the last equality: Let us assume we are given a regular version of the conditional distribution of $P(X_n \in \cdot|S_{n-1} = s)$. Then from Lemma 4.16, we know that

$$E[\mathbf{1}_A(S_{n-1} + X_n)|S_{n-1} = s] = \int_{\mathbb{R}} \mathbf{1}_A(s + x)P(X_n \in dx|S_{n-1} = s) = E[\mathbf{1}_A(s + X_n)]$$

by independence. This implies that

$$P(S_n \in A|S_{n-1}) = E[\mathbf{1}_A(s + X_n)]|_{s=S_{n-1}}.$$

For the second equality in (5.1), we need a similar formula for the general conditional expectation, allowing to "integrate out independent terms", which can be easily proved by the usual arguments starting with indicator functions.

Exercise 5.4. How can one realize a discrete time Markov chain on the probability space $([0,1], \mathcal{B}([0,1]), \lambda|_{[0,1]})^{\otimes\mathbb{N}}$? I.e., given the initial distribution π and the transition matrix P, explicitly construct the random variables $(X_n)_{n\in\mathbb{N}}$ such that they have the correct distributions and conditional distributions.

Next we are going to give examples in continuous time, which fall into a very important sub-class of continuous time Markov processes.

Definition 5.5. A stochastic process $(X_t)_{t\geq 0}$ is called *Lévy process* iff

(i) $X_0 = 0$ a.s.,

(ii) the process has *independent increments*, i.e., for any $n \in \mathbb{N}$ and $0 \leq t_1 < t_2 < \cdots < t_n$ the random variables $X_{t_2} - X_{t_1}, X_{t_3} - X_{t_2}, \ldots, X_{t_n} - X_{t_{n-1}}$ are independent,

(iii) the process has *stationary increments*, i.e., for any $t \geq 0$ and any $h > 0$ the increment $X_{t+h} - X_t$ has the same distribution as X_h (i.e., the distribution of the increment only depends on h but not on t),

(iv) the paths $t \mapsto X_t(\omega)$ are a.s. right continuous with left limits.

In a sense, the notion of a Lévy process is the appropriate generalization of the notion of a random walk as in Example 5.2 to the continuous time setting, except that we did not require $S_0 = 0$ there – even though this would be a common assumption. In particular, the same argument as in the example shows that any Lévy process is a Markov process. In a way that can be made precise, there are two "extremal" Lévy processes, namely the *Poisson process* and the *Brownian motion*.

Definition 5.6. A Lévy process $(N_t)_{t\geq 0}$ is called *Poisson process* with parameter $\lambda > 0$ iff for every $h > 0$ the random variable N_h has the Poisson distribution with parameter λh, i.e.,

$$P(N_h = k) = \frac{(\lambda h)^k}{k!} e^{-\lambda h}.$$

In particular, this means that N_t is an increasing process only taking integer values. The Poisson process is often used in mathematical models. A representative use of the Poisson process can be found in insurance mathematics.

Example 5.7. One way to model insurance business is to use separate models for the number and times of claims on the one hand, and for the size of claims on the other hand. If claims occur largely independent of each other (i.e., without big clusters), then the arrival of claims can be modeled by a Poisson process with a suitable (possibly time-dependent) intensity λ.[2] The claim sizes are assumed to be i.i.d. random variables which are independent of the claim numbers. More precisely, they are represented by an i.i.d. sequence $(X_n)_{n\in\mathbb{N}}$ of positive random variables independent of the Poisson process $(N_t)_{t\geq 0}$. Then the accumulated claim size of the insurance company up to time t is modeled by

$$S_t := \sum_{n=1}^{N_t} X_n. \tag{5.2}$$

[2] Up to the usual imperfections, this assumption might be reasonable for car insurance, but certainly not reasonable for insurances against earthquakes or storms. Mathematically, the claim is justified by Theorem 5.10 below.

The process $(S_t)_{t\geq 0}$ is known as a *compound Poisson process* and is again a Lévy process. Its distribution is known as compound Poisson distribution. On the other hand, the insurance company also gets premia. Assume that premia are paid continuously with a rate c. Then the overall loss L_t of the insurance company at time t satisfies

$$L_t = S_t - ct.$$

This model for the aggregate losses is known as *Lundberg model*. Note that L_t is still a Lévy process.

Exercise 5.8.

- Verify that the compound Poisson process (5.2) is a Lévy process and compute its expected value and variance, assuming that they exist.
- Compute the characteristic function of the compound Poisson process.
- Show that the loss process $(L_t)_{t\geq 0}$ is a Lévy process.

Definition 5.9. A Lévy process $(B_t)_{t\geq 0}$ is called *Brownian motion* iff for any $h > 0$ we have $B_t \sim \mathcal{N}(0, t)$ and the paths $t \mapsto B_t(\omega)$ are a.s. continuous.

5.2 Poisson process

The reason why the Poisson process is so often applied is the following observation.

Theorem 5.10. *Given a counting process (i.e., an integer-valued increasing process) $(N_t)_{t\geq 0}$ with independent and stationary increments satisfying*

- $P(N_0 = 0) = 1$,
- $P(N_h = 1) = \lambda h + o(h)$ *for* $h \to 0$,
- $P(N_h > 1) = o(h)$ *for* $h \to 0$

for some $\lambda > 0$. If the paths are chosen in such a way that the process is right continuous, then $(N_t)_{t\geq 0}$ is a Poisson process.

Proof. We need to prove that $p_n(t) := P(N_t = n)$ gives the probability function of the Poisson distribution with parameter λt. Let us first consider $n = 0$. By the assumptions, we have

$$p_0(t + h) = P(N_{t+h} - N_t = 0)P(N_t = 0) = (1 - \lambda h + o(h))p_0(t),$$

implying that

$$\dot{p_0}(t) = \lim_{h\to 0} \frac{p_0(t + h) - p_0(t)}{h} = \lim_{h\to\infty} \left(-\lambda + \frac{o(h)}{h}\right) p_0(t) = -\lambda p_0(t).$$

Solving the ODE with the initial value $p_0(0) = 1$ we obtain the desired

$$p_0(t) = e^{-\lambda t}. \tag{5.3}$$

Now we tackle $n > 0$. By independence and stationarity of the increments, we again have

$$\begin{aligned} p_n(t+h) &= \sum_{j=0}^{n} P(N_{t+h} - N_t = j | N_t = n-j) P(N_t = n-j) \\ &= \sum_{j=0}^{n} P(N_{t+h} - N_t = j) P(N_t = n-j) \\ &= \sum_{j=0}^{n} P(N_h = j) p_{n-j}(t) \\ &= p_n(t) P(N_h = 0) + p_{n-1}(t) P(N_h = 1) + \sum_{j=2}^{n} P(N_h = j) p_{n-j}(t). \end{aligned}$$

We estimate the sum by

$$\sum_{j=2}^{n} P(N_h = j) p_{n-j}(t) \leq \sum_{j=2}^{n} P(N_h = j) \leq P(N_h > 1) = o(h),$$

giving us

$$p_n(t+h) = p_n(t)(1 - \lambda h + o(h)) + p_{n-1}(t)(\lambda h + o(h)) + o(h).$$

As before, we obtain

$$\dot{p_n}(t) = -\lambda p_n(t) + \lambda p_{n-1}(t), \quad p_n(0) = 0,$$

or, multiplying with $e^{\lambda t}$,

$$\frac{d}{dt}\left(e^{\lambda t} p_n(t)\right) = \lambda e^{\lambda t} p_{n-1}(t),$$

which can be solved iteratively starting with (5.3). By induction, we get $e^{\lambda t} p_n(t) = (\lambda t)^n / n!$ or

$$p_n(t) = \frac{(\lambda t)^n}{n!} e^{-\lambda t}. \qquad \square$$

Basically, the main conditions of Theorem 5.10 mean that the probability of two "nearly simultaneous" jumps is small compared to the probability of one jump.

Remark 5.11. Instead of recording the number of jumps N_t until time t, we can also look at the jump times. Indeed, let

$$T_n \coloneqq \inf \{\, t \mid N_t \geq n \,\}, \quad n \in \mathbb{N}_0. \tag{5.4}$$

and consider the inter-arrival times

$$\tau_n \coloneqq T_n - T_{n-1}, \quad n \in \mathbb{N}. \tag{5.5}$$

This leads to the following characterization of the Poisson process:

- The inter-arrival times $(\tau_n)_{n\in\mathbb{N}}$ of a Poisson process $(N_t)_{t\geq 0}$ with intensity λ are independent and exponentially distributed with parameter λ (i.e., with expectation $1/\lambda$).

- Given a sequence of i.i.d. exponential random variables $(\sigma_n)_{n\in\mathbb{N}}$, define

$$M_t := \sup\left\{ n \;\middle|\; \sum_{k=1}^{n} \sigma_k \le t \right\}.$$

  Then $(M_t)_{t\ge 0}$ is a Poisson process.

Remark 5.11 gives a way to construct a Poisson process. Indeed, we know how to construct a probability space on which a sequence $(\tau_n)_{n\in\mathbb{N}}$ of i.i.d. exponential random variables can be defined. Therefore, we can construct the Poisson process on this probability space.

Exercise 5.12. Let $(N_t)_{t\ge 0}$ be a Poisson process with intensity λ. Then the conditional distribution of the first jump time T_1 given that $N_t = 1$ is the uniform distribution on $[0, t]$.

Indeed, the assertion of Exercise 5.12 can be generalized: given that $N_t = n$, the jump times $(T_1, \dots, T_n)$ are distributed according to the order statistics of n independent uniform distributions on $[0, t]$. This means that we sample n times from the uniform distribution and order these n random variables according to their size. Then we have obtained a sample from the distribution of $(T_1, \dots, T_n)$ conditioned on $N_t = n$.

5.3 Construction of stochastic processes

As before, by *construction of a stochastic process* $(X_i)_{i\in I}$ we mean the construction of a probability space $(\Omega, \mathcal{F}, P)$ such that the stochastic process $(X_i)_{i\in I}$ can be defined on $(\Omega, \mathcal{F}, P)$. Usually, this will be done by referring to already known constructions. For instance, in the case of the Poisson process, we established an explicit formula of $(N_t)_{t\ge 0}$ in terms of the inter-arrival times, which form an i.i.d. sequence of exponential random variables, see Remark 5.11. However, we know how to construct a probability space supporting an i.i.d. sequence of exponentials as infinite product space of $(]0,\infty[, \mathcal{B}(]0,\infty[), f(x)dx)$ with $f(x) = \exp(-\lambda x)/\lambda$. Thus, the Poisson process can be defined on this product space.

We are going to give constructions for two special cases: first we are going to treat a general Markov process in discrete time, where we assume the transition kernels to be given. Then we extend this to continuous-time processes where all finite-dimensional marginals are given.

More precisely, given a sequence κ_n of Markov kernels on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, we want to construct a Markov process $(X_n)_{n\in\mathbb{N}}$ such that

$$\kappa_n(x, A) = P(X_{n+1} \in A | X_n = x), \quad \forall n \in \mathbb{N},\ x \in \mathbb{R},\ A \in \mathcal{B}(\mathbb{R}).^3 \tag{5.6}$$

By the Markov property, we intuitively expect that those transition probabilities should already carry enough information to construct the process. But before going in this direction, we have to do discuss a bit more about Markov kernels.

Let $(E_n, \mathcal{E}_n)$ be measurable spaces, $n \in \mathbb{N}$. Let κ_1 be a stochastic kernel defined on $E_1 \times \mathcal{E}_2$ and κ_2 a stochastic kernel defined on $(E_1 \times E_2) \times \mathcal{E}_3$. Then we can define a Markov kernel $\kappa_1 \otimes \kappa_2$on $E_1 \times (\mathcal{E}_2 \otimes \mathcal{E}_3)$ by

$$\kappa_1 \otimes \kappa_2(x, A) := \int_{E_2} \kappa_1(x, dy) \int_{E_3} \kappa_2((x, y), dz) \mathbf{1}_A(y, z), \quad A \in \mathcal{E}_2 \otimes \mathcal{E}_3,\ x \in E_1. \tag{5.7}$$

[3] In fact, in the following we will never use specific properties of the measurable space $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, so the result will be valid on any measurable space.

Lemma 5.13. *The map $\kappa_1 \otimes \kappa_2$ defined in (5.7) is a Markov kernel.*

Proof. With a similar argument as in Fubini's theorem, one sees that

$$(x, y) \mapsto \int_{E_3} \kappa_2((x, y), dz)\mathbf{1}_A(y, z)$$

is $\mathcal{E}_1 \otimes \mathcal{E}_2$-measurable for any $A \in \mathcal{E}_2 \otimes \mathcal{E}_3$.[4] Integrating again with respect to κ_1 yields, with the same argument, $\mathcal{E}_1$-measurability of

$$x \mapsto \kappa_1 \otimes \kappa_2(x, A) \coloneqq \int_{E_2} \kappa_1(x, dy) \int_{E_3} \kappa_2((x, y), dz)\mathbf{1}_A(y, z).$$

On the other hand, for fixed x, the set function $A \mapsto \kappa_1 \otimes \kappa_2(x, A)$ is non-negative with $\kappa_1 \otimes \kappa_2(x, E_2 \times E_3) = 1$ and $\kappa_1 \otimes \kappa_2(x, \emptyset) = 0$. Finite additivity of the set function follows immediately by linearity of the integral, and σ-additivity is then implied by monotone convergence. □

Of course, κ_2 may be defined on E_2 only by treating it as a constant function in the first part of the first argument. Moreover, given a probability measure μ on $(E_1, \mathcal{E}_1)$ and a Markov kernel κ on $E_1 \times \mathcal{E}_2$, then we can define a probability measure $\mu \otimes \kappa$ on $\mathcal{E}_1 \otimes \mathcal{E}_2$ by

$$\mu \otimes \kappa(A_1 \times A_2) \coloneqq \int_{A_1} \mu(dx)\kappa(x, A_2), \quad A_1 \in \mathcal{E}_1,\ A_2 \in \mathcal{E}_2, \tag{5.8}$$

extending it to $\mathcal{E}_1 \otimes \mathcal{E}_2$ by Caratheodory. (By Lemma 5.13, $\mu \otimes \kappa$ is a Markov kernel, which does not depend on the first component.) Note that we have a generalized *Fubini theorem* for products of measures and Markov kernels: for any Markov kernel κ defined on $E_1 \times \mathcal{E}_2$, any probability measure μ on $\mathcal{E}_1$ and any (integrable or non-negative) measurable function f defined on $E_1 \times E_2$, we have

$$\int_{E_1 \times E_2} f(x, y)(\mu \otimes \kappa)(d(x, y)) = \int_{E_1} \left(\int_{E_2} f(x, y)\kappa(x, dy) \right) \mu(dx).^5 \tag{5.9}$$

Intuitively, equation (5.9), which also holds, mutatis mutandis, for products of Markov kernels, says that it does not matter in which order we integrate.

Iteratively, given Markov kernels κ_n on $\left(\prod_{i=1}^n E_i\right) \times \mathcal{E}_{n+1}$, the Markov kernel $\kappa_1 \otimes \cdots \otimes \kappa_n$ is defined on $E_1 \times \left(\bigotimes_{i=2}^{n+1} \mathcal{E}_i\right)$ and $\mu \otimes \kappa_1 \otimes \cdots \kappa_n$ is a probability measure on $\bigotimes_{i=1}^{n+1} \mathcal{E}_i$.

Regarding stochastic processes, let $\mu = P_{X_1}$ and

$$\kappa_n((x_1, \ldots, x_n), A) = P(X_{n+1} \in A | X_1 = x_1, \ldots, X_n = x_n).$$

Then

$$\mu \otimes \kappa_1 \otimes \cdots \otimes \kappa_{n-1} = P_{(X_1, \ldots, X_n)}.$$

Of course, in the case of a Markov process, $\kappa_n((x_1, \ldots, x_n), A) = \kappa_n(x_n, A)$ depends only on x_n, $x_1, \ldots, x_n \in \mathbb{R}$, $A \in \mathcal{B}(\mathbb{R})$. In fact, the following theorem asserts that we can go to the limit $n \to \infty$ in this construction.

[4] Given a measurable function $f : E_1 \times E_2 \to \mathbb{R}_{\geq 0}$ and a Markov kernel $\kappa : E_1 \times \mathcal{E}_2 \to [0, 1]$, then the map $x \mapsto \int f(x, y)\kappa(x, dy)$ is measurable. This is obvious for $f = \mathbf{1}_{A_1}\mathbf{1}_{A_2}$ with $A_i \in \mathcal{E}_i$. Then it is easy to show that the system $\mathcal{D}$ of all sets $A \in \mathcal{E}_1 \otimes \mathcal{E}_2$ such that we have measurability with $f = \mathbf{1}_A$ is a Dynkin system containing the measurable rectangles, implying that $\mathcal{D} = \mathcal{E}_1 \otimes \mathcal{E}_2$. Then one can extend the result to any measurable function f as above by the construction of the integral.

[5] For the proof note that the equality holds by definition for $f = \mathbf{1}_{A_1}\mathbf{1}_{A_2}$ with $A_i \in \mathcal{E}_i$. Then one can extend the formula to $f = \mathbf{1}_A$ and finally to general measurable f by a series of approximating arguments.

Theorem 5.14 (Ionescu-Tulcea)**.** *Given a sequence of measurable spaces $(E_n, \mathcal{E}_n)$, $n \in \mathbb{N}$, a probability measure P_1 on $(E_1, \mathcal{E}_1)$ and a sequence of Markov kernels κ_n defined on $\left(\prod_{i=1}^n E_i\right) \times \mathcal{E}_{n+1}$. Set*

$$(\Omega_n, \mathcal{F}_n) := \bigotimes_{i=1}^n (E_i, \mathcal{E}_i),\ n \in \mathbb{N}, \quad (\Omega, \mathcal{F}) := \bigotimes_{n\in\mathbb{N}} (E_n, \mathcal{E}_n),$$

and define probability measures P_n on $(\Omega_n, \mathcal{F}_n)$ by

$$P_n := P_1 \otimes \kappa_1 \otimes \cdots \otimes \kappa_{n-1}.$$

Then there is a unique probability measure P on $(\Omega, \mathcal{F})$ such that

$$\forall n \in \mathbb{N},\ \forall (A_1 \times \cdots \times A_n) \in \bigotimes_{i=1}^n \mathcal{E}_i :\ P\left(A_1 \times \cdots \times A_n \times \prod_{i=n+1}^{\infty} E_i\right) = P_n(A_1 \times \cdots \times A_n).$$

Example 5.15. We want to construct a Markov process $(X_n)_{n\in\mathbb{N}}$ on $\mathbb{R}$. We know that there is always a regular version of the conditional distribution. Assume we are given the Markov kernels

$$\kappa_n(x, A) = P(X_{n+1} \in A | X_n = x), \quad x \in \mathbb{R},\ A \in \mathcal{B}(\mathbb{R}).$$

Moreover, we are given the initial distribution μ of X_1. By Theorem 5.14, there is a (unique) probability measure P on the measurable space

$$(\Omega, \mathcal{F}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))^{\otimes\mathbb{N}}$$

such that the process defined by the projections π_n satisfies the requirements, i.e.,

$$P(\pi_1 \in A) = \mu(A),\ P(\pi_{n+1} \in A | \pi_n = x) = \kappa_n(x, A), \quad x \in \mathbb{R},\ A \in \mathcal{B}(\mathbb{R}),\ n \in \mathbb{N}.$$

Indeed, we can interpret κ_n as Markov kernel on $\mathbb{R}^n \times \mathcal{B}(\mathbb{R})$ as in the theorem.

Proof of Theorem 5.14. Uniqueness of P is clear since P is uniquely determined on the cylinder sets, which form a generator of $\bigotimes_{n\in\mathbb{N}} \mathcal{E}_n$ closed under intersections. So we only have to prove existence.

Let $\mathcal{A}$ denote the system of cylinder sets, i.e., $\mathcal{A} = p^{-1}_{\{1,\dots,n\}}(A)$ for some $n \in \mathbb{N}$ and $A \in \mathcal{F}_n$, where $p_{\{1,\dots,n\}} : \Omega \to \Omega_n$ is the projection map. A set-function P is defined on $\mathcal{A}$ by

$$P(p^{-1}_{\{1,\dots,n\}}(A)) := P_n(A), \quad A \in \mathcal{F}_n.$$

It is clear that $0 \le P \le 1$ and that $P(\Omega) = 1$. Moreover, finite additivity is also clear, since for a finite sequence $A_1, \dots, A_M$ of cylinder sets we can find a common index n such that $A_1, \dots, A_M \in p^{-1}_{\{1,\dots,n\}}(\mathcal{F}_n)$. By Carathéodory's extension theorem, we can extend P to a probability measure on $\mathcal{F}$ provided that P is upper continuous at $\emptyset$, i.e., for any sequence $(A_n)_{n\in\mathbb{N}}$ of cylinder sets with $A_n \searrow \emptyset$ we have $\lim_{n\to\infty} P(A_n) = 0$.

So assume we are giving a decreasing sequence of cylinder sets $A_n = p^{-1}_{\{1,\dots,n\}}(B_n)$ with $\alpha := \inf_n P(A_n) > 0$ and $B_n \in \mathcal{F}_n$.[6] For $1 \le m < n$ and $(x_1, \dots, x_m) \in E_1 \times \cdots \times E_m$ set

$$h_{m,n}(x_1, \dots, x_m) := \left(\delta_{\{(x_1,\dots,x_m)\}} \otimes \bigotimes_{k=m}^{n-1} \kappa_k\right)(B_n),$$

[6]This is a notational simplification. In general, we would have to assume that $A_n = p^{-1}_{\{1,\dots,k_n\}}(B_n)$ for some $B_n \in \mathcal{F}_{k_n}$ with a sequence k_n which we might assume to be increasing. It is not difficult to extend the following arguments to this general situation, but they are more transparent in the simpler case.

$h_m(x_1,\dots,x_m) \coloneqq \inf_{n>m} h_{m,n}(x_1,\dots,x_m)$. We will show that for any n there is a $\rho_n \in E_n$ such that for every m, $h_m(\rho_1,\dots,\rho_m) \geq \alpha$. Since $B_{n+1} \supset B_n \times E_{n+1}$, we have

$$\begin{aligned}
h_{m,n+1}(x_1,\dots,x_m) &= \left(\delta_{\{(x_1,\dots,x_m)\}} \otimes \bigotimes_{k=m}^{n} \kappa_k\right)(B_{n+1}) \\
&\leq \left(\delta_{\{(x_1,\dots,x_m)\}} \otimes \bigotimes_{k=m}^{n} \kappa_k\right)(B_n \times E_{n+1}) \\
&= \left(\delta_{\{(x_1,\dots,x_m)\}} \otimes \bigotimes_{k=m}^{n-1} \kappa_k\right)(B_n) \\
&= h_{m,n}(x_1,\dots,x_m).
\end{aligned}$$

Thus, $h_m = \lim_{n\to\infty} h_{m,n}$ is a decreasing limit. Noting that by the (generalized) Fubini theorem

$$\begin{aligned}
\int_{\Omega_m} h_{m,n} dP_m &= \int_{\Omega_m} \Bigg(\int_{\Omega_n} \delta_{\{x_1,\dots,x_m\}}(d(y_1,\dots,y_m)) \bigotimes_{k=m}^{n-1} \kappa_k((y_1,\dots,y_m), d(y_{m+1},\dots,y_n)) \\
&\qquad \mathbf{1}_{B_n}(y_1,\dots,y_n) \Bigg) P_m(d(x_1,\dots,x_m)) \\
&= \int_{\Omega_n} \left(\int_{\Omega_m} \delta_{\{x_1,\dots,x_m\}}(d(y_1,\dots,y_m)) P_m(d(x_1,\dots,x_m)) \right) \\
&\qquad \bigotimes_{k=m}^{n-1} \kappa_k((y_1,\dots,y_m), d(y_{m+1},\dots,y_n)) \mathbf{1}_{B_n}(y_1,\dots,y_n) \\
&= \int_{\Omega_n} P_m(d(y_1,\dots,y_m)) \bigotimes_{k=m}^{n-1} \kappa_k((y_1,\dots,y_m), d(y_{m+1},\dots,y_n)) \mathbf{1}_{B_n}(y_1,\dots,y_n) \\
&= P_n(B_n),
\end{aligned}$$

where we used that by Fubini's theorem

$$\begin{aligned}
\int_{y\in A} \int_{x\in\Omega_m} \delta_{\{(x_1,\dots,x_m)\}}(dy) P_m(dx) &= \int_{x\in\Omega_m} \int_{y\in A} \delta_{\{(x_1,\dots,x_m)\}}(dy) P_m(dx) \\
&= \int_{x\in\Omega_m} \mathbf{1}_A(x) P_m(dx) = P_m(A)
\end{aligned}$$

for $A \in \mathcal{F}_m$, which can be (by abuse of notation) expressed as

$$\int_{\Omega_m} \delta_{\{(x_1,\dots,x_m)\}}(d(y_1,\dots,y_m)) P_m(d(x_1,\dots,x_m)) = P_m(d(y_1,\dots,y_m)).$$

By the monotone convergence theorem we have

$$\int_{\Omega_m} h_m(x) P_m(dx) = \inf_{n>m} \int_{\Omega_m} h_{m,n}(x) P_m(dx) = \inf_{n>m} P_n(B_n) \geq \alpha.$$

Thus, $h_0 \geq \alpha$. On the other hand, we have

$$\begin{aligned}\int_{E_{m+1}} & h_{m+1}(\rho_1, \dots, \rho_m, y)\kappa_m((\rho_1, \dots, \rho_m), dy) \\ &= \inf_{n>m+1} \int_{E_{m+1}} h_{m+1,n}(\rho_1, \dots, \rho_m, y)\kappa_m((\rho_1, \dots, \rho_m), dy) \\ &= \inf_{n>m+1} h_{m,n}(\rho_1, \dots, \rho_m) \geq h_m(\rho_1, \dots, \rho_m) \geq \alpha,\end{aligned}$$

so that we can, indeed, find $\rho_{m+1} \in E_{m+1}$ with $h_{m+1}(\rho_1, \dots, \rho_{m+1}) \geq \alpha$.

Thus, we have constructed a sequence $(\rho_n)_{n\in\mathbb{N}} \in \Omega$ such that

$$\alpha \leq h_{m,m}(\rho_1, \dots, \rho_m) = \mathbf{1}_{B_m}(\rho_1, \dots, \rho_m),$$

implying that $(\rho_n)_{n\in\mathbb{N}} \in \bigcap_{n\in\mathbb{N}} A_n \neq \emptyset$. □

The other situation we are going to treat is a general stochastic process $(X_i)_{i\in I}$ given by its marginal distributions. Let us assume that the state space of the process is $\mathbb{R}$, i.e., X_i is an $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$-valued random variable, $i \in I$.[7] For any finite subset $J \subset I$, we are given the distribution μ_J of the $\mathbb{R}^{|J|}$-valued random variable $(X_i)_{i\in J}$. In which cases is it possible to construct a probability space $(\Omega, \mathcal{F}, P)$ and random variables $X_i : \Omega \to \mathbb{R}$, $i \in I$, such that we have

$$\forall J \subset I,\ |J| < \infty :\ P_{(X_i)_{i\in J}} = \mu_J? \tag{5.10}$$

If such a process $(X_i)_{i\in I}$ exists, we call the family of probability measures $(\mu_J)_{J\in F(I)}$ the family of *finite-dimensional marginal distributions* of the process, where we have used the notation $F(I) := \{\, J \subset I \mid |J| < \infty \,\}$. Of course, there needs to be some kind of consistency requirement. After all, given $J_1 \subset J_2 \in F(I)$, then μ_{J_1} is already determined by μ_{J_2}: denoting with $p_{J_2,J_1} : \mathbb{R}^{J_2} \to \mathbb{R}^{J_1}$ to projection mapping $(x_i)_{i\in J_2}$ to $(x_i)_{i\in J_1}$, we have

$$(\mu_{J_2})_{p_{J_2,J_1}} = \mu_{J_1}. \tag{5.11}$$

It turns out, however, that this *consistency condition* is already sufficient. For the statement of the theorem let us also introduce the notation p_J for the projection $(x_i)_{i\in I} \mapsto (x_i)_{i\in J}$ for $J \subset I$.

Theorem 5.16 (Kolmogorov's extension theorem)**.** *Given a polish space E with the Borel σ-algebra $\mathcal{B}(E)$, an index set I and a consistent family $(\mu_J)_{J\in F(I)}$ of probability measures on $(E, \mathcal{B}(E))^{\otimes J}$, respectively, i.e., a family satisfying (5.11) for any $J_1 \subset J_2 \in F(I)$.*[8] *Then there is a unique probability measure μ on $(E, \mathcal{B}(E))^{\otimes I}$ such that*

$$\forall J \in F(I) :\ \mu_{p_J} = \mu_J.$$

μ is called the projective limit *of the probability measures (μ_J) and denoted by*

$$\mu = \varprojlim_{J\uparrow I} \mu_J.$$

[7] We may treat processes with a polish state space in the same way, but one can show that it is not possible to generalize this approach to general measurable state spaces.

[8] We could easily choose different polish spaces E_i, $i \in I$. Since the proof remains the same (except for more complicated notations), and the stochastic process with changing state spaces seem to be rather exotic, we only treat the standard case here.

Put differently, choose $(\Omega, \mathcal{F}, P) = (E^I, \mathcal{B}(E)^{\otimes I}, \mu)$ *and* $X_i = \pi_i := p_{\{i\}}$*, then the stochastic process* $(X_i)_{i\in I}$ *has the finite dimensional marginals* $(\mu_J)_{J\in F(I)}$*. The process* $(X_i)_{i\in I} = (\pi_i)_{i\in I}$ *defined like this is known as the* canonical process*.*

Proof. We prove the theorem in two stages.
1. Countable index set I.
If the index set I is countable, then we may assume $I = \mathbb{N}$. For $n \in \mathbb{N}$ consider $\mu_n := \mu_{\{1,\dots,n\}}$, a probability measure on $(E^n, \mathcal{B}(E)^{\otimes n})$. Note that E^n is again polish and $\mathcal{B}(E)^{\otimes n} = \mathcal{B}(E^n)$. Therefore, there is a regular conditional distribution

$$\kappa_{n-1}((x_1,\dots,x_{n-1}), A) = \mu_n(\pi_n \in A | \pi_1 = x_1, \dots, \pi_{n-1} = x_{n-1}),$$

$A \in \mathcal{B}(E)$, $x_1, \dots, x_{n-1} \in E$. By Lemma 4.16, the Markov kernel κ_{n-1} defined on $E^{n-1} \times \mathcal{B}(E)$ satisfies for all $A \in \mathcal{B}(E^n)$:

$$\begin{aligned}\mu_n(A) &= \int_{E^n} \mathbf{1}_A(x_1,\dots,x_n)\kappa_{n-1}((x_1,\dots,x_{n-1}), dx_n)(\mu_n)_{p_{\{1,\dots,n\},\{1,\dots,n-1\}}}(d(x_1,\dots,x_{n-1})) \\ &= \int_{E^n} \mathbf{1}_A(x_1,\dots,x_n)\kappa_{n-1}((x_1,\dots,x_{n-1}), dx_n)\mu_{n-1}(d(x_1,\dots,x_{n-1})),\end{aligned}$$

where we used the consistency condition (5.11). Thus, we have iteratively

$$\mu_n = \mu_1 \otimes \kappa_1 \otimes \cdots \kappa_{n-1}, \quad n \in \mathbb{N},$$

and the assertion follows from Theorem 5.14.
2. Uncountable index set I
Let $G(I) := \{ J \subset I \mid J \text{ countable} \}$. Recall that

$$\mathcal{B}(E)^{\otimes I} = \bigcup_{J\in G(I)} p_J^{-1}\left(\mathcal{B}(E)^{\otimes J}\right) = \bigcup_{J\in G(I)} \sigma(p_J).^{9} \tag{5.12}$$

For any $J \in G(I)$, by the first step of the proof we can construct a probability measure μ_J on $(E^J, \mathcal{B}(E^J))$ with $\mu_K = (\mu_J)_{p_{J,K}}$, $K \in F(I)$, $K \subset J$. Thus, given two index sets $J, J' \in G(I)$, we know that

$$\mu_J(p_{J,K}^{-1}(A)) = \mu_{J'}(p_{J',K}^{-1}(A)) = \mu_K(A)$$

for all cylinder sets $A \in \mathcal{B}(E)^{\otimes K}$ with $K \subset J \cap J'$. Since the cylinder sets form a generator closed under intersections, this means that the probability measures $\widetilde{\mu}_J$ and $\widetilde{\mu}_{J'}$ defined on $\sigma(p_J)$ and $\sigma(p_{J'})$, respectively, by $\widetilde{\mu}_J(A) = \mu_J(B)$ with $A = p_J^{-1}(B)$ (and similarly for J') coincide on $\sigma(p_J) \cap \sigma(p_{J'})$, i.e.,

$$\forall A \in \sigma(p_J) \cap \sigma(p_{J'}) : \ \widetilde{\mu}_J(A) = \widetilde{\mu}_{J'}.$$

Using (5.12), we may therefore define a set function μ on $\mathcal{B}(E)^{\otimes I}$ by

$$\mu(A) := \widetilde{\mu}_J(A) \text{ for } J \in G(I) \text{ with } A \in \sigma(p_J), \quad A \in \mathcal{B}(E)^{\otimes I}.$$

We need to show that μ is a probability measure. Obviously, $0 \le \mu \le 1$ and $\mu(E^I) = 1$. Given a sequence of disjoint measurable sets A_n, let $J_n \in G(I)$ such that $A_n \in \sigma(p_{J_n})$, $n \in \mathbb{N}$. Then $J := \bigcup_{n\in\mathbb{N}} J_n \in G(I)$ and $\forall n : \ A_n \in \sigma(p_J)$, $\bigcup_{n\in\mathbb{N}} A_n \in \sigma(p_J)$, implying that

$$\mu\left(\bigcup_{n\in\mathbb{N}} A_n\right) = \widetilde{\mu}_J\left(\bigcup_{n\in\mathbb{N}} A_n\right) = \sum_{n\in\mathbb{N}} \widetilde{\mu}_J(A_n) = \sum_{n\in\mathbb{N}} \mu(A_n). \qquad \square$$

[9] Indeed, the right hand side is a σ-algebra, since countable unions of countable sets are countable.

Chapter 6

Brownian motion

Recall that a (standard, one-dimensional) Brownian motion is a stochastic process $(B_t)_{t\geq 0}$ with a.s. continuous paths, $B_0 = 0$ and independent stationary increments with distribution $B_t - B_s \sim \mathcal{N}(0, t-s)$ for $t > s \geq 0$, see Definition 5.9. We note that one could also consider Brownian motions with drifts, i.e., processes $X_t = \mu t + \sigma B_t$ and multi-dimensional Brownian motions, which are vector-valued processes whose individual components are Brownian motions such that the increments are multi-dimensional Gaussians. The first question is whether such a process exists.

6.1 Construction of the Brownian motion

In light of Section 5.3, the obvious way to construct a probability space $(\Omega, \mathcal{F}, P)$ carrying the Brownian motion would be following Kolmogorov's extension theorem (Theorem 5.16). Indeed, given any finite ensemble of time points $0 \leq t_1 < t_2 < \cdots < t_n$, it easily follows from the definition of Brownian motion that

$$(B_{t_1}, \ldots, B_{t_n}) \sim \mathcal{N}(0, \Sigma), \quad \Sigma = (\sigma_{i,j})_{i,j=1}^n \text{ with } \sigma_{i,j} = \min(t_i, t_j),\ 1 \leq i, j \leq n. \tag{6.1}$$

It is also easy to see that the family of distributions defined by (6.1) is consistent in the sense of (5.11). Thus, Theorem 5.16 indeed implies the existence of a probability space $(\Omega, \mathcal{F}, P)$ and of a family $(B_t)_{t\geq 0}$ of random variables defined thereon such that

- $B_0 = 0$,
- the increments $B_{t_n} - B_{t_{n-1}}, B_{t_{n-1}} - B_{t_{n-2}}, \ldots, B_{t_1} - 0$ are independent and $B_t - B_s =^d B_{t-s}$ whenever $0 < t_1 < \cdots < t_n$ and $t > s > 0$,
- $B_t \sim \mathcal{N}(0, t)$.

Exercise 6.1. Verify that one can indeed define such a probability space carrying a process satisfying those conditions using Theorem 5.16.

Continuity of paths, however, is a property that cannot directly be obtained by Kolmogorov's extension theorem. Indeed, there is no reason that the process $(B_t)_{t\geq 0}$ obtained by Kolmogorov's extension theorem is continuous.[1] Another famous theorem of

[1] Recall that the probability space constructed in Theorem 5.16 is $(\mathbb{R}, \mathcal{B}(\mathbb{R}))^{\otimes[0,\infty[}$ and the process is just the evaluation map $B_t(\omega) = \omega(t)$. By (5.12), any set $A \in \mathcal{B}(\mathbb{R})^{\otimes[0,\infty[}$ is already determined by the function

Kolmogorov, the *Kolmogorov-Chentsov theorem* provides the existence of a (Hölder-) continuous *version* of the process $(B_t)_{t\geq 0}$, i.e., the existence of a process $(W_t)_{t\geq 0}$ (defined on the probability space $(\Omega, \mathcal{F}, P)$) such that the paths $t \mapsto W_t(\omega)$ are continuous (for every ω) and

$$\forall t \geq 0 : \ P(B_t = W_t) = 1. \tag{6.2}$$

Obviously, if two processes are versions of each other, then properties of finite-dimensional marginal distributions are the same. Thus, $(W_t)_{t\geq 0}$ satisfies all the requirements of Definition 5.9 and is a Brownian motion.

Example 6.2. We construct two processes $(X_t)_{t\in[0,1]}$ and $(Y_t)_{t\in[0,1]}$ which are versions of each other but not almost surely equal. We choose the probability space $([0,1], \mathcal{B}([0,1]), \lambda|_{[0,1]})$ and set $X_t(\omega) \equiv 1$ and $Y_t(\omega) := 1 - \mathbf{1}_{\{t\}}(\omega)$. Then for every t we have $P(X_t = Y_t) = 1$, but $P(\forall t : \ X_t = Y_t) = 0$. Note that each path of X but no path of Y is continuous.

We will give a direct construction avoiding the Kolmogorov-Chentsov theorem. Let $(Y_{k,n})_{0\leq k<2^n,\ n\in\mathbb{N}_0}$ be a sequence of independent standard normal random variables. Since this is a countable number of independent random variables, we know that we construct the sequence on the probability space

$$(\Omega, \mathcal{F}, P) = (\mathbb{R}, \mathcal{B}(\mathbb{R}), \mu)^{\otimes\mathbb{N}},$$

where $\mu(dx) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$. For any $n \in \mathbb{N}_0$ define a stochastic process $(X_n(t))_{t\in[0,1]}$ by

$$\forall \frac{k}{2^n} \leq t < \frac{k+1}{2^n} : \ X_n(t) = \frac{\frac{k+1}{2^n} - t}{2^{-n}} X_n\left(\frac{k}{2^n}\right) + \frac{t - \frac{k}{2^n}}{2^{-n}} X_n\left(\frac{k+1}{2^n}\right), \tag{6.3a}$$

$$X_0(0) = 0, \quad X_0(1) = Y_{0,0}, \tag{6.3b}$$

$$X_{n+1}\left(\frac{k}{2^n}\right) = X_n\left(\frac{k}{2^n}\right), \ X_{n+1}\left(\frac{2k+1}{2^{n+1}}\right) = X_n\left(\frac{2k+1}{2^{n+1}}\right) + 2^{-(n+2)/2} Y_{2k+1,n+1}, \tag{6.3c}$$

Where we always use $n \in \mathbb{N}_0$ and, for fixed n, $0 \leq k < 2^n$. Moreover, for understanding (6.3c) note that

$$X_n\left(\frac{2k+1}{2^{n+1}}\right) = \frac{X_n\left(\frac{k}{2^n}\right) + X_n\left(\frac{k+1}{2^n}\right)}{2}.$$

We will show that $B_t := \lim_{n\to\infty} X_n(t)$ is a Brownian motion on $[0,1]$. Finally, given a sequence of independent Brownian motions B_n defined on the intervals $[0,1]$, $n \in \mathbb{N}$, we can easily see that the concatenated process

$$B_t := \sum_{k=1}^{\lfloor t \rfloor} B_k(1) + B_{\lfloor t \rfloor + 1}(t - \lfloor t \rfloor)$$

defines a Brownian motion on $[0, \infty[$. Note that this Brownian motion can also be defined on the same probability space $(\Omega, \mathcal{F}, P)$, but, for simplicity, one can also take $(\Omega, \mathcal{F}, P)^{\otimes\mathbb{N}}$.

Before proving the main result, we have to provide a few auxiliary lemmas.

values $\omega(t_i)$ for a sequence of time points $(t_i)_{i\in\mathbb{N}}$. Since the set $C([0,\infty[: \mathbb{R})$ of continuous functions does not have this property, it cannot be a measurable set with respect to the product σ-algebra. Thus, the question whether $(B_t)_{t\geq 0}$ has almost surely continuous paths, is only reasonable in the sense of an outer probability. There is certainly a way around this technicality, but our construction will avoid the problem altogether.

Lemma 6.3. *Given a sequence of (d-dimensional) L^p-random variables converging a.s. to some random variable X, $p \geq 1$. Assume that there is $r > p$ such that $\sup_n E\left[|X_n|^r\right] < \infty$. Then $X \in L^p$ and the convergence holds in L^p.*

Proof. By Fatou's lemma, we have

$$E\left[|X|^r\right] \leq \liminf_{n\to\infty} E\left[|X_n|^r\right] \leq \sup_n E\left[|X_n|^r\right] < \infty,$$

implying that $\sup_n E\left[|X_n - X|^r\right] < \infty$. By Hölder's inequality, we have

$$E\left[|X_n - X|^p\right] \leq \epsilon^p + E\left[|X_n - X|^p \mathbf{1}_{\{|X_n - X| > \epsilon\}}\right] \leq \epsilon^p + E\left[|X_n - X|^r\right]^{p/r} P\left(|X_n - X| > \epsilon\right)^{(r-p)/r}.$$

Thus, for any $\epsilon > 0$ we have $\limsup_{n\to\infty} E\left[|X_n - X|^p\right] \leq \epsilon^p$, implying the assertion. □

Exercise 6.4. Using Borel-Cantelli, show that any sequence X_n converging to a random variable X in probability has an almost surely converging subsequence. Then adapt the proof of the lemma and show that we can replace the assumption of almost sure convergence by convergence in probability.

Lemma 6.5. *Given a sequence $(X_n)_{n\in\mathbb{N}}$ of d-dimensional centered Gaussian random variables converging almost surely to a d-dimensional random variable X. Then $X_n \to X$ in L^2 and X is also Gaussian.*

Proof. From the exercises we know that $X_n \to X$ in L^2 implies that X has the normal distribution. Since almost sure convergence implies convergence in distribution, which in turn implies convergence of the characteristic functions, we know that the covariance matrices Σ_n of X_n converge to some matrix Σ. Note that for some matrix A_n such that $\Sigma_n = A_n^T A_n$ (existence of which is guaranteed by the Cholesky factorization) we have $X_n =^d A_n Z$ for any $Z \sim \mathcal{N}(0, I_d)$. On the other hand, using the operator norm corresponding to the Euclidean distance on $\mathbb{R}^d$, we have $|A_n Z| \leq \|A_n\| \, |Z|$. In particular, we have

$$E\left[|X_n|^3\right] \leq \|A_n\|^3 \, E\left[|Z|^3\right] = \lambda_{\max}(\Sigma_n)^{3/2} E\left[|Z|^3\right],$$

using that the operator-norm of A_n is the square root of the largest eigenvalue of Σ_n. Thus, we have $\sup_n E[|X_n|^3] < \infty$ and the result follows from Lemma 6.3. □

Lemma 6.6. *For any n, the vector $(X_n(k/2^n))_{k=0}^n$ is a centered Gaussian random vector with covariance $E\left[X_n(k/2^n)X_n(l/2^n)\right] = \min(k/2^n, l/2^n)$.*

Proof. We prove the assertion by induction. It is true for $n = 0$. Assume it has been established for n. By (6.3), the vector $(X_{n+1}(k/2^{n+1}))$ is a linear combination of the independent centered Gaussian vectors $(X_n(k/2^n))$ and $(Y_{k,n+1})$. By general properties of the normal distribution, it is thus a centered Gaussian random vector. Let us show the covariance formula for $k = l$ (the general case works in the same way). If k is even, then $X_{n+1}(k/2^{n+1}) = X_n((k/2)/2^n)$ and the formula holds by induction. Otherwise, letting $k = 2m + 1$, we have

$$E\left[X_{n+1}\left(\frac{2m+1}{2^{n+1}}\right)^2\right] = E\left[\left(\frac{X_n(\frac{m}{2^n}) + X_n\left(\frac{m+1}{2^n}\right)}{2} + 2^{-(n+2)/2} Y_{2m+1,n+1}\right)^2\right]$$

$$= \frac{1}{4}\left(\frac{m}{2^n} + \frac{m+1}{2^n} + \frac{2m}{2^n}\right) + 2^{-(n+2)} = \frac{2m+1}{2^{n+1}}. \quad \square$$

Theorem 6.7. *The process $(B_t)_{t\in[0,1]}$ defined by $B_t := \lim_{n\to\infty} X_n(t)$ exists (in the a.s. sense) and satisfies the properties of a Brownian motion.*

Proof. First note that Markov's inequality for a random variable $Y \sim \mathcal{N}(0,1)$ implies that there is a well-known constant C such that for any $a > 0$, $P(|Y| > a) \le Ca^{-8}$. Moreover, since X_n and X_{n+1} are linearly interpolated between the grid points (see (6.3a) together with (6.3c)), we have

$$\sup_{t\in[k/2^n,(k+1)/2^n]} |X_{n+1}(t) - X_n(t)| > \epsilon \Longrightarrow 2^{-(n+2)/2} \left|Y_{2k+1,n+1}\right| > \epsilon.$$

Thus,

$$\begin{aligned} P\left(\sup_{t\in[0,1]} |X_{n+1}(t) - X_n(t)| > 2^{-n/4}\right) &= P\left(\bigcup_{k=0}^{2^n-1} \left\{ \sup_{t\in[k/2^n,(k+1)/2^n]} |X_{n+1}(t) - X_n(t)| > 2^{-n/4} \right\}\right) \\ &\le P\left(\bigcup_{k=0}^{2^n-1} \left\{ 2^{-(n+2)/2} \left|Y_{2k+1,n+1}\right| > 2^{-n/4} \right\}\right) \\ &\le 2^n P\left(|Y| > 2^{(n+4)/4}\right) \\ &\le C2^{-n-8}. \end{aligned}$$

Thus, the first part of the Borel-Cantelli lemma (Theorem 2.26) implies that there is a P-nullset N such that for every $\omega \in N^c$ the sequence of continuous functions $X_n(\cdot)(\omega)$ is a Cauchy sequence in $C([0,1])$. Thus, the limiting process B_t exists and $t \mapsto B_t(\omega)$ is continuous for every $\omega \in N^c$ – and for every $\omega \in \Omega$ if we choose $B_t(\omega)$ appropriately for $\omega \in N$.

We are left to prove the requirements on the finite-dimensional marginals as laid down in Definition 5.9. First note that $B_0 \equiv 0$ since $X_n(0) \equiv 0$ for any n. Now given $0 < t_1 < \cdots < t_K$, we consider

$$(B_{t_K} - B_{t_{K-1}}, \ldots, B_{t_2} - B_{t_1}, B_{t_1}) = \lim_{n\to\infty} \underbrace{(X_n(t_K) - X_n(t_{K-1}), \ldots, X_n(t_2) - X_n(t_1), X_n(t_1))}_{=:Z_n} \text{ a.s.}$$

By the interpolation property (6.3a), we can describe Z_n as a linear map applied to the random vector

$$\overline{Z}_n := (X_n(\lfloor t_1 \rfloor), X_n(\lceil t_1 \rceil), \ldots, X_n(\lfloor t_K \rfloor), X_n(\lceil t_K \rceil)),$$

where $\lfloor t_j \rfloor \le t_j \le \lceil t_j \rceil$ are the elements of $\{ k/2^n \mid k = 0, \ldots, 2^n \}$ closest to t_j, $j = 1, \ldots, K$. Therefore, Z_n is a centered normal random variable with covariance matrix Σ_n. By Lemma 6.5, this implies that $(B_{t_K} - B_{t_{K-1}}, \ldots, B_{t_2} - B_{t_1}, B_{t_1})$ is centered normal with covariance matrix $\Sigma := \lim_{n\to\infty} \Sigma_n$. For any $s \le t$ we have

$$\begin{aligned} E[X_n(s)X_n(t)] &= E\left[\left(\frac{\lceil s \rceil - s}{2^{-n}} X_n(\lfloor s \rfloor) + \frac{s - \lfloor s \rfloor}{2^{-n}} X_n(\lceil s \rceil)\right)\left(\frac{\lceil t \rceil - t}{2^{-n}} X_n(\lfloor t \rfloor) + \frac{t - \lfloor t \rfloor}{2^{-n}} X_n(\lceil t \rceil)\right)\right] \\ &= \frac{\lceil s \rceil - s}{2^{-n}} \frac{\lceil t \rceil - t}{2^{-n}} \lfloor s \rfloor + \frac{\lceil s \rceil - s}{2^{-n}} \frac{t - \lfloor t \rfloor}{2^{-n}} \lfloor s \rfloor + \\ &\quad + \frac{s - \lfloor s \rfloor}{2^{-n}} \frac{\lceil t \rceil - t}{2^{-n}} \min(\lceil s \rceil, \lfloor t \rfloor) + \frac{s - \lfloor s \rfloor}{2^{-n}} \frac{t - \lfloor t \rfloor}{2^{-n}} \lceil s \rceil \\ &= s + O(2^{-n}), \end{aligned}$$

using that $|s - \lfloor s \rfloor| = O(2^{-n})$, $|s - \lceil s \rceil| = O(2^{-n})$ and likewise for t. Thus, Σ is, indeed, of the form guaranteeing the distributional properties of the Brownian motion. □

6.2 Properties of the Brownian motion

Lemma 6.8. *Given a Brownian motion $(B_t)_{t\geq 0}$ defined on $(\Omega, \mathcal{F}, P)$ and constants $c, s > 0$.*

(i) The process $c^{-1}B_{c^2 t}$, $t \geq 0$, is again a Brownian motion.

(ii) The process $W_t := B_{t+s} - B_s$, $t \geq 0$, is again a Brownian motion.

(iii) The process $W_t := tB_{1/t}$, $t > 0$, with $W_0 := 0$ is again a Brownian motion.

Proof. Left as an exercise. □

Property (i) can be interpreted as a scaling property of the Brownian motion. In particular, it implies that $B_t \sim \sqrt{t}$ for t small.

Lemma 6.9. *The Brownian motion is a martingale in its natural filtration.*

Proof. Left as an exercise. □

Fix some interval $[s, s+t]$ and consider a *partition* $\mathcal{D}$ of the interval, i.e., a finite set $\mathcal{D} = \{ s = t_0 < \cdots t_n = t \}$. The *mesh* of such a partition is defined by

$$\|\mathcal{D}\| := \sup_k |t_{k+1} - t_k| .$$

Theorem 6.10. *Given a sequence $\mathcal{D}_n$ of partitions of the interval $[s, s+t]$ such that $\|\mathcal{D}_n\| \to 0$. We write $\mathcal{D}_n = \{ t_0^{(n)} < \cdots < t_{m_n}^{(n)} \}$. Then the random variables*

$$S_n := \sum_{k=0}^{m_n - 1} \left(B_{t_{k+1}^{(n)}} - B_{t_k^{(n)}} \right)^2 \xrightarrow[n\to\infty]{L^2} t.$$

Proof. By property (ii) in Lemma 6.8, we may assume that $s = 0$. Dropping superscripts, we see that

$$\begin{aligned} S_n - t &= \sum_{i=0}^{m-1} \left((B_{t_{i+1}} - B_{t_i})^2 - (t_{i+1} - t_i) \right) \\ &= \sum_{i=0}^{m-1} \left((B_{t_{i+1}} - B_{t_i})^2 - E\left[(B_{t_{i+1}} - B_{t_i})^2 \right] \right). \end{aligned}$$

Since we have decomposed $S_n - t$ into a sum of independent centered random variables, we obtain

$$\begin{aligned} E\left[(S_n - t)^2 \right] &= \sum_{i=0}^{m-1} E\left[\left((B_{t_{i+1}} - B_{t_i})^2 - E\left[(B_{t_{i+1}} - B_{t_i})^2 \right] \right)^2 \right] \\ &= \sum_{i=0}^{m-1} E\left[\left((B_{t_{i+1}} - B_{t_i})^2 - (t_{i+1} - t_i) \right)^2 \right]. \end{aligned}$$

We note that $B_{t_{i+1}} - B_{t_i} \sim \sqrt{t_{i+1} - t_i} X$ with $X \sim \mathcal{N}(0,1)$. Thus,

$$\begin{aligned} E\left[(S_n - t)^2\right] &= \sum_{i=0}^{m-1} (t_{i+1} - t_i)^2 E\left[(X^2-1)^2\right] \\ &\leq E\left[(X^2-1)^2\right] \sum_{i=0}^{m-1} \|\mathcal{D}_n\| (t_{i+1} - t_i) \\ &= 2t \|\mathcal{D}_n\|, \end{aligned}$$

which converges to 0 for $n \to \infty$. □

Remark. Under certain assumptions on the sequence of partitions, we can even obtain almost sure convergence to t. However, it can be shown that almost sure convergence does not hold for general sequences of partitions.

Exercise 6.11. Assume that the sequence of partitions satisfies $\sum_n \|\mathcal{D}_n\| < \infty$. Show that $S_n \to t$ a.s.
Hint: Use the Borel-Cantelli lemma.

The limit of the random variables S_n is known as the *quadratic variation* of the process $(B_t)_{t\geq 0}$ on the interval $[s, s+t]$. Specializing the result to $s = 0$, the theorem can be rephrased as saying that the Brownian motion has quadratic variation t. Recall that the *total variation* of a function f on an interval $[s, s+t]$ is defined by

$$V_{s,s+t}(f) := \sup_{\mathcal{D}} \sum_{k=0}^{m-1} |f(t_{k+1}) - f(t_k)|.$$

Corollary 6.12. *The total variation of the Brownian motion on any interval $[s, s+t]$ of positive length t is a.s. infinite.*

Proof. Assume that the Brownian motion had a finite total variation on a set A with $P(A) > 0$. Take a sequence of partitions $\mathcal{D}_n$ with $\sum_n \|\mathcal{D}_n\| < \infty$. Then

$$\begin{aligned} S_n = \sum_{k=0}^{m-1} (B_{t_{k+1}} - B_{t_k})^2 &\leq \sup_{k=0,\dots,m-1} \left|B_{t_{k+1}} - B_{t_k}\right| \sum_{k=0}^{m-1} \left|B_{t_{k+1}} - B_{t_k}\right| \\ &\leq \sup_{k=0,\dots,m-1} \left|B_{t_{k+1}} - B_{t_k}\right| V_{s,s+t}(B). \end{aligned}$$

As continuous functions on the bounded interval $[s, s+t]$, the maps $u \mapsto B_u(\omega)$ are uniformly continuous, for all $\omega \in \Omega$. Thus, $\sup_{k=0,\dots,m-1} \left|B_{t_{k+1}}(\omega) - B_{t_k}(\omega)\right|$ converges to 0 as $\|\mathcal{D}_n\| \to 0$. On the other hand, for $\omega \in A$, we have $V_{s,s+t}(B)(\omega) < \infty$, implying that

$$\forall \omega \in A : \; S_n(\omega) \xrightarrow[n\to\infty]{} 0,$$

in contradiction to Theorem 6.10 together with Exercise 6.11. □

As already discussed above, Brownian motion is a Markov process and a martingale with respect to its natural filtration $\mathcal{F}_t := \sigma(B_s : 0 \leq s \leq t)$, $t \geq 0$. Most of the properties of discrete-time martingales discussed in Section 4.3 and 4.4 also hold in the continuous case, at least in the case of continuous processes – otherwise technical subtleties frequently occur. We recall some of these properties and definitions applied

to Brownian motion. As in discrete time, a random variable $\tau : \Omega \to [0, \infty]$ is called *stopping time* iff

$$\forall t \geq 0 : \ \{ \tau \leq t \} \in \mathcal{F}_t. \tag{6.4}$$

Let $A \subset \mathbb{R}$ be either open or closed, then the hitting time

$$\tau_A \coloneqq \inf \{ t \in [0, \infty[\mid B_t \in A \} \tag{6.5}$$

is a stopping time.[2] Given a stopping time τ, we again define the σ-algebra of the τ-past by

$$\mathcal{F}_\tau \coloneqq \{ A \in \mathcal{F} \mid \forall t \geq 0 : \ A \cap \{ \tau \leq t \} \in \mathcal{F}_t \}. \tag{6.6}$$

As in the discrete case, the random variable B_τ is $\mathcal{F}_\tau$-measurable, whenever $P(\tau < \infty) = 1$.

Lemma 6.13. *Let τ be a stopping time with $P(\tau < \infty) = 1$. Consider the process $(X_t)_{t \geq 0}$ defined by*

$$X_t \coloneqq B_{\tau + t} - B_\tau.$$

Then $(X_t)_{t \geq 0}$ is again a Brownian motion. Moreover, the filtration $(\mathcal{G}_t)_{t \geq 0}$ generated by $(X_t)_{t \geq 0}$ is independent of $\mathcal{F}_\tau$.

We omit the proof of the lemma.[3] The property given in Lemma 6.13 is known as *strong Markov property*, because it implies that for any stopping time τ and any $t \geq 0$ and any $A \in \mathcal{B}(\mathbb{R})$, we have

$$P(B_{\tau+t} \in A | \mathcal{F}_\tau) = P(B_{\tau+t} \in A | B_\tau) \text{ on } \{ \tau < \infty \}$$

almost surely, which is a direct generalization of the Markov property to stopping times.

Exercise 6.14. Reflection principle, see Karatzas and Shreve.

Example 6.15. In the following, we shall study very particular hitting times. For $a > 0$ consider $\tau_a \coloneqq \tau_{[a, \infty[}$ and for $a < 0$ let $\tau_a \coloneqq \tau_{]-\infty, a]}$ be the first times that the Brownian motion reaches the level a. Moreover, for $a < b$ let $\tau_{a,b} \coloneqq \tau_{]a,b[^c}$ be the first time that the Brownian motion leaves the interval $]a, b[$. By continuity, note that $B_{\tau_a} = a$ a.s. Form Lemma 6.9 and the exercises, we know that $(B_t)_{t \geq 0}$ and $\left(B_t^2 - t\right)_{t \geq 0}$ are martingales. We assume that the optional sampling theorem holds for both martingales with the stopping time τ.[4] Then, we have

$$E[B_\tau] = 0 = aP(B_\tau = a) + bP(B_\tau = b),$$
$$E\left[B_\tau^2\right] = E[\tau] = a^2 P(B_\tau = a) + b^2 P(B_\tau = b).$$

From the first equation we get

$$P(B_\tau = b) = \frac{|a|}{|a| + |b|}, \ P(B_\tau = a) = \frac{|b|}{|a| + |b|},$$

and inserting into the second equation we get $E[B_\tau^2] = E[\tau] = |a|\,|b|$.

[2] As an example for the above mentioned subtleties, we note that continuity is essential for τ_A being a stopping time for closed A.

[3] The proof is rather straightforward when τ takes only finitely many values. Otherwise, the stopping time is approximated by discrete stopping times.

[4] While τ is an unbounded stopping time, the claim still seems justified since the processes $B_{\min(\tau,t)}$ and $B^2_{\min(\tau,t)}$ are bounded. Thus, we can approximate τ by bounded stopping times τ_n, using the dominated convergence theorem. For bounded stopping times, we can extend the proof of the optional sampling theorem by approximation with stopping times taking only finitely many values. For details see, for instance, Karatzas and Shreve or Breiman.

6.3 Donsker's invariance principle

Let $(Y_n)_{n\in\mathbb{N}}$ be an i.i.d. sequence of square integrable random variables with $\sigma^2 := \mathrm{var}[Y_1]$. Consider the random walk $S_n := Y_1 + \cdots + Y_n$. We rescale the random walk and extend it to a process $X_n(t)$ defined on $[0,1]$ by

$$X_n(t) := \frac{S_{\lfloor nt \rfloor}}{\sigma\sqrt{n}}, \quad t \in [0,1], \tag{6.7}$$

where $S_0 := 0$ and $\lfloor t \rfloor$ denotes the largest integer smaller or equal to t. By the central limit theorem, we know that $X_n(1) \xrightarrow[n\to\infty]{d} \mathcal{N}(0,1)$, but much more is true: we even have that the process $(X_n(t))_{t\in[0,1]}$ converges weakly to $(B_t)_{t\in[0,1]}$ for $n \to \infty$. This result is known as *Donsker's theorem*, *invariance principle* (owing to the fact that the limiting process does not depend on the random walk) and *functional central limit theorem*. But first, we will show that the random walk S_n can be *embedded* into the Brownian motion.

Theorem 6.16 (Skorohod's embedding theorem)**.** *Given any random variable Y with $E[Y] = 0$ and $E[Y^2] < \infty$. Then there is a probability space $(\Omega, \mathcal{F}, P)$ and a Brownian motion $(B_t)_{t\geq 0}$ random variables $U \leq 0 \leq V$ defined thereon such that $B_{\tau_{U,V}}$ has the same distribution as Y.*

Proof. Let us first assume that Y takes only two values $u \leq 0 \leq v$ with corresponding probabilities p and $q = 1 - p$. Then we take $U \equiv u$ and $V \equiv v$ and Example 6.15 implies that $B_{\tau_{U,V}}$ has the same distribution as Y.

In the second step, assume that Y takes values in some finite set. This means that there are numbers $y_1, \ldots, y_K$ such that $P(Y \in \{y_1, \ldots, y_K\}) = 1$ with $p_i := P(Y = y_i)$. We claim that there are numbers $u_1, \ldots, u_n \leq 0$ and $v_1, \ldots, v_n \geq 0$ and positive numbers $q_1, \ldots, q_n$ and $r_1, \ldots, r_n$ such that

$$\begin{gathered} q_1 + \cdots + q_n + r_1 + \cdots + r_n = 1, \\ \forall i \in \{1, \ldots, n\} : \; q_i u_i + r_i v_i = 0, \\ q_1 \delta_{\{u_1\}} + \cdots + q_n \delta_{\{u_n\}} + r_1 \delta_{\{v_1\}} + \cdots + r_n \delta_{\{v_n\}} = P_Y. \end{gathered}$$

This means that $\{u_1, \ldots, u_n, v_1, \ldots, v_n\} = \{y_1, \ldots, y_K\}$ and that for every $1 \leq i \leq K$ we have

$$\sum_{j:\, u_j = y_i} q_j + \sum_{j:\, v_j = y_i} r_j = p_i,$$

where we note that one of these sums will be 0. Note that we do not require the individual us and vs to be distinct! The claim is easily proved by induction on K. Then define random variables (U, V) independent of the Brownian motion by setting

$$P((U,V) = (u_i, v_i)) = q_i + r_i, \quad i = 1, \ldots, n.$$

By independence of the Brownian motion and the random variables (U, V) and the Example 6.15, we have

$$P(B_{\tau_{U,V}} = u_i | (U,V) = (u_i, v_i)) = \frac{|v_i|}{|u_i| + |v_i|} = \frac{v_i}{\frac{r_i}{q_i} v_i + v_i} = \frac{q_i}{r_i + q_i}.$$

Thus,

$$P(B_{\tau_{U,V}} = u_i) = \sum_{j:\, u_j = u_i} P(B_{\tau_{U,V}} = u_j | (U,V) = (u_j, v_j)) P((U,V) = (u_j, v_j))$$
$$= \sum_{j:\, u_j = u_i} \frac{q_j}{r_j + q_j}(r_j + q_j) = \sum_{j:\, u_j = u_i} q_j = p_i,$$

and likewise for positive values v_i. Therefore, we see that, indeed, $B_{\tau_{U,V}}$ has the same distribution as Y.

In the final step, let Y be any square integrable centered random variable. We can approximate Y by centered random variables Y_n with finite support in the sense that $Y_n \xrightarrow[n\to\infty]{d} Y$. Consider the corresponding random boundaries (U_n, V_n) such that $B_{\tau_{U_n,V_n}}$ has the same distribution as Y_n. Without giving the proof, we claim that at least along a subsequence we have

$$(U_n, V_n) \xrightarrow[n\to\infty]{d} (U, V)$$

for some random variables (U, V).[5] Applying Example 6.15 to the factorized conditional probabilities, we get with $\tau := \tau_{U,V}$ and some interval $I \subset]0, \infty[$ that

$$P(B_\tau \in I | U, V) = \mathbf{1}_I(V) \frac{|U|}{|U| + |V|}$$

and by the same reasoning we have for $\tau_n := \tau_{U_n,V_n}$ that

$$P(B_{\tau_n} \in I | U_n, V_n) = \mathbf{1}_I(V_n) \frac{|U_n|}{|U_n| + |V_n|}.$$

Thus, if the distribution functions of Y and V are continuous at the boundaries of I, we have

$$P(Y \in I) = \lim_{n\to\infty} P(Y_n \in I) = \lim_{n\to\infty} P(B_{\tau_n} \in I)$$
$$= \lim_{n\to\infty} E\left[\mathbf{1}_I(V_n) \frac{|U_n|}{|U_n| + |V_n|}\right] = E\left[\mathbf{1}_I(V) \frac{|U|}{|U| + |V|}\right] = E[B_\tau \in I].$$

We can likewise argue for intervals $I \subset]-\infty, 0[$. Since the set of points at which the distribution functions of Y, V and U are continuous is certainly dense, we have obtained the result. □

Theorem 6.17. *Given an i.i.d. sequence of random variables $(Y_n)_{n\in\mathbb{N}}$ with $E[Y_1] = 0$ and $\sigma^2 := E[Y_1^2] < \infty$. Then there is a probability space, on which we can define a Brownian motion $(B_t)_{t\geq 0}$ and an i.i.d. sequence $(\tau_n)_{n\in\mathbb{N}}$ of non-negative, integrable random variables independent of the Brownian motion, such that $E[\tau_1] = \sigma^2$ and such that the sequence $\left(B_{\sum_{k=1}^n \tau_k}\right)_{n\in\mathbb{N}}$ has the same distribution as the random walk $(S_n)_{n\in\mathbb{N}}$ with $S_n := Y_1 + \cdots + Y_n$.*

Proof. By Theorem 6.16, there are random variables (U, V) such that $B_{\tau_{U,V}}$ has the same distribution as Y_1 when (U, V) are independent of the Brownian motion B. Take one probability space $(\Omega_1, \mathcal{F}_1, P_1)$ on which one can define a Brownian motion $(B_t)_{t\geq 0}$ and

[5] The proof depends on compactness results in the weak topology.

one probability space $(\Omega_2, \mathcal{F}_2, P_2)$ on which one can define an i.i.d. sequence of random variables (U_n, V_n) having the same distribution as (U, V). Then we choose

$$(\Omega, \mathcal{F}, P) \coloneqq (\Omega_1, \mathcal{F}_1, P_1) \otimes (\Omega_2, \mathcal{F}_2, P_2),$$

on which we can jointly define a Brownian motion $(B_t)_{t\geq 0}$ and an independent i.i.d. sequence $((U_n, V_n))_{n\in\mathbb{N}}$. We define a filtration

$$\mathcal{F}_t \coloneqq \sigma\left(\{ B_s \mid s \leq t \} \cup \{ U_1, V_1 \}\right).$$

Obviously, both $(B_t)_{t\geq 0}$ and $(B_t^2 - t)_{t\geq 0}$ are martingales with respect to the filtration, and $\tau_n \coloneqq \tau_{U_n,V_n}$ is a stopping time, $n \in \mathbb{N}$. Without proof, we collect the following results:

- The strong Markov property holds with respect to the stopping times τ_n, i.e., $B_s^{(1)} \coloneqq B_{\tau_1+s} - B_{\tau_1}$ is a Brownian motion, and the filtration $\mathcal{F}_s^{(1)}$ generated by $B^{(1}$ and U_2, V_2 is independent of $\mathcal{F}_{\tau_1}$.
- We set $\tau_2 \coloneqq \inf\left\{ t \geq 0 \;\middle|\; B_t^{(1)} \in]U_2, V_2[^c \right\}$.
- This construction can be iterated, leading to stopping times τ_{n+1} and Brownian motions $B_s^{(n+1)} \coloneqq B_{\tau_{n+1}+s}^{(n)} - B_{\tau_{n+1}}^{(n)}$ such that the filtration $\mathcal{F}_s^{(n+1)}$ generated by $B^{(n+1)}$ and U_{n+2}, V_{n+2} is independent of $\mathcal{F}_{\tau_{n+1}}^{(n)}$.
- The optional sampling theorem holds for the martingales $B^{(n)}$ and $(B^{(n)})^2 - t$ (with $B^{(0)} \coloneqq B$) and the stopping time τ_{n+1}, i.e.,

$$E\left[(B_{\tau_{n+1}}^{(n)})^2\right] = E[\tau_{n+1}].$$

Then we see that S_1 and B_{τ_1} have the same distribution by Theorem 6.16. Moreover, we have

$$\sigma^2 = E[S_1^2] = E[B_{\tau_1}^2] = E[\tau_1].$$

By the other properties obtained above, the stopping times τ_n are i.i.d. and $B_{\tau_{n+1}}^{(n)}$ has the same distribution as Y_n. Note that

$$\begin{aligned} B_{\tau_{n+1}}^{(n)} &= B_{\tau_n+\tau_{n+1}}^{(n-1)} - B_{\tau_n}^{(n-1)} \\ &= B_{\tau_{n-1}+\tau_n+\tau_{n+1}}^{(n-2)} - B_{\tau_{n-1}}^{(n-2)} - B_{\tau_n+\tau_{n-1}}^{(n-2)} + B_{\tau_{n-1}}^{(n-2)} \\ &= B_{\tau_{n-1}+\tau_n+\tau_{n+1}}^{(n-2)} - B_{\tau_n+\tau_{n-1}}^{(n-2)} \\ &= B_{\tau_1+\cdots+\tau_{n+1}} - B_{\tau_1+\cdots+\tau_n}. \end{aligned}$$

Therefore,

$$S_n = \sum_{k=1}^{n} Y_k =^d \sum_{k=1}^{n} B_{\tau_k}^{(k-1)} = \sum_{k=1}^{n} (B_{\tau_1+\cdots+\tau_k} - B_{\tau_1+\cdots+\tau_{k-1}}) = B_{\tau_1+\cdots+\tau_n},$$

where "$=^d$" denotes equality in distribution. □

Theorem 6.18. *Given a centered, square integrable i.i.d. sequence $(Y_n)_{n\in\mathbb{N}}$ and the corresponding random walk $(S_n)_{n\in\mathbb{N}}$ as above. Let $X_n(t) \coloneqq \frac{1}{\sigma\sqrt{n}} S_{\lfloor nt \rfloor}$ for $n \in \mathbb{N}$ and $t \in [0, 1]$. Then there is a probability space $(\Omega, \mathcal{F}, P)$ with a Brownian motion $(B_t)_{t\in[0,1]}$ and a sequence of processes $(\overline{X}_n(t))_{t\in[0,1]}$ such that $X_n(\cdot)$ and $\overline{X}_n(\cdot)$ have the same distributions and such that for any subsequence (n_k) increasing fast enough we have*

$$\lim_{k\to\infty} \sup_{0\leq t\leq 1} \left|\overline{X}_{n_k}(t) - B_t\right| = 0 \text{ a.s.}$$

From the theorem we easily obtain the main result of the section.

Corollary 6.19 (Donsker's theorem)**.** *In the setting of Theorem 6.18, the centered, scaled, continuously interpolated random walks $(X_n(t))_{t\in[0,1]}$ converge weakly to the Brownian motion $(B_t)_{t\in[0,1]}$, i.e., for any bounded continuous function $f : C([0,1];\mathbb{R}) \to \mathbb{R}$ we have*

$$\lim_{n\to\infty} E\left[f((X_n(t))_{t\in[0,1]})\right] = E\left[f\left((B_t)_{t\in[0,1]}\right)\right].$$

Proof. For any subsequence (n_k) increasing fast enough, Theorem 6.18 and continuity of f imply that

$$\lim_{k\to\infty} f\left((\overline{X}_{n_k}(t))_{t\in[0,1]}\right) = f\left((B_t)_{t\in[0,1]}\right) \text{ a.s.}$$

Consequently, along any such subsequence we have

$$\lim_{k\to\infty} E\left[f\left((X_{n_k}(t))_{t\in[0,1]}\right)\right] = \lim_{k\to\infty} E\left[f\left((\overline{X}_{n_k}(t))_{t\in[0,1]}\right)\right] = E\left[f\left((B_t)_{t\in[0,1]}\right)\right],$$

which implies the convergence of the full sequence. □

Proof of Theorem 6.18. Without loss of generality we may assume that $\sigma = 1$. We construct $(\Omega, \mathcal{F}, P)$ as in Theorem 6.17 and define the Brownian motion $(B_t)_{t\geq 0}$ thereon. By Lemma 6.8, $t \mapsto \sqrt{n}B_{t/n}$ is also a Brownian motion, independent of $(U_k, V_k)_{k\in\mathbb{N}}$. Thus, we can also embed the random walk in the Brownian motion $\sqrt{n}B_{t/n}$ using stopping times $(\tau_k^{(n)})_{k\in\mathbb{N}}$ such that S_k has the same distribution as

$$\sqrt{n}B\left(\frac{\tau_1^{(n)} + \cdots + \tau_k^{(n)}}{n}\right).$$

Consequently, $X_n(t)$ has the same distribution as

$$\overline{X}_n(t) \coloneqq B\left(\frac{\tau_1^{(n)} + \cdots + \tau_{\lfloor nt\rfloor}^{(n)}}{n}\right).$$

Note that for fixed n the sequence of random variables $\tau_k^{(n)}$ is an i.i.d. sequence of random variables with $E\left[\tau_k^{(n)}\right] = 1$. On the other hand, for fixed k, the random variables $\tau_k^{(n)}$ are identically distributed but not independent. We claim that for any subsequence (n_k) increasing sufficiently fast we have

$$Z_{n_k} \coloneqq \sup_{0\leq t\leq 1}\left|\frac{\tau_1^{(n_k)} + \cdots + \tau_{\lfloor n_k t\rfloor}^{(n_k)}}{n_k} - t\right| \xrightarrow[k\to\infty]{\text{a.s.}} 0. \tag{6.8}$$

Accepting (6.8), continuity of the paths of the Brownian motion B implies that, indeed,

$$\lim_{k\to\infty} \sup_{0\leq t\leq 1}\left|\overline{X}_{n_k}(t) - B_t\right| = 0 \text{ a.s.}$$

We are left to prove (6.8). If we can prove that $Z_n \to 0$ in probability, then the Borel-Cantelli lemma implies that $Z_{n_k} \to 0$ a.s. for subsequences increasing fast enough. For the proof of convergence in probability, note that

$$\begin{aligned} Z_n = \sup_{0\leq t\leq 1}\left|\frac{\tau_1^{(n)} + \cdots + \tau_{\lfloor nt\rfloor}^{(n)}}{n} - t\right| &= \sup_{0\leq t\leq 1}\left|\frac{\tau_1^{(n)} + \cdots + \tau_{\lfloor nt\rfloor}^{(n)} - \lfloor nt\rfloor}{n} + \frac{\lfloor nt\rfloor - nt}{n}\right| \\ &\leq \sup_{0\leq t\leq 1}\left|\frac{\overline{\tau}_1^{(n)} + \cdots + \overline{\tau}_{\lfloor nt\rfloor}^{(n)}}{n}\right| + \frac{1}{n} \leq \sup_{0\leq t\leq 1} t\left|\frac{\overline{\tau}_1^{(n)} + \cdots + \overline{\tau}_{\lfloor nt\rfloor}^{(n)}}{\lfloor nt\rfloor}\right| + \frac{1}{n} \end{aligned}$$

with $\overline{\tau}_k^{(n)} := \tau_k^{(n)} - 1$. Ignoring the $1/n$-term, we note that there are two different regimes going on: for t small, the term is small because of the factor t in front, whereas for t close to 1, the term is small because of some law of large numbers. Indeed, for any fixed $\epsilon > 0$ we can further bound Z_n by

$$\begin{aligned} Z_n &\le \epsilon \sup_{0\le t\le\epsilon} \left| \frac{\overline{\tau}_1^{(n)} + \cdots + \overline{\tau}_{\lfloor nt\rfloor}^{(n)}}{\lfloor nt\rfloor} \right| + \sup_{\epsilon\le t\le 1} \left| \frac{\overline{\tau}_1^{(n)} + \cdots + \overline{\tau}_{\lfloor nt\rfloor}^{(n)}}{\lfloor nt\rfloor} \right| + \frac{1}{n} \\ &\le \epsilon \sup_{k\ge 1} \left| \frac{\overline{\tau}_1^{(n)} + \cdots + \overline{\tau}_k^{(n)}}{k} \right| + \sup_{k\ge\lfloor\epsilon n\rfloor} \left| \frac{\overline{\tau}_1^{(n)} + \cdots + \overline{\tau}_k^{(n)}}{k} \right| + \frac{1}{n} \\ &=^d \epsilon \sup_{k\ge 1} \left| \frac{\overline{\tau}_1^{(1)} + \cdots + \overline{\tau}_k^{(1)}}{k} \right| + \sup_{k\ge\lfloor\epsilon n\rfloor} \left| \frac{\overline{\tau}_1^{(1)} + \cdots + \overline{\tau}_k^{(1)}}{k} \right| + \frac{1}{n}. \end{aligned}$$

We start with the middle term. By the law of large numbers,

$$\lim_{k\to\infty} \frac{\overline{\tau}_1^{(1)} + \cdots + \overline{\tau}_k^{(1)}}{k} = 0 \text{ a.s.},$$

implying that the second term converges to 0 in probability for $n \to \infty$. Thus, for any $x > 0$, we have

$$\limsup_{n\to\infty} P(Z_n > x) \le P\left(\epsilon \sup_{k\ge 1} \left| \frac{\overline{\tau}_1^{(1)} + \cdots + \overline{\tau}_k^{(1)}}{k} \right| > x \right).$$

Letting $\epsilon \to 0$, we see that Z_n converges to 0 in probability. □

Appendix A

Collection of results from elementary probability

Multi-dimensional normal distribution

Definition A.1. A measure ν on $\mathcal{B}(\mathbb{R})$ is called *Gaussian* or *normal* if its Fourier transform is of the form

$$\widehat{\nu}(u) = e^{iau - \frac{1}{2}\sigma^2 u^2},$$

for some constants $a \in \mathbb{R}$ and $\sigma \geq 0$. A measure ν on $\mathcal{B}(\mathbb{R}^n)$ is Gaussian or normal if and only if the image measures ν_h under all linear functionals $h : \mathbb{R}^n \to \mathbb{R}$ are one-dimensional Gaussian.

Remark A.2. If $\sigma > 0$, then a Gaussian measure in the above sense has the density $\frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-a)^2/(2\sigma^2)}$. Thereof ore, the above definition corresponds to the usual definition in the case of a positive variance, but also includes the Dirac measures $\delta_{\{a\}}$. The same is true in the multi-dimensional case: if the covariance matrix Σ is invertible, then the Gaussian measure has the usual density. Again, also singular measures are included in the definition.

We write $\mathcal{N}(\mu, \Sigma)$ for the Gaussian distribution with expectation $\mu \in \mathbb{R}^n$ and covariance matrix $\Sigma \in \mathbb{R}^{n\times n}$. Thus, an n-dimensional random vector X has a multi-dimensional normal distribution if and only if $\langle \lambda \, , X \rangle$ has a one-dimensional normal distribution, for each $\lambda \in \mathbb{R}^n$. We collect some properties of the normal distribution.

Lemma A.3. *Given an n-dimensional random vector X.*

(i) $X \sim \mathcal{N}(\mu, \Sigma)$ if and only if for every $u \in \mathbb{R}^n$: $\varphi_X(u) = \exp\left(i \langle u \, , \mu \rangle - \frac{1}{2} \langle u \, , \Sigma u \rangle\right)$. Moreover, the normal measure $\mathcal{N}(\mu, \Sigma)$ exists for any $\mu \in \mathbb{R}^n$ and any symmetric, positive semi-definite matrix $\Sigma \in \mathbb{R}^{n\times n}$.

(ii) Let $X \sim \mathcal{N}(\mu, \Sigma)$ and $A \in \mathbb{R}^{k\times n}$, $b \in \mathbb{R}^k$, then $AX + b \sim \mathcal{N}(A\mu + b, A\Sigma A^T)$.

Inequalities

In what follows, X and Y are real random variables satisfying the required integrability conditions. The first inequality holds because the covariance is a positive semi-definite bi-linear form (on $L^2(\Omega)$).

Lemma A.4 (Cauchy-Schwarz inequality)**.** *We have*

$$\mathrm{Cov}[X, Y]^2 \leq V[X]V[Y].$$

More generally, we also have the Hölder inequality.

Lemma A.5 (Hölder's inequality)**.** *Given $X \in L^p(\Omega, \mathcal{F}, P)$ and $Y \in L^q(\Omega, \mathcal{F}, P)$ with $1 = 1/p + 1/q$, $p, q \geq 1$ (including the possibility of $p = \infty$), then $XY \in L^1$ and*

$$E[|XY|] \leq E[|X|^p]^{1/p} E[|Y|^q]^{1/q}.$$

Lemma A.6 (Markov's[1] inequality)**.** *Given an increasing function $u : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ such that $u(|X|)$ is integrable. Then for every $\epsilon > 0$*

$$P(|X| \geq \epsilon) \leq \frac{E[u(|X|)]}{u(\epsilon)}.$$

Corollary A.7 (Chebychev's[2] inequality)**.** *If X is square integrable, then for every $\epsilon > 0$*

$$P(|X - E[X]| \geq \epsilon) \leq \frac{V[X]}{\epsilon^2}.$$

Lemma A.8 (Jensen's inequality)**.** *Given a convex function f such that $f(X)$ is integrable, then*

$$E[f(X)] \geq f(E[X]).$$

[1] German: Markow.
[2] German: Tschebyschow.

Appendix B

Characteristic functions

Given a probability measure μ on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$, we recall that the *characteristic function* is defined by

$$\widehat{\mu}(u) \coloneqq \int_{\mathbb{R}^n} e^{i\langle u,x\rangle}\mu(dx), \quad u \in \mathbb{R}^n. \tag{B.1}$$

In particular, for a random variable X defined on the probability space $(\Omega, \mathcal{F}, P)$ with values in $\mathbb{R}^n$, we set $\varphi_X \coloneqq \widehat{P_X}$. Note that the integral of a $\mathbb{C}$-valued measurable function f is simply defined by

$$\int f d\mu = \int \Re f d\mu + i \int \Im f d\mu,$$

and that a function f is $\mathcal{B}(\mathbb{C})$-$\mathcal{B}(\mathbb{R}^n)$-measurable if and only if $\Re f$ and $\Im f$ are $\mathcal{B}(\mathbb{R})$-$\mathcal{B}(\mathbb{R}^n)$-measurable. This can be seen from the fact that $\mathcal{B}(\mathbb{C}) \equiv \mathcal{B}(\mathbb{R}^2) = \mathcal{B}(\mathbb{R}) \otimes \mathcal{B}(\mathbb{R})$. Thus,

$$f_u \coloneqq x \mapsto e^{i\langle u,x\rangle}$$

is measurable and μ-integrable for every $u \in \mathbb{R}^n$, since $|f_u| = 1$.

Example B.1. If μ has a density f with respect to the n-dimensional Lebesgue measure, then

$$\widehat{\mu}(u) = \int_{\mathbb{R}^n} e^{i\langle u,x\rangle} f(x)dx$$

is the *Fourier transform* of the function f. Thus, we can invert the characteristic function to get back the density by

$$f(x) = \frac{1}{(2\pi)^n} \int_{\mathbb{R}^n} e^{-i\langle x,u\rangle}\widehat{\mu}(u)du.$$

We will next show that the situation is typical: indeed, the characteristic function $\widehat{\mu}$ characterizes the probability measure μ.

Theorem B.2. *A probability measure μ is uniquely determined by its characteristic function. More precisely, given two probability measures μ and ν on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ with $\widehat{\mu} = \widehat{\nu}$, then $\mu = \nu$.*

We note that the theorem is actually a theorem from analysis, and it is not surprising that its proof is more analytic than probabilistic. Before giving the proof, we collect some auxiliary results. We start with one of the main theorems from functional analysis.

Lemma B.3 (Stone-Weierstrass theorem)**.** *Let E be a compact metric space and $\mathbb{K} = \mathbb{R}$ or $\mathbb{K} = \mathbb{C}$. Consider $C \subset C(E, \mathbb{K})$ be a point-separating algebra*[1]*, which is also closed under complex conjugation in case $\mathbb{K} = \mathbb{C}$. Then C is dense in $C_b(E, \mathbb{K})$ with respect to $\|\cdot\|_\infty$.*

From measure theory, we recall the following result.

Lemma B.4. *Let E be a metric space and μ a probability measure on $(E, \mathcal{B}(E))$. Then for every $A \in \mathcal{B}(E)$ we have $\mu(A) = \sup\{\, \mu(K) \mid K \subset A \text{ compact} \,\}$.*

Lemma B.5. *The family $C_b(\mathbb{R}^n, \mathbb{R})$ separates probability measures on $\mathbb{R}^n$, i.e., given two probability measures μ and ν on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ with $\int f d\mu = \int f d\nu$ for every bounded continuous function f, then necessarily $\mu = \nu$.*

Proof. By Lemma B.4, it suffices to prove that $\mu(K) = \nu(K)$ for every compact set $K \subset \mathbb{R}^n$. Given such a compact set K, one can find a sequence ρ_n of uniformly bounded, continuous functions converging to $\mathbf{1}_K$.[2] By dominated convergence, we can conclude that

$$\mu(K) = \lim_{n\to\infty} \int \rho_n d\mu = \lim_{n\to\infty} \int \rho_n d\nu = \nu(K). \qquad \square$$

If $\mathbb{R}^n$ were a compact set, then we could easily conclude by combining Lemma B.3 with Lemma B.5. Since this is not the case, we need to invest a bit more work.

Proof of Theorem B.2. Assume we are given probability measures μ and ν such that $\widehat{\mu} = \widehat{\nu}$. Moreover, let $f \in C_b(\mathbb{R}^n, \mathbb{C})$ and $0 < \epsilon < 2$. We can find $N \in \mathbb{N}$ large enough that

$$(1 + 2\|f\|_\infty)(\mu + \nu)(\mathbb{R}^n \setminus [-N, N]^n) < \epsilon/2.$$

Let $C_N \coloneqq \left\{ f_u|_{[-N,N]^n} \;\middle|\; u \in 2\pi\mathbb{Z}^n \right\}$ and $\mathcal{D}_N$ the generated algebra. Then $\mathcal{D}_N$ is dense in $C([-N, N]^n, \mathbb{C})$ by Lemma B.3, implying that we can find $g \in \mathcal{D}_N$ such that

$$\sup\{\, |f(x) - g(x)| \mid x \in [-N, N]^n \,\} < \epsilon/4.$$

Now we interpret g as a function on $\mathbb{R}^n$. By linearity of the integral, $\int g d\mu = \int g d\nu$. Since g is $\mathbb{Z}^n$-periodic, we have $\|f - g\|_\infty \le \|f\|_\infty + \|g\|_\infty \le 2\|f\|_\infty + \epsilon/2 < 1 + 2\|f\|_\infty$. Finally, we have

$$\begin{aligned}
\left| \int f d\mu - \int f d\nu \right| &\le \left| \int_{[-N,N]^n} (f - g) d\mu \right| + \left| \int_{[-N,N]^n} (f - g) d\nu \right| + \\
&\quad + \left| \int_{\mathbb{R}^n \setminus [-N,N]^n} (f - g) d\mu \right| + \left| \int_{\mathbb{R}^n \setminus [-N,N]^n} (f - g) d\nu \right| \\
&\le \frac{\epsilon}{4} + \frac{\epsilon}{4} + \|f - g\|_\infty \mu(\mathbb{R}^n \setminus [-N, N]^n) + \|f - g\|_\infty \nu(\mathbb{R}^n \setminus [-N, N]^n) \\
&\le \epsilon.
\end{aligned}$$

[1] C is an algebra, i.e., $1 \in C$, $f, g \in C$ imply that $fg \in C$, $f + g \in C$ and $f \in C$, $\alpha \in \mathbb{K}$ imply that $\alpha f \in C$. Moreover, for $x \neq y \in E$ we can find $f \in C$ with $f(x) \neq f(y)$.

[2] Find an open set U with $\operatorname{dist}(K, U^c) < 1/n$. From topology we know that we can find a continuous function ρ_n which is 1 on K, 0 on U^c and bounded by 1 in general.

Since ϵ was arbitrary, we have $\int f d\mu = \int f d\nu$ for every $f \in C_b(\mathbb{R}^n, \mathbb{C})$. By Lemma B.5, this shows that $\mu = \nu$. □

We collect some simple properties of the Fourier transform.

Lemma B.6. *Given a probability measure μ on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$.*

(i) $\widehat{\mu}$ is continuous.

(ii) For all $u \in \mathbb{R}^n$ we have $\left|\widehat{\mu}(u)\right| \leq 1 = \widehat{\mu}(0)$.

(iii) $\widehat{\mu}$ is positive semi-definite*, i.e., for every collection $u_1, \ldots, u_k \in \mathbb{R}^n$, the matrix $A \in \mathbb{C}^{k\times k}$ given by $A_{i,j} = \widehat{\mu}(u_i - u_j)$ is positive semi-definite.*

The proof is left as an exercise.

Remark B.7. In fact, one can prove that $\widehat{\mu}$ is uniformly continuous. A famous theorem by S. Bochner shows that, conversely, any continuous, positive semi-definite function $f : \mathbb{R}^n \to \mathbb{C}$ is the Fourier transform of a finite measure on $\mathcal{B}(\mathbb{R}^n)$, which is a probability measure if $f(0) = 1$.

In analysis, it is well known that derivatives of the Fourier transforms are related to the *moments* of the underlying function or measure. For simplicity, we give the statement in dimension $n = 1$ only.

Theorem B.8. *Assume that the probability measure μ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ has a finite n'th moment, i.e., $M_n := \int x^n \mu(dx)$ exists as a real number. Then $\widehat{\mu}$ is n times continuously differentiable and*

$$\widehat{\mu}^{(k)}(0) = i^k M_k, \quad k \leq n.$$

Proof. Since $\partial_u^k f_u(x) = (ix)^k f_u(x)$, we have $\left|\partial_u^k f_u(x)\right| = |x|^k$, which is μ-integrable for $k \leq n$. Now fix $k < n$ and assume that we have already established that

$$\widehat{\mu}^{(k)}(u) = i^k \int_{\mathbb{R}} x^k e^{iux} \mu(dx) = \int_{\mathbb{R}} \frac{\partial^k}{\partial u^k} f_u(x) \mu(dx).$$

Since $\frac{\partial^{k+1}}{\partial u^{k+1}} f_u$ is uniformly bounded by an integrable function, we may apply Lemma B.9 to get

$$\widehat{\mu}^{(k+1)}(u) = \frac{d}{du} \int_{\mathbb{R}} \frac{\partial^k}{\partial u^k} f_u(x) \mu(dx) = \int_{\mathbb{R}} \frac{\partial^{k+1}}{\partial u^{k+1}} f_u(x) \mu(dx) = i^{k+1} \int_{\mathbb{R}} x^{k+1} e^{iux} \mu(dx),$$

implying the statement for $u = 0$. (Strictly speaking, we need to apply Lemma B.9 for both the real and imaginary parts of the integral.) □

Lemma B.9. *Let $(A, \mathcal{A}, \mu)$ be some measure space, $I \subset \mathbb{R}$ an open interval, $I \neq \emptyset$, and $f : A \times I \to \mathbb{R}$ a function satisfying:*

(i) $\forall y \in I : \; x \mapsto f(x, y) \in L^1(\mu)$.

(ii) For almost every $x \in \Omega$ the map $y \mapsto f(x, y)$ is differentiable with derivative denoted by $f'(x, y)$.

(iii) $g := x \mapsto \sup_{y\in I} |f'(x, y)| \in L^1(\mu)$.

Then for every $y \in I$ the function $x \mapsto f'(x,y) \in L^1(\mu)$ and

$$\frac{d}{dy}\int_A f(x,y)\mu(dx) = \int_A f'(x,y)\mu(dx).$$

Proof. Left as an exercise. □

Example B.10. Let $X \sim \mathcal{N}(\mu, \sigma^2)$. We want to compute its characteristic function. Note that $X = \mu + \sigma Y$, implying that $\varphi_X(u) = e^{i\mu u}\varphi_Y(\sigma u)$. Thus, it suffices to compute φ_Y. By Lemma B.9, we have

$$\begin{aligned}
\frac{d}{du}\varphi_Y(u) &= \int_{\mathbb{R}} ixe^{iux}\frac{1}{\sqrt{2\pi}}e^{-x^2/2}dx \\
&= -i\int_{\mathbb{R}} e^{iux}\frac{\partial}{\partial x}\frac{1}{\sqrt{2\pi}}e^{-x^2/2}dx \\
&= i\int_{\mathbb{R}} \frac{\partial}{\partial x}e^{iux}\frac{1}{\sqrt{2\pi}}e^{-x^2/2}dx - i\underbrace{\left[e^{iux}\frac{1}{\sqrt{2\pi}}e^{-x^2/2}\right]_{-\infty}^{+\infty}}_{=0} \\
&= -u\int_{\mathbb{R}} e^{iux}\frac{1}{\sqrt{2\pi}}e^{-x^2/2}dx = -u\varphi_Y(u),
\end{aligned}$$

where we used integration by parts. But the ordinary differential equation

$$\dot{\varphi_Y}(u) = -u\varphi_Y(u), \quad \varphi_Y(0) = 1,$$

has the unique solution $\varphi_Y(u) = e^{-u^2/2}$. Thus, we have showed that

$$\varphi_X(u) = e^{iu\mu - \frac{1}{2}\sigma^2 u^2}. \tag{B.2}$$

Now assume that $Z \sim \mathcal{N}(\mu, \Sigma)$ is a d-dimensional normal random variable. This means that $\langle \lambda, Z\rangle \sim \mathcal{N}(\langle \lambda, \mu\rangle, \langle \lambda, \Sigma\lambda\rangle)$ for every $\lambda \in \mathbb{R}^d$. In particular, for $u \in \mathbb{R}^d$,

$$\varphi_Z(u) = E[e^{i\langle u, Z\rangle}] = \varphi_{\langle u, Z\rangle}(1) = e^{i\langle u, \mu\rangle - \frac{1}{2}\langle u, \Sigma u\rangle}, \tag{B.3}$$

using (B.2).

Appendix C

Weak convergence and the central limit theorem

In the following, we are going to discuss probability measures on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ or random variables defined on some probability space $(\Omega, \mathcal{F}, P)$ taking values in $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. Therefore, we are going to omit these qualifications in the current chapter. Note that most of the results are also true for polish spaces $(E, \mathcal{B}(E))$.

From measure theory, we recall the following types of convergence of random variables X_n:

- X_n converges to a random variable X almost surely, iff $P\left(\lim_{n\to\infty} X_n = X\right) = 1$;
- for $p \geq 1$ we say that X_n converges to X in L^p iff $X_n, X \in L^p(\Omega, \mathcal{F}, P)$ and

$$\lim_{n\to\infty} \|X - X_n\|_{L^p} = \lim_{n\to\infty} \left(E\left[|X - X_n|^p\right]\right)^{1/p} = 0.$$

We know that those two types of convergence are "independent" in the sense that none of them implies the other. Both of them imply *convergence in measure* or *convergence in probability*, i.e., the statement that

$$\forall \epsilon > 0 : \lim_{n\to\infty} P\left(|X - X_n| \geq \epsilon\right) = 0. \tag{C.1}$$

Lemma C.1. *Given a sequence of $\mathbb{R}^d$-valued random variables X_n and another $\mathbb{R}^d$-valued random variable X on $(\Omega, \mathcal{F}, P)$. Assume that either $X_n \to X$ almost surely or $X_n \to X$ in $L^p(\Omega, \mathcal{F}, P)$, $p \geq 1$. Then $X_n \to X$ in probability.*

On the other hand, if $X_n \to X$ in probability, then there is a subsequence $(n_k)_{k\in\mathbb{N}}$ such that $X_{n_k} \to X$ almost surely.

Proof. From Bürger's lecture notes we already know that almost sure convergence implies convergence in probability. Therefore, we may assume that $X_n \to X$ in L^p. For fixed $\epsilon > 0$, Markov's inequality implies that

$$\lim_{n\to\infty} P\left(|X - X_n| \geq \epsilon\right) \leq \lim_{n\to\infty} \frac{E\left[|X - X_n|^p\right]}{\epsilon^p} = 0.$$

For the second part, fix $k \in \mathbb{N}$. By convergence in probability, we can find an $n_k \geq n_{k-1}$ such that

$$P(A_k) \leq \frac{1}{k^2}, \text{ where } A_k := \left\{ \left|X_{n_k} - X\right| \geq \frac{1}{k} \right\}.$$

In particular, we have $\sum_k P(A_k) \leq \sum_k 1/k^2 < \infty$. Thus, Theorem 2.26 implies that

$$P\left(\limsup_{k\to\infty} A_k\right) = 0.$$

However, $\omega \notin \limsup_{k\to\infty} A_k$ implies that $\left|X_{n_k}(\omega) - X(\omega)\right| < 1/k$ for all k large enough. Thus,

$$\forall \omega \notin \limsup_{k\to\infty} A_k : \lim_{k\to\infty} X_{n_k}(\omega) = X(\omega). \qquad \square$$

Weak convergence is even weaker then convergence in probability, and is, in fact, more a type of convergence of measures.

Definition C.2. Given a sequence of probability measures μ_n and a single probability measure μ on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. We say that μ_n converges *weakly* to μ, symbolically $\mu_n \xrightarrow{d} \mu$ iff

$$\forall f \in C_b(\mathbb{R}^d) : \lim_{n\to\infty} \int f d\mu_n = \int f d\mu.$$

A sequence of random variables taking values in $\mathbb{R}^d$ converges *weakly* or *in distribution* to a random variable X on $\mathbb{R}^d$ iff the distributions of X_n converge weakly to the distribution of X. In this case, we write $X_n \xrightarrow{d} X$.

We note that the random variables X_n do not need to be defined on one common probability space $(\Omega, \mathcal{F}, P)$, since only the distribution matters. Therefore, statements like $X_n \xrightarrow[n\to\infty]{d} \mu$ make sense, too. By Lemma B.5, the weak limit of a sequence of probability measures is unique if it exists. On the other hand, the weak limit of a sequence of random variables is, of course, by no means unique: any random variable with the limiting distribution is a limit.

Lemma C.3. *On the level of distribution functions, weak convergence of probability measures μ_n on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ with c.d.f. F_n to a probability measure μ with c.d.f. F is equivalent to* weak convergence of the distribution functions

$$\forall x \in S(F) : \lim_{n\to\infty} F_n(x) = F(x),$$

where $S(F)$ is the set of points of continuity of F.

Proof. During the proof we say that an interval I with endpoints a and b is an *interval of continuity* of F iff $a, b \in S(F)$. We first prove that weak convergence implies weak convergence of the distribution functions. Fix $\epsilon > 0$. Let I be an interval of continuity of F and $I_\delta \supset I$ an interval of continuity such that the interval $I_\delta \setminus I$ has length δ and such that the left (right) endpoint of I_δ is strictly smaller (larger) then the left (right) endpoint of I[1] and with $\mu(I_\delta \setminus I) < \epsilon$ – here we use that I is an interval of continuity. Then we can find a continuous function f taking values in $[0, 1]$ such that

$$f(x) = \begin{cases} 1, & x \in I, \\ 0, & x \in I_\delta^c. \end{cases}$$

Then we have for n sufficiently large

$$\mu_n(I) \leq \int f d\mu_n < \int f d\mu + \epsilon \leq \mu(I_\delta) + \epsilon < \mu(I) + 2\epsilon.$$

[1] This is not meant to preclude the possibility to choose intervals $I =]-\infty, a]$.

By constructing a similar interval $I_\delta \subset I$, we can also show the revers inequality, i.e., for n sufficiently large

$$\mu_n(I) > \mu(I) - 2\epsilon.$$

This implies that $|\mu_n(I) - \mu(I)| < 2\epsilon$ for n sufficiently large and any interval of continuity I. In particular, choosing $I =]-\infty, x]$ with $x \in S(F)$ gives the claim.

In the other direction, assume that f is a continuous function bounded by $M > 0$. Let A be a finite interval of continuity for F with $\mu(A) > 1 - \epsilon$. By uniform continuity of f on A we can approximate f on A by a step function g with $|f(x) - g(x)| < \epsilon$ for $x \in A$ and we can choose g such that the intervals, on which g takes constant values, are intervals of continuity of F. Outside of A, we set $g = 0$, implying that $|g(x) - f(x)| \leq M$ for $x \in A^c$. We note that

$$\int g d\mu_n \xrightarrow[n\to\infty]{} \int g d\mu.$$

On the other hand, we have

$$\left| \int f d\mu - \int g d\mu \right| \leq \epsilon\mu(A) + M\epsilon \leq (1 + M)\epsilon,$$

and for n sufficiently large

$$\left| \int f d\mu_n - \int g d\mu_n \right| \leq \epsilon\mu_n(A) + M\mu_n(A^c) \leq \epsilon\mu_n(A) + 2\epsilon M \leq (1 + 2M)\epsilon.$$

Combining the inequalities we have for n sufficiently large that

$$\begin{aligned} \left| \int f d\mu - \int f d\mu_n \right| &\leq \left| \int f d\mu - \int g d\mu \right| + \left| \int g d\mu - \int g d\mu_n \right| + \left| \int g d\mu_n - \int f d\mu_n \right| \\ &< 3(M + 1)\epsilon, \end{aligned}$$

implying convergence of $\int f d\mu_n$ to $\int f d\mu$. □

Remark C.4. Weak convergence topologizes the set $\mathcal{M}_1(\mathbb{R}^d)$ of probability measures on $\mathcal{B}(\mathbb{R}^d)$. In fact, one can show that the *weak topology* – which is a weak* topology in the sense of functional analysis – can be metricized by Prokhorov's [German: Prohorov?] metric defined by $d_P(\mu, \nu) := \max\left(d'_P(\mu, \nu), d'_P(\nu, \mu)\right)$ with

$$d'_P(\mu, \nu) := \inf\left\{ \epsilon > 0 \;\middle|\; \forall A \in \mathcal{B}(\mathbb{R}^d) :\ \mu(A) \leq \nu\left(B_\epsilon(A)\right) + \epsilon \right\},$$

where B_ϵ denotes the open ball with radius ϵ.

Lemma C.5 (Slutzky)**.** *Let (X_n) and (Y_n) be sequences of $\mathbb{R}^d$-valued random variables such that $X_n - Y_n \xrightarrow[n\to\infty]{} 0$ in probability and such that $Y_n \xrightarrow[n\to\infty]{d} Y$ for some random variable Y. Then we have $X_n \xrightarrow[n\to\infty]{d} Y$ as well. In particular, convergence in probability implies convergence in distribution.*

Proof. For simplicity, we consider the case $d = 1$ only.[2] Let F_n and G_n denote the distribution functions of X_n and Y_n, respectively. We want to show that $F_n(x) \to G(x)$

[2]The extension to the multi-dimensional case can later be justified by the Cramér-Wold trick, see Lemma C.10.

for $x \in S(G)$, where G denotes the distribution function of Y. Notice that for any $\epsilon > 0$ and $x \in \mathbb{R}$

$$\{ X_n \leq x \} \subset \{ Y_n \leq x + \epsilon \} \cup \{ X_n - Y_n \leq -\epsilon \} \subset \{ Y_n \leq x + \epsilon \} \cup \{ |X_n - Y_n| \geq \epsilon \}.$$

Thus,

$$F_n(x) \leq G_n(x + \epsilon) + P(|X_n - Y_n| \geq \epsilon). \tag{C.2}$$

Now assume that $x \in S(G)$, $x + \epsilon \in S(G)$. For n large enough we have

$$F_n(x) \leq G(x + \epsilon) + 2\epsilon.$$

Since $x \in S(G)$, for any $\delta > 0$ and any n large enough, this implies

$$F_n(x) \leq G(x) + \delta.$$

By turning around the roles of F_n and G_n in (C.2) and changing $x \to x - \epsilon$, we can also get for n large enough

$$F_n(x) \geq G(x - \epsilon) - \epsilon,$$

implying finally that $F_n(x) \to G(x)$, $x \in S(G)$.

For the second part assume that $X_n \to X$ in probability and set $Y_n \equiv X$. Then $X_n \to X$ in distribution by the first part. □

Corollary C.6. *Given a sequence of $\mathbb{R}^d$-valued random variables (X_n) and an $\mathbb{R}^d$-valued random variable X. Assume either of the following three conditions:*

a) $X_n \to X$ almost surely;

b) $X_n \to X$ in $L^p(\Omega, \mathcal{F}, P)$ for some $p \geq 1$;

c) $X_n \to X$ in probability.

Then $X_n \xrightarrow{d} X$.

Proof. We already know by Lemma C.1 that a) and b) imply c). By Lemma C.5, c) implies convergence in distribution. □

Remark C.7. In fact, Corollary C.6 c) can be partially inverted: if we have a sequence of random variables $X_n \xrightarrow[n\to\infty]{d} X_0$, then one can find a probability space $(\Omega', \mathcal{F}', P')$ and random variables $X'_n : \Omega' \to \mathbb{R}^d$ with $P_{X_n} = P'_{X'_n}$, $n \in \mathbb{N}_0$, such that $X'_n \to X'_0$ in (P') probability.

As a probabilistic property, weak convergence should be reflected by properties of the characteristic functions.

Theorem C.8 (Lévy's continuity theorem)**.** *Assume we are given a sequence μ_n of probability measures on $\mathbb{R}^d$.*

(i) If there is a probability measure μ with $\mu_n \xrightarrow[n\to\infty]{d} \mu$, then the sequence $\widehat{\mu_n}$ converges (uniformly on compact sets) to $\widehat{\mu}$.

(ii) Conversely, assume that $\widehat{\mu_n} \xrightarrow[n\to\infty]{} f$ *pointwise for some function* $f : \mathbb{R}^d \to \mathbb{C}$ *(partially) continuous*[3] *at* 0*. Then there is a probability measure* μ *on* $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ *with* $f = \widehat{\mu}$ *and* $\mu_n \xrightarrow[n\to\infty]{d} \mu$*.*

We skip the proof of this important theorem, since it relies on non-trivial results from functional analysis regarding relative sequential compactness in the metric space of probability measures.

Theorem C.9 (Central limit theorem)**.** *Let* X_n *be a sequence of i.i.d. random variables taking values in* $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ *and assume that* $\mu := E[X_1] \in \mathbb{R}^d$ *and* $\Sigma := \operatorname{cov}[X_1] \in \mathbb{R}^{d\times d}$ *exist. Then*

$$\frac{\sum_{i=1}^n (X_i - \mu)}{\sqrt{n}} \xrightarrow[n\to\infty]{d} \mathcal{N}(0, \Sigma).$$

In particular, in the one-dimensional case, we can reformulate the result as

$$\frac{\sum_{i=1}^n (X_i - \mu)}{\sqrt{n}\sigma} \xrightarrow[n\to\infty]{d} \mathcal{N}(0, 1),$$

provided that $\sigma^2 := \operatorname{var}[X_1] > 0$*.*

Proof. We first give the proof in dimension $d = 1$ for $\sigma > 0$ – otherwise, there is nothing to prove. For ease of notation, we assume that $\mu = 0$. Since X_1 is square integrable, Theorem B.8 implies that

$$\varphi_{X_1}(u) = 1 - \frac{\sigma^2}{2}u^2 + o(u^2).$$

By Theorem 2.25 (ii) the characteristic function of the scaled partial sum $S_n^* := \frac{1}{\sqrt{n}\sigma}\sum_{l=1}^n X_l$ satisfies

$$\varphi_{S_n^*}(u) = \prod_{l=1}^n \varphi_{\frac{X_l}{\sqrt{n}\sigma}}(u) = \left(\varphi_{X_1}\left(\frac{u}{\sqrt{n}\sigma}\right)\right)^n = \left(1 - \frac{u^2}{2n} + o(n^{-1})\right)^n \xrightarrow[n\to\infty]{} e^{-u^2/2}.$$

Since $e^{-u^2/2}$ is the characteristic function of the standard normal distribution, Theorem C.8 gives the result. □

For the proof for the multi-dimensional case, we are going to use the following auxiliary result.

Lemma C.10 (Cramér-Wold)**.** *Given a sequence* X_n *of random variables in* $\mathbb{R}^d$*. Then the sequence converges weakly to some random variable* X *if and only if for every* $u \in \mathbb{R}^d$ *the sequence of one-dimensional random variables* $\langle u, X_n\rangle$ *converges to some random variable* X^u*. In that case,* $\langle u, X\rangle$ *has the same distribution as* X^u*.*

Proof. We only need to prove that convergence of $\langle u, X_n\rangle$ in distribution for every $u \in \mathbb{R}^d$ implies weak convergence of the sequence X_n. By Lévy's continuity theorem, we know that for every $u \in \mathbb{R}^d$ the sequence of functions

$$\varphi_{\langle u, X_n\rangle}(t) = E[e^{i\langle tu, X_n\rangle}] = \varphi_{X_n}(tu)$$

[3] This means that the functions $x_i \mapsto f(x)$ is continuous for every component i.

converges to a continuous function $f_u(t)$. Clearly, there is a function $f : \mathbb{R}^d \to \mathbb{C}$ such that $f_u(t) = f(tu)$. Now take $1 \le l \le d$ and $u = e_l$. Again, by Theorem C.8 we have

$$1 = f(0) = f_{e_l}(0) = \lim_{t\to 0} f_{e_l}(t) = \lim_{t\to 0} f(te_l),$$

implying partial continuity of f. Thus, we can appeal to Lévy's continuity theorem for the last time and get weak convergence of the sequence X_n to a random variable X with characteristic function f.

Note that this, in turn, implies weak convergence of $\langle u\,, X_n\rangle$ to $\langle u\,, X\rangle$ because of continuity of $x \mapsto \langle u\,, x\rangle$. □

Proof of the multi-dimensional part of Theorem C.9. Now we assume that X_n are d-dimensional (again with $\mu = 0$). By the one-dimensional version, we know that

$$\frac{\sum_{i=1}^n \langle u\,, X_n\rangle}{\sqrt{n}} \xrightarrow[n\to\infty]{d} \mathcal{N}(0, \sigma_u^2),$$

where $\sigma_u^2 = \operatorname{var}[\langle u\,, X_1\rangle] = u^T \operatorname{cov}[X_1]u$ for every $u \in \mathbb{R}^d$. By Lemma C.10, this implies weak convergence of $\frac{1}{\sqrt{n}} \sum_{i=1}^n X_i$ to some random variable X such that

$$\forall u \in \mathbb{R}^d : \ \langle u\,, X\rangle \sim \mathcal{N}(0, u^T \Sigma u).$$

In particular, we obtain

$$\varphi_X(u) = E[e^{i\langle u\,, X\rangle}] = e^{-\frac{1}{2}u^T\Sigma u},$$

implying that $X \sim \mathcal{N}(0, \Sigma)$ by Theorem B.2. □

Remark C.11. The *Berry-Esseen theorem* gives a non-asymptotic quantitative bound for the speed of convergence in the central limit theorem. Given i.i.d. real random variables X_n with $E[X_1] = 0$ (w.l.o.g.) and $\operatorname{var}[X_1] = \sigma^2 > 0$ such that $\gamma := E\left[|X_1|^3\right] < \infty$. Once again, let $S_n^* := (X_1 + \cdots + X_n)/(\sqrt{n}\sigma)$. There is a universal constant C (i.e., a constant independent of the sequence X_n and the parameters) such that for all $n \in \mathbb{N}$

$$\sup_{x\in\mathbb{R}} \left|P\left(S_n^* \le x\right) - \Phi(x)\right| \le \frac{C\gamma}{\sigma^3\sqrt{n}},$$

where Φ denotes the c.d.f. of the standard normal distribution. The constant C is not known, but one has an upper bound $C \le 0.4785$.

Remark C.12. The central limit theorem can be generalized to non-identically distributed independent families of (real) random variables (X_n). In this case, we say that the central limit theorem holds if

$$\frac{\sum_{j=1}^n (X_j - E[X_j])}{\sqrt{\sum_{j=1}^n \operatorname{var}[X_j]}} \xrightarrow[n\to\infty]{d} \mathcal{N}(0, 1).$$

It turns out that the relevant sufficient condition for the central limit theorem is $\lim_{n\to\infty} L_n(\epsilon) = 0$ for every $\epsilon > 0$, where

$$L_n(\epsilon) := \frac{1}{s_n^2} \sum_{j=1}^n \int_{\{\,|x-E[X_j]|\ge \epsilon s_n\,\}} (x - E[X_j])^2 P_{X_j}(dx) \tag{C.3}$$

with $s_n^2 := \sum_{j=1}^n \mathrm{var}[X_n]$. This condition is known as *Lindeberg's condition*. Moreover, Lindeberg's condition is rather sharp. Indeed, if the central limit theorem holds and the random variables satisfy *Feller's condition*, i.e.,

$$\lim_{n\to\infty} \left(\max_{j=1,\dots,n} \frac{\mathrm{var}[X_j]}{s_n} \right) = 0,$$

then they satisfy Lindeberg's condition.