Measure Theoretic Probability

P.J.C. Spreij

this version: October 26, 2023

Preface

In these notes we explain the measure theoretic foundations of modern probability. The notes are used during a course that had as one of its principal aims a swift introduction to measure theory as far as it is needed in modern probability, e.g. to define concepts as conditional expectation and to prove limit theorems for martingales.

Everyone with a basic notion of mathematics and probability would understand what is meant by $f(x)$ and $\mathbb{P}(A)$. In the former case we have the value of some function f evaluated at its argument. In the second case, one recognizes the probability of an event A. Look at the notations, they are quite similar and this suggests that also $\mathbb{P}$ is a function, defined on some domain to which A belongs. This is indeed the point of view that we follow. We will see that $\mathbb{P}$ is a function -a special case of a *measure*- on a collection of sets, that satisfies certain properties, a *σ-algebra*. In general, a σ-algebra Σ will be defined as a suitable collection of subsets of a given set S. A measure μ will then be a map on Σ, satisfying some defining properties. This gives rise to considering a triple, to be called a measure space, (S, Σ, μ). We will develop probability theory in the context of measure spaces and because of tradition and some distinguished features, we will write $(\Omega, \mathcal{F}, \mathbb{P})$ for a *probability space* instead of (S, Σ, μ). Given a measure space we will develop in a rather abstract sense *integrals* of functions defined on S. In a probabilistic context, these integrals have the meaning of *expectations*. The general setup provides us with two big advantages. In the definition of expectations, we don't have to distinguish anymore between random variables having a *discrete distribution* and those who have what is called a *density*. In the first case, expectations are usually computed as sums, whereas in the latter case, Riemann integrals are the tools. We will see that these are special cases of the more general notion of *Lebesgue integral*. Another advantage is the availability of *convergence theorems*. In analytic terms, we will see that integrals of functions converge to the integral of a *limit* function, given appropriate conditions and an appropriate concept of convergence. In a probabilistic context, this translates to convergence of expectations of random variables. We will see many instances, where the foundations of the theory can be fruitfully applied to fundamental issues in probability theory. These lecture notes are the result of teaching the course *Measure Theoretic Probability* for a number of years.

To a large extent this course was initially based on the book *Probability with Martingales* by D. Williams, but also other texts have been used. In particular we consulted *An Introduction to Probability Theory and Its Applications, Vol. 2* by W. Feller, *Convergence of Stochastic Processes* by D. Pollard, *Real and Complex Analysis* by W. Rudin, *Real Analysis and Probability* by R.M. Dudley, *Foundations of Modern Probability* by O. Kallenberg and *Essentials of stochastic finance* by A.N. Shiryaev.

These lecture notes have first been used in Fall 2008. Among the students who then took the course was Ferdinand Rolwes, who corrected (too) many typos and other annoying errors. Later, Delyan Kalchev, Jan Rozendaal, Arjun Sudan, Willem van Zuijlen, Hailong Bao and Johan du Plessis corrected quite some remaining errors. I am grateful to them all.

Amsterdam, May 2014
Peter Spreij

Contents

1 σ-algebras and measures

In this chapter we lay down the measure theoretic foundations of probability theory. We start with some general notions and show how these are instrumental in a probabilistic environment.

1.1 σ-algebras

Definition 1.1 Let S be a non-empty set. A collection $\Sigma_0 \subset 2^S$ is called an *algebra* (on S) if

(i) $S \in \Sigma_0$,

(ii) $E \in \Sigma_0 \Rightarrow E^c \in \Sigma_0$,

(iii) $E, F \in \Sigma_0 \Rightarrow E \cup F \in \Sigma_0$.

Notice that always $\emptyset$ belongs to an algebra, since $\emptyset = S^c$. Of course property (iii) extends to finite unions by induction. Moreover, in an algebra we also have $E, F \in \Sigma_0 \Rightarrow E \cap F \in \Sigma_0$, since $E \cap F = (E^c \cup F^c)^c$. Furthermore $E \setminus F = E \cap F^c \in \Sigma_0$.

Definition 1.2 Let S be a non-empty set. A collection $\Sigma \subset 2^S$ is called a *σ-algebra* (on S) if it is an algebra and $\bigcup_{n=1}^{\infty} E_n \in \Sigma$ as soon as $E_n \in \Sigma$ $(n = 1, 2 \ldots)$.

If Σ is a σ-algebra on S, then (S, Σ) is called a *measurable space* and the elements of Σ are called *measurable* sets. We shall 'measure' them in the next section.

If $\mathcal{C}$ is any collection of subsets of S, then by $\sigma(\mathcal{C})$ we denote the smallest σ-algebra containing $\mathcal{C}$. This means that $\sigma(\mathcal{C})$ is the intersection of all σ-algebras that contain $\mathcal{C}$ (see Exercise 1.1). If $\Sigma = \sigma(\mathcal{C})$, we say that $\mathcal{C}$ generates Σ. The union of two σ-algebras Σ_1 and Σ_2 on a set S is usually not a σ-algebra. We write $\Sigma_1 \vee \Sigma_2$ for $\sigma(\Sigma_1 \cup \Sigma_2)$.

One of the most relevant σ-algebras of this course is $\mathcal{B} = \mathcal{B}(\mathbb{R})$, the Borel sets of $\mathbb{R}$. Let $\mathcal{O}$ be the collection of all open subsets of $\mathbb{R}$ with respect to the usual topology (in which all intervals (a, b) are open). Then $\mathcal{B} := \sigma(\mathcal{O})$. Of course, one similarly defines the Borel sets of $\mathbb{R}^d$, and in general, for a topological space $(S, \mathcal{O})$, one defines the Borel-sets as $\sigma(\mathcal{O})$. Borel sets can in principle be rather 'wild', but it helps to understand them a little better, once we know that they are generated by simple sets.

Proposition 1.3 *Let $\mathcal{I} = \{(-\infty, x] : x \in \mathbb{R}\}$. Then $\sigma(\mathcal{I}) = \mathcal{B}$.*

Proof We prove the two obvious inclusions, starting with $\sigma(\mathcal{I}) \subset \mathcal{B}$. Since $(-\infty, x] = \cap_n (-\infty, x + \frac{1}{n}) \in \mathcal{B}$, we have $\mathcal{I} \subset \mathcal{B}$ and then also $\sigma(\mathcal{I}) \subset \mathcal{B}$, since $\sigma(\mathcal{I})$ is the smallest σ-algebra that contains $\mathcal{I}$. (Below we will use this kind of arguments repeatedly).

For the proof of the reverse inclusion we proceed in three steps. First we observe that $(-\infty, x) = \cup_n (-\infty, x - \frac{1}{n}] \in \sigma(\mathcal{I})$. Knowing this, we conclude

that $(a,b) = (-\infty,b) \setminus (-\infty,a] \in \sigma(\mathcal{I})$. Let then G be an arbitrary open set. Since G is open, for every $x \in G$ there exists a rational $\varepsilon_x > 0$ such that $(x - 2\varepsilon_x, x + 2\varepsilon_x) \subset G$. Consider $(x - \varepsilon_x, x + \varepsilon_x)$ and choose a rational q_x in this interval, note that $|x - q_x| \leq \varepsilon_x$. It follows that $x \in (q_x - \varepsilon_x, q_x + \varepsilon_x) \subset (x - 2\varepsilon_x, x + 2\varepsilon_x) \subset G$. Hence $G \subset \cup_{x\in G}(q_x - \varepsilon_x, q_x + \varepsilon_x) \subset G$, and so $G = \cup_{x\in G}(q_x - \varepsilon_x, q_x + \varepsilon_x)$. But the union here is in fact a countable union, since there are only countably many q_x and ε_x. (Note that the arguments above can be used for any metric space with a countable dense subset to get that an open G is a countable union of open balls.) It follows that $G \in \sigma(\mathcal{I})$, hence $\mathcal{O} \subset \sigma(\mathcal{I})$, and therefore (recall $\mathcal{B}$ is the smallest σ-algebra containing $\mathcal{O}$) $\mathcal{B} \subset \sigma(\mathcal{I})$. □

An obvious question to ask is whether every subset of $\mathbb{R}$ belongs to $\mathcal{B} = \mathcal{B}(\mathbb{R})$. The answer is no, as we will show that the cardinality of $\mathcal{B}(\mathbb{R})$ is the same as the cardinality of $\mathbb{R}$, from which the negative answer follows.

Let $\mathcal{E}$ be a countable collection of subsets of some set S that contains $\emptyset$. We show that necessarily the cardinality of $\sigma(\mathcal{E})$ is at most $2^{\aleph_0}$. To that end we define collections $\mathcal{E}_\alpha$, for any ordinal number α less than ω, the first uncountable ordinal number. To start, we put $\mathcal{E}_0 = \mathcal{E}$. Let $0 < \alpha < \omega$ ($<$ denotes the usual ordering of the ordinal numbers) and assume that the collections $\mathcal{E}_\beta$ are defined for all $\beta < \alpha$. Put $\mathcal{E}_\alpha^0 = \cup_{\beta<\alpha}\mathcal{E}_\beta$. We define $\mathcal{E}_\alpha$ as the collection of sets that can be written as a countable union $\cup_{n=1}^\infty E_n$, with $E_n \in \mathcal{E}_\alpha^0$ or $E_n^c \in \mathcal{E}_\alpha^0$. Finally, we define $\mathcal{E}_\omega := \cup_{\alpha<\omega}\mathcal{E}_\alpha$.

The first thing we will prove is that $\mathcal{E}_\omega \subset \sigma(\mathcal{E})$. Trivially, $\mathcal{E}_0 \subset \sigma(\mathcal{E})$. Suppose now that $\mathcal{E}_\beta \subset \sigma(\mathcal{E})$ for all $\beta < \alpha$. Then also $\mathcal{E}_\alpha^0 \subset \sigma(\mathcal{E})$, and if $E = \cup_{n=1}^\infty E_n \in \mathcal{E}_\alpha$, it follows that $E \in \sigma(\mathcal{E})$. Hence $\mathcal{E}_\alpha \subset \sigma(\mathcal{E})$ and we conclude that $\mathcal{E}_\omega \subset \sigma(\mathcal{E})$. Note that this also yields $\sigma(\mathcal{E}_\omega) = \sigma(\mathcal{E})$. We will now show that $\mathcal{E}_\omega$ is a σ-algebra, from which it then follows that $\mathcal{E}_\omega = \sigma(\mathcal{E})$.

It is obvious that $\emptyset \in \mathcal{E}_\omega$. Let $E \in \mathcal{E}_\omega$, then there is some $\alpha < \omega$ for which $E \in \mathcal{E}_\alpha$. But $E^c = \cup_{n=1}^\infty E_n$ with $E_n = E^c$, so that $E^c \in \mathcal{E}_\beta$ for all $\beta > \alpha$, and thus $E^c \in \mathcal{E}_\omega$. Similarly, we look at unions. Let $E_n \in \mathcal{E}_\omega$ ($n \in \mathbb{N}$), so there are $\alpha_n < \omega$ such that $E_n \in \mathcal{E}_{\alpha_n}$. Properties of ordinal numbers yield the existence of $\beta < \omega$ such that $\alpha_n \leq \beta$ for all n. It follows that $\cup_{n=1}^\infty E_n \in \mathcal{E}_\beta \subset \mathcal{E}_\omega$. We conclude that $\mathcal{E}_\omega$ is a σ-algebra.

The next thing to show is that the cardinality of $\mathcal{E}_\omega$ is at most $2^{\aleph_0}$. The construction of $\mathcal{E}_1$ from $\mathcal{E}_0 = \mathcal{E}$ implies that the cardinality of $\mathcal{E}_1$ is at most $\aleph_0^{\aleph_0}$, which is equal to $2^{\aleph_0}$. Let $\alpha < \omega$ and assume that the cardinality of $\mathcal{E}_\beta$ is less than or equal to $2^{\aleph_0}$ for all $1 \leq \beta < \alpha$, a property which then also holds for $\mathcal{E}_\alpha^0$. The construction of $\mathcal{E}_\alpha$ from $\mathcal{E}_\alpha^0$ yields, by the same argument as used above for $\mathcal{E}_1$, that also $\mathcal{E}_\alpha$ has cardinality less than or equal to $2^{\aleph_0}$. This shows that the set $I(\omega) := \{\alpha < \omega : \mathcal{E}_\alpha \text{ has cardinality less than or equal to } 2^{\aleph_0}\}$ is what is called an *inductive* set. Since the ordinal numbers with the ordering $<$ is well-ordered, $I(\omega) = \{\alpha : \alpha < \omega\}$. It follows that also the cardinality of $\mathcal{E}_\omega$ is at most equal to $2^{\aleph_0}$.

Turning back to the initial question on the cardinality of $\mathcal{B}(\mathbb{R})$, we apply the above result. Take $\mathcal{E}$ as the set of intervals (a,b) with $a,b \in \mathbb{Q}$ augmented with

the empty set and conclude that $\mathcal{B}(\mathbb{R})$ has cardinality at most equal to $2^{\aleph_0}$.

1.2 Measures

Let Σ_0 be an algebra on a set S, and Σ be a σ-algebra on S. We consider mappings $\mu_0 : \Sigma_0 \to [0, \infty]$ and $\mu : \Sigma \to [0, \infty]$. Note that ∞ is allowed as a possible value.

We call μ_0 *finitely additive* if $\mu_0(\emptyset) = 0$ and if $\mu_0(E \cup F) = \mu_0(E) + \mu_0(F)$ for every pair of disjoint sets E and F in Σ_0. Of course this addition rule then extends to arbitrary finite unions of disjoint sets. The mapping μ_0 is called *σ-additive* or *countably additive*, if $\mu_0(\emptyset) = 0$ and if $\mu_0(\cup_n E_n) = \sum_n \mu_0(E_n)$ for every sequence (E_n) of disjoint sets of Σ_0 whose union is also in Σ_0. σ-additivity is defined similarly for μ, but then we don't have to require that $\cup_n E_n \in \Sigma$. This is true by definition.

Definition 1.4 Let (S, Σ) be a measurable space. A countably additive mapping $\mu : \Sigma \to [0, \infty]$ is called a *measure*. The triple (S, Σ, μ) is called a *measure space*.

Some extra terminology follows. A measure is called finite if $\mu(S) < \infty$. It is called σ-finite, if we can write $S = \cup_n S_n$, where the S_n are measurable sets and $\mu(S_n) < \infty$. If $\mu(S) = 1$, then μ is called a *probability measure*.

Measures are used to 'measure' (measurable) sets in one way or another. Here is a simple example. Let $S = \mathbb{N}$ and $\Sigma = 2^{\mathbb{N}}$ (we often take the power set as the σ-algebra on a countable set). Let τ (we write τ instead of μ for this special case) be the *counting measure*: $\tau(E) = |E|$, the cardinality of E. One easily verifies that τ is a measure, and it is σ-finite, because $\mathbb{N} = \cup_n \{1, \ldots, n\}$.

A very simple measure is the *Dirac measure*. Consider a measurable space (S, Σ) and single out a specific $x_0 \in S$. Define $\delta(E) = \mathbf{1}_E(x_0)$, for $E \in \Sigma$ ($\mathbf{1}_E$ is the indicator function of the set E, $\mathbf{1}_E(x) = 1$ if $x \in E$ and $\mathbf{1}_E(x) = 0$ if $x \notin E$). Check that δ is a measure on Σ.

Another example is *Lebesgue measure*, whose existence is formulated below. It is the most natural candidate for a measure on the Borel sets on the real line.

Theorem 1.5 *There exists a unique measure λ on $(\mathbb{R}, \mathcal{B})$ with the property that for every interval $I = (a, b]$ with $a < b$ it holds that $\lambda(I) = b - a$.*

The proof of this theorem is deferred to later, see Theorem 2.6. For the time being, we take this existence result for granted. One remark is in order. One can show that $\mathcal{B}$ is not the largest σ-algebra for which the measure λ can coherently be defined. On the other hand, on the power set of $\mathbb{R}$ it is impossible to define a measure that coincides with λ on the intervals. We'll come back to this later.

Here are the first elementary properties of a measure.

Proposition 1.6 *Let (S, Σ, μ) be a measure space. Then the following hold true (all the sets below belong to Σ).*

(i) *If* $E \subset F$, *then* $\mu(E) \leq \mu(F)$.

(ii) $\mu(E \cup F) \leq \mu(E) + \mu(F)$.

(iii) $\mu(\cup_{k=1}^{n} E_k) \leq \sum_{k=1}^{n} \mu(E_k)$

If μ *is finite, we also have*

(iv) *If* $E \subset F$, *then* $\mu(F \setminus E) = \mu(F) - \mu(E)$.

(v) $\mu(E \cup F) = \mu(E) + \mu(F) - \mu(E \cap F)$.

Proof The set F can be written as the disjoint union $F = E \cup (F \setminus E)$. Hence $\mu(F) = \mu(E) + \mu(F \setminus E)$. Property (i) now follows and (iv) as well, provided μ is finite. To prove (ii), we note that $E \cup F = E \cup (F \setminus (E \cap F))$, a disjoint union, and that $E \cap F \subset F$. The result follows from (i). Moreover, (v) also follows, if we apply (iv). Finally, (iii) follows from (ii) by induction. □

Measures have certain continuity properties.

Proposition 1.7 *Let* (E_n) *be a sequence in* Σ.

(i) *If the sequence is increasing, with limit* $E = \cup_n E_n$, *then* $\mu(E_n) \uparrow \mu(E)$ *as* $n \to \infty$.

(ii) *If the sequence is decreasing, with limit* $E = \cap_n E_n$ *and if* $\mu(E_n) < \infty$ *from a certain index on, then* $\mu(E_n) \downarrow \mu(E)$ *as* $n \to \infty$.

Proof (i) Define $D_1 = E_1$ and $D_n = E_n \setminus \cup_{k=1}^{n-1} E_k$ for $n \geq 2$. Then the D_n are disjoint, $E_n = \cup_{k=1}^{n} D_k$ for $n \geq 1$ and $E = \cup_{k=1}^{\infty} D_k$. It follows that $\mu(E_n) = \sum_{k=1}^{n} \mu(D_k) \uparrow \sum_{k=1}^{\infty} \mu(D_k) = \mu(E)$.

To prove (ii) we assume without loss of generality that $\mu(E_1) < \infty$. Define $F_n = E_1 \setminus E_n$. Then (F_n) is an increasing sequence with limit $F = E_1 \setminus E$. So (i) applies, yielding $\mu(E_1) - \mu(E_n) \uparrow \mu(E_1) - \mu(E)$. The result follows. □

Corollary 1.8 *Let* (S, Σ, μ) *be a measure space. For an arbitrary sequence* (E_n) *of sets in* Σ, *we have* $\mu(\cup_{n=1}^{\infty} E_n) \leq \sum_{n=1}^{\infty} \mu(E_n)$.

Proof Exercise 1.2. □

Remark 1.9 The finiteness condition in the second assertion of Proposition 1.7 is essential. Consider $\mathbb{N}$ with the counting measure τ. Let $F_n = \{n, n+1, \ldots\}$, then $\cap_n F_n = \emptyset$ and so it has measure zero. But $\tau(F_n) = \infty$ for all n.

1.3 Null sets

Consider a measure space (S, Σ, μ) and let $E \in \Sigma$ be such that $\mu(E) = 0$. If N is a subset of E, then it is fair to suppose that also $\mu(N) = 0$. But this can only be guaranteed if $N \in \Sigma$. Therefore we introduce some new terminology. A set $N \subset S$ is called a *null set* or μ-null set, if there exists $E \in \Sigma$ with $E \supset N$ and $\mu(E) = 0$. The collection of null sets is denoted by $\mathcal{N}$, or $\mathcal{N}_\mu$ since it depends on μ. In Exercise 1.5 you will be asked to show that $\mathcal{N}$ is a σ-algebra and to extend μ to $\bar{\Sigma} = \Sigma \vee \mathcal{N}$. If the extension is called $\bar{\mu}$, then we have a new measure space $(S, \bar{\Sigma}, \bar{\mu})$, which is *complete*, all $\bar{\mu}$-null sets belong to the σ-algebra $\bar{\Sigma}$.

1.4 π- and d-systems

In general it is hard to grab what the elements of a σ-algebra Σ are, but often collections $\mathcal{C}$ such that $\sigma(\mathcal{C}) = \Sigma$ are easier to understand. In 'good situations' properties of Σ can easily be deduced from properties of $\mathcal{C}$. This is often the case when $\mathcal{C}$ is a π-system, to be defined next.

Definition 1.10 A collection $\mathcal{I}$ of subsets of S is called a π-system, if $I_1, I_2 \in \mathcal{I}$ implies $I_1 \cap I_2 \in \mathcal{I}$.

It follows that a π-system is closed under finite intersections. In a σ-algebra, all familiar set operations are allowed, at most countably many. We will see that it is possible to disentangle the defining properties of a σ-algebra into taking finite intersections and the defining properties of a d-system. This is the content of Proposition 1.12 below.

Definition 1.11 A collection $\mathcal{D}$ of subsets of S is called a d-system, if the following hold.

(i) $S \in \mathcal{D}$.

(ii) If $E, F \in \mathcal{D}$ such that $E \subset F$, then $F \setminus E \in \mathcal{D}$.

(iii) If $E_n \in \mathcal{D}$ for $n \in \mathbb{N}$, and $E_n \subset E_{n+1}$ for all n, then $\cup_n E_n \in \mathcal{D}$.

Proposition 1.12 *Σ is a σ-algebra iff it is a π-system and a d-system.*

Proof Let Σ be a π-system and a d-system. We check the defining conditions of a σ-algebra. (i) Since Σ is a d-system, $S \in \Sigma$. (ii) Complements of sets in Σ are in Σ as well, again because Σ is a d-system. (iii) If $E, F \in \Sigma$, then $E \cup F = (E^c \cap F^c)^c \in \Sigma$, because we have just shown that complements remain in Σ and because Σ is a π-system. Then Σ is also closed under finite unions. Let $E_1, E_2, \ldots$ be a sequence in Σ. We have just showed that the sets $F_n = \cup_{i=1}^n E_i \in \Sigma$. But since the F_n form an increasing sequence, also their union is in Σ, because Σ is a d-system. But $\cup_n F_n = \cup_n E_n$. This proves that Σ is a σ-algebra. Of course the other implication is trivial. $\square$

If $\mathcal{C}$ is a collection of subsets of S, then by $d(\mathcal{C})$ we denote the smallest d-system that contains $\mathcal{C}$. Note that it always holds that $d(\mathcal{C}) \subset \sigma(\mathcal{C})$. In one important case we have equality. This is known as Dynkin's lemma.

Lemma 1.13 *Let $\mathcal{I}$ be a π-system. Then $d(\mathcal{I}) = \sigma(\mathcal{I})$.*

Proof Suppose that we would know that $d(\mathcal{I})$ is a π-system as well. Then Proposition 1.12 yields that $d(\mathcal{I})$ is a σ-algebra, and so it contains $\sigma(\mathcal{I})$. Since the reverse inclusion is always true, we have equality. Therefore we will prove that indeed $d(\mathcal{I})$ is a π-system.

Step 1. Put $\mathcal{D}_1 = \{B \in d(\mathcal{I}) : B \cap C \in d(\mathcal{I}), \forall C \in \mathcal{I}\}$. We claim that $\mathcal{D}_1$ is a d-system. Given that this holds and because, obviously, $\mathcal{I} \subset \mathcal{D}_1$, also $d(\mathcal{I}) \subset \mathcal{D}_1$. Since $\mathcal{D}_1$ is defined as a subset of $d(\mathcal{I})$, we conclude that these

two collections are the same. We now show that the claim holds. Evidently $S \in \mathcal{D}_1$. Let $B_1, B_2 \in \mathcal{D}_1$ with $B_1 \subset B_2$ and $C \in \mathcal{I}$. Write $(B_2 \setminus B_1) \cap C$ as $(B_2 \cap C) \setminus (B_1 \cap C)$. The last two intersections belong to $d(\mathcal{I})$ by definition of $\mathcal{D}_1$ and so does their difference, since $d(\mathcal{I})$ is a d-system. For $B_n \uparrow B$, $B_n \in \mathcal{D}_1$ and $C \in \mathcal{I}$ we have $(B_n \cap C) \in d(\mathcal{I})$ which then converges to $B \cap C \in d(\mathcal{I})$. So $B \in \mathcal{D}_1$.

Step 2. Put $\mathcal{D}_2 = \{C \in d(\mathcal{I}) : B \cap C \in d(\mathcal{I}), \forall B \in d(\mathcal{I})\}$. We claim, again, (and *you check*) that $\mathcal{D}_2$ is a d-system. The key observation is that $\mathcal{I} \subset \mathcal{D}_2$. Indeed, take $C \in \mathcal{I}$ and $B \in d(\mathcal{I})$. The latter collection is nothing else but $\mathcal{D}_1$, according to step 1. But then $B \cap C \in d(\mathcal{I})$, which means that $C \in \mathcal{D}_2$. It now follows that $d(\mathcal{I}) \subset \mathcal{D}_2$, but then we must have equality, because $\mathcal{D}_2$ is defined as a subset of $d(\mathcal{I})$. The equality $\mathcal{D}_2 = d(\mathcal{I})$ and the definition of $\mathcal{D}_2$ together imply that $d(\mathcal{I})$ is a π-system, as desired. □

Sometimes another version of Lemma 1.13 is useful.

Corollary 1.14 *The assertion of Lemma 1.13 is equivalent to the following statement. Let $\mathcal{I}$ be a π-system and $\mathcal{D}$ be a d-system. If $\mathcal{I} \subset \mathcal{D}$, then $\sigma(\mathcal{I}) \subset \mathcal{D}$.*

Proof Suppose that $\mathcal{I} \subset \mathcal{D}$. Then $d(\mathcal{I}) \subset \mathcal{D}$. But $d(\mathcal{I}) = \sigma(\mathcal{I})$, according to Lemma 1.13. Conversely, let $\mathcal{I}$ be a π-system. Then $\mathcal{I} \subset d(\mathcal{I})$. By hypothesis, one also has $\sigma(\mathcal{I}) \subset d(\mathcal{I})$, and the latter is always a subset of $\sigma(\mathcal{I})$. □

All these efforts lead to the following very useful theorem. It states that any finite measure on Σ is characterized by its action on a rich enough π-system. We will meet many occasions where this theorem is used.

Theorem 1.15 *Let $\mathcal{I}$ be a π-system and $\Sigma = \sigma(\mathcal{I})$. Let μ_1 and μ_2 be finite measures on Σ with the properties that $\mu_1(S) = \mu_2(S)$ and that μ_1 and μ_2 coincide on $\mathcal{I}$. Then $\mu_1 = \mu_2$ (on Σ).*

Proof The whole idea behind the proof is to find a good d-system that contains $\mathcal{I}$. The following set is a reasonable candidate. Put $\mathcal{D} = \{E \in \Sigma : \mu_1(E) = \mu_2(E)\}$. The inclusions $\mathcal{I} \subset \mathcal{D} \subset \Sigma$ are obvious. If we can show that $\mathcal{D}$ is a d-system, then Corollary 1.14 gives the result. The fact that $\mathcal{D}$ is a d-system is straightforward to check, we present only one verification. Let $E, F \in \mathcal{D}$ such that $E \subset F$. Then (use Proposition 1.6 (iv)) $\mu_1(F \setminus E) = \mu_1(F) - \mu_1(E) = \mu_2(F) - \mu_2(E) = \mu_2(F \setminus E)$ and so $F \setminus E \in \mathcal{D}$. □

Remark 1.16 In the above proof we have used the fact that μ_1 and μ_2 are finite. If this condition is violated, then the assertion of the theorem is not valid in general. Here is a counterexample. Take $\mathbb{N}$ with the counting measure $\mu_1 = \tau$ and let $\mu_2 = 2\tau$. A π-system that generates $2^{\mathbb{N}}$ is given by the sets $G_n = \{n, n+1, \ldots\}$ $(n \in \mathbb{N})$.

1.5 Probability language

In Probability Theory, one usually writes $(\Omega, \mathcal{F}, \mathbb{P})$ instead of (S, Σ, μ), and one then speakes of a *probability space*. On one hand this is merely change of notation and language. We still have that Ω is a set, $\mathcal{F}$ a σ-algebra on it, and $\mathbb{P}$ a measure, but in this case, $\mathbb{P}$ is a *probability* measure (often also simply called probability), $\mathbb{P}(\Omega) = 1$. In probabilistic language, Ω is often called the set of outcomes and elements of $\mathcal{F}$ are called *events*. So by definition, an event is a measurable subset of the set of all outcomes.

A probability space $(\Omega, \mathcal{F}, \mathbb{P})$ can be seen as a mathematical model of a random experiment. Consider for example the experiment consisting of tossing two coins. Each coin has individual outcomes 0 and 1. The set Ω can then be written as $\{00, 01, 10, 11\}$, where the notation should be obvious. In this case, we take $\mathcal{F} = 2^\Omega$ and a choice of $\mathbb{P}$ could be such that $\mathbb{P}$ assigns probability $\frac{1}{4}$ to all singletons. Of course, from a purely mathematical point of view, other possibilities for $\mathbb{P}$ are conceivable as well.

A more interesting example is obtained by considering an infinite sequence of coin tosses. In this case one should take $\Omega = \{0, 1\}^{\mathbb{N}}$ and an element $\omega \in \Omega$ is then an infinite sequence $(\omega_1, \omega_2, \ldots)$ with $\omega_n \in \{0, 1\}$. It turns out that one cannot take the power set of Ω as a σ-algebra, if one wants to have a nontrivial probability measure defined on it. As a matter of fact, this holds for the same reason that one cannot take the power set on $(0, 1]$ to have a consistent notion of Lebesgue measure. This has everything to do with the fact that one can set up a bijective correspondence between $(0, 1)$ and $\{0, 1\}^{\mathbb{N}}$. Nevertheless, there is a good candidate for a σ-algebra $\mathcal{F}$ on Ω. One would like to have that sets like 'the 12-th outcome is 1' are events. Let $\mathcal{C}$ be the collection of all such sets, $\mathcal{C} = \{\{\omega \in \Omega : \omega_n = s\}, n \in \mathbb{N}, s \in \{0, 1\}\}$. We take $\mathcal{F} = \sigma(\mathcal{C})$ and all sets $\{\omega \in \Omega : \omega_n = s\}$ are then events. One can show that there indeed exists a probability measure $\mathbb{P}$ on this $\mathcal{F}$ with the nice property that for instance the set $\{\omega \in \Omega : \omega_1 = \omega_2 = 1\}$ (in the previous example it would have been denoted by $\{11\}$) has probability $\frac{1}{4}$.

Having the interpretation of $\mathcal{F}$ as a collection of events, we now introduce two special events. Consider a sequence of events $E_1, E_2, \ldots$ and define

$$\limsup E_n := \bigcap_{m=1}^{\infty} \bigcup_{n=m}^{\infty} E_n$$
$$\liminf E_n := \bigcup_{m=1}^{\infty} \bigcap_{n=m}^{\infty} E_n.$$

Note that the sets $F_m = \cap_{n \geq m} E_n$ form an increasing sequence and the sets $D_m = \cup_{n \geq m} E_n$ form a decreasing sequence. Clearly, $\mathcal{F}$ is closed under taking limsup and liminf. The terminology is explained by (i) of Exercise 1.4. In probabilistic terms, $\limsup E_n$ is described as the event that the E_n occur *infinitely often*, abbreviated by E_n i.o. Likewise, $\liminf E_n$ is the event that the E_n occur *eventually*. The former interpretation follows by observing that $\omega \in \limsup E_n$

iff for all m, there exists $n \geq m$ such that $\omega \in E_n$. In other words, a particular outcome ω belongs to $\limsup E_n$ iff it belongs to some (infinite) subsequence of (E_n).

The terminology to call $\bigcup_{m=1}^{\infty} \bigcap_{n=m}^{\infty} E_n$ the lim inf of the sequence is justified in Exercise 1.4. In this exercise, *indicator functions* of events are used, of which we here recall the definition. If E is an event, then the function $\mathbf{1}_E$ is defined by $\mathbf{1}_E(\omega) = 1$ if $\omega \in E$ and $\mathbf{1}_E(\omega) = 0$ if $\omega \notin E$.

1.6 Exercises

1.1 Prove the following statements.

(a) The intersection of an arbitrary family of d-systems is again a d-system.
(b) The intersection of an arbitrary family of σ-algebras is again a σ-algebra.
(c) If $\mathcal{C}_1$ and $\mathcal{C}_2$ are collections of subsets of Ω with $\mathcal{C}_1 \subset \mathcal{C}_2$, then $d(\mathcal{C}_1) \subset d(\mathcal{C}_2)$.

1.2 Prove Corollary 1.8.

1.3 Prove the claim that $\mathcal{D}_2$ in the proof of Lemma 1.13 forms a d-system.

1.4 Consider a measure space (S, Σ, μ). Let (E_n) be a sequence in Σ.

(a) Show that $\mathbf{1}_{\liminf E_n} = \liminf \mathbf{1}_{E_n}$.
(b) Show that $\mu(\liminf E_n) \leq \liminf \mu(E_n)$. (Use Proposition 1.7.)
(c) Show also that $\mu(\limsup E_n) \geq \limsup \mu(E_n)$, provided that μ is finite.

1.5 Let (S, Σ, μ) be a measure space. Call a subset N of S a (μ, Σ)-null set if there exists a set $N' \in \Sigma$ with $N \subset N'$ and $\mu(N') = 0$. Denote by $\mathcal{N}$ the collection of all (μ, Σ)-null sets. Let Σ^* be the collection of subsets E of S for which there exist $F, G \in \Sigma$ such that $F \subset E \subset G$ and $\mu(G \setminus F) = 0$. For $E \in \Sigma^*$ and F, G as above we define $\mu^*(E) = \mu(F)$.

(a) Show that Σ^* is a σ-algebra and that $\Sigma^* = \Sigma \vee \mathcal{N} (= \sigma(\mathcal{N} \cup \Sigma))$.
(b) Show that μ^* restricted to Σ coincides with μ and that $\mu^*(E)$ doesn't depend on the specific choice of F in its definition.
(c) Show that the collection of (μ^*, Σ^*)-null sets is $\mathcal{N}$.

1.6 Let $\mathcal{G}$ and $\mathcal{H}$ be two σ-algebras on Ω. Let $\mathcal{C} = \{G \cap H : G \in \mathcal{G},\, H \in \mathcal{H}\}$. Show that $\mathcal{C}$ is a π-system and that $\sigma(\mathcal{C}) = \sigma(\mathcal{G} \cup \mathcal{H})$.

1.7 Let Ω be a countable set. Let $\mathcal{F} = 2^{\Omega}$ and let $p : \Omega \to [0, 1]$ satisfy $\sum_{\omega \in \Omega} p(\omega) = 1$. Put $\mathbb{P}(A) = \sum_{\omega \in A} p(\omega)$ for $A \in \mathcal{F}$. Show that $\mathbb{P}$ is a probability measure.

1.8 Let Ω be a countable set. Let $\mathcal{A}$ be the collection of $A \subset \Omega$ such that A or its complement has finite cardinality. Show that $\mathcal{A}$ is an algebra. What is $d(\mathcal{A})$?

1.9 Show that a finitely additive map $\mu : \Sigma_0 \to [0, \infty]$ is countably additive if $\mu(H_n) \to 0$ for every decreasing sequence of sets $H_n \in \Sigma_0$ with $\bigcap_n H_n = \emptyset$. If μ is countably additive, do we necessarily have $\mu(H_n) \to 0$ for every decreasing sequence of sets $H_n \in \Sigma_0$ with $\bigcap_n H_n = \emptyset$?

1.10 Consider the collection Σ_0 of subsets of $\mathbb{R}$ that can be written as a *finite* union of *disjoint* intervals of type $(a, b]$ with $-\infty \leq a \leq b < \infty$ or (a, ∞). Show that Σ_0 is an algebra and that $\sigma(\Sigma_0) = \mathcal{B}(\mathbb{R})$.

2 Existence of Lebesgue measure

In this chapter we construct the Lebesgue measure on the Borel sets of $\mathbb{R}$. To that end we need the concept of outer measure. Somewhat hidden in the proof of the construction is the extension of a countably additive function on an algebra to a measure on a σ-algebra. There are different versions of extension theorems, originally developed by Carathéodory. Although of crucial importance in measure theory, we will confine our treatment of extension theorems mainly aimed at the construction of Lebesgue measure on $(\mathbb{R}, \mathcal{B})$. However, see also the end of this section.

2.1 Outer measure and construction

Definition 2.1 Let S be a set. An *outer measure* on S is a mapping $\mu^* : 2^S \to [0, \infty]$ that satisfies

(i) $\mu^*(\emptyset) = 0$,

(ii) μ^* is monotone, i.e. $\mu^*(E) \leq \mu^*(F)$ if $E \subset F$,

(iii) μ^* is subadditive, i.e. $\mu^*(\bigcup_{n=1}^{\infty} E_n) \leq \sum_{n=1}^{\infty} \mu^*(E_n)$, valid for any sequence of sets E_n.

Definition 2.2 Let μ^* be an outer measure on a set S. A set $E \subset S$ is called *μ-measurable* if

$$\mu^*(F) = \mu^*(E \cap F) + \mu^*(E^c \cap F), \forall F \subset S.$$

The class of μ-measurable sets is denoted by Σ_μ.

Theorem 2.3 *Let μ^* be an outer measure on a set S. Then Σ_μ is a σ-algebra and the restricted mapping $\mu : \Sigma_\mu \to [0, \infty]$ of μ^* is a measure on Σ_μ.*

Proof It is obvious that $\emptyset \in \Sigma_\mu$ and that $E^c \in \Sigma_\mu$ as soon as $E \in \Sigma_\mu$. Let $E_1, E_2 \in \Sigma_\mu$ and $F \subset S$. The trivial identity

$$F \cap (E_1 \cap E_2)^c = (F \cap E_1^c) \cup (F \cap (E_1 \cap E_2^c))$$

yields with the subadditivity of μ^*

$$\mu^*(F \cap (E_1 \cap E_2)^c) \leq \mu^*(F \cap E_1^c) + \mu^*(F \cap (E_1 \cap E_2^c)).$$

Add to both sides $\mu^*(F \cap (E_1 \cap E_2))$ and use that $E_1, E_2 \in \Sigma_\mu$ to obtain

$$\mu^*(F \cap (E_1 \cap E_2)) + \mu^*(F \cap (E_1 \cap E_2)^c) \leq \mu^*(F).$$

From subadditivity the reversed version of this equality immediately follows as well, which shows that $E_1 \cap E_2 \in \Sigma_\mu$. We conclude that Σ_μ is an algebra.

Pick disjoint $E_1, E_2 \in \Sigma_\mu$, then $(E_1 \cup E_2) \cap E_1^c = E_2$. If $F \subset S$, then by $E_1 \in \Sigma_\mu$

$$\begin{aligned} \mu^*(F \cap (E_1 \cup E_2)) &= \mu^*(F \cap (E_1 \cup E_2) \cap E_1) + \mu^*(F \cap (E_1 \cup E_2) \cap E_1^c) \\ &= \mu^*(F \cap E_1) + \mu^*(F \cap E_2). \end{aligned}$$

By induction we obtain that for every sequence of disjoint set E_i in Σ_μ it holds that for every $F \subset S$

$$\mu^*(F \cap \bigcup_{i=1}^{n} E) = \sum_{i=1}^{n} \mu^*(F \cap E_i). \tag{2.1}$$

If $E = \bigcup_{i=1}^{\infty} E_i$, it follows from (2.1) and the monotonicity of μ^* that

$$\mu^*(F \cap E) \geq \sum_{i=1}^{\infty} \mu^*(F \cap E_i).$$

Since subadditivity of μ^* immediately yields the reverse inequality, we obtain

$$\mu^*(F \cap E) = \sum_{i=1}^{\infty} \mu^*(F \cap E_i). \tag{2.2}$$

Let $U_n = \bigcup_{i=1}^{n} E_i$ and note that $U_n \in \Sigma_\mu$. We obtain from (2.1) and (2.2) and monotonicity

$$\begin{aligned}
\mu^*(F) &= \mu^*(F \cap U_n) + \mu^*(F \cap U_n^c) \\
&\geq \sum_{i=1}^{n} \mu^*(F \cap E_i) + \mu^*(F \cap E^c) \\
&\to \sum_{i=1}^{\infty} \mu^*(F \cap E_i) + \mu^*(F \cap E^c) \\
&= \mu^*(F \cap E) + \mu^*(F \cap E^c).
\end{aligned}$$

Combined with $\mu^*(F) \leq \mu^*(F \cap E) + \mu^*(F \cap E^c)$, which again is the result of subadditivity, we see that $E \in \Sigma_\mu$. If follows that Σ_μ is a σ-algebra, since every countable union of sets in Σ_μ can be written as a union of disjoint sets in Σ_μ (use that we already know that Σ_μ is an algebra). Finally, take $F = S$ in (2.2) to see that μ^* restricted to Σ_μ is a measure. □

We will use Theorem 2.3 to show the existence of Lebesgue measure on $(\mathbb{R}, \mathcal{B})$. Let E be a subset of $\mathbb{R}$. By $\mathcal{I}(E)$ we denote a cover of E consisting of at most countably many open intervals. For any interval I, we denote by $\lambda_0(I)$ its ordinary length. We now define a function λ^* defined on $2^{\mathbb{R}}$ by putting for every $E \subset \mathbb{R}$

$$\lambda^*(E) = \inf_{\mathcal{I}(E)} \sum_{I_k \in \mathcal{I}(E)} \lambda_0(I_k). \tag{2.3}$$

Lemma 2.4 *The function λ^* defined by (2.3) is an outer measure on $\mathbb{R}$ and satisfies $\lambda^*(I) = \lambda_0(I)$.*

Proof Properties (i) and (ii) of Definition 2.1 are obviously true. We prove subadditivity. Let $E_1, E_2, \ldots$ be arbitrary subsets of $\mathbb{R}$ and $\varepsilon > 0$. By definition of λ^*, there exist covers $\mathcal{I}(E_n)$ of the E_n such that for all n

$$\lambda^*(E_n) \geq \sum_{I \in \mathcal{I}(E_n)} \lambda_0(I) - \varepsilon 2^{-n}. \tag{2.4}$$

Because $\cup_n \mathcal{I}(E_n)$ is a countable open cover of $\cup_n E_n$,

$$\begin{aligned} \lambda^*(\cup_n E_n) &\leq \sum_n \sum_{I \in \mathcal{I}(E_n)} \lambda_0(I) \\ &\leq \sum_n \lambda^*(E_n) + \varepsilon, \end{aligned}$$

in view of (2.4). Subadditivity follows upon letting $\varepsilon \to 0$.

Turning to the next assertion, we observe that $\lambda^*(I) \leq \lambda_0(I)$ is almost immediate (I an arbitrary interval). The reversed inequality is a little harder to prove. Without loss of generality, we may assume that I is compact. Let $\mathcal{I}(I)$ be a cover of I. We aim at proving

$$\lambda_0(I) \leq \sum_{I_k \in \mathcal{I}(I)} \lambda_0(I_k), \text{for every interval } I. \tag{2.5}$$

If this holds, then by taking the infimum on the right hand of (2.5), it follows that $\lambda_0(I) \leq \lambda^*(I)$. To prove (2.5) we proceed as follows. The covering intervals are open. By compactness of I, there exists a finite subcover of I, $\{I_1, \ldots, I_n\}$ say. So, it is sufficient to show (2.5), which we do by induction. If $n = 1$, this is trivial. Assume it is true for covers with at most $n-1$ elements. Assume that $I = [a, b]$. Then b is an element of some $I_k = (a_k, b_k)$. Note that the interval $I \setminus I_k$ (possibly empty) is covered by the remaining intervals, and by hypothesis we have $\lambda_0(I \setminus I_k) \leq \sum_{j \neq k} \lambda_0(I_j)$. But then we deduce $\lambda_0(I) = (b - a_k) + (a_k - a) \leq (b_k - a_k) + (a_k - a) \leq \lambda_0(I_k) + \lambda_0(I \setminus I_k) \leq \sum_j \lambda_0(I_j)$. □

Lemma 2.5 *Any interval $I_a = (-\infty, a]$ ($a \in \mathbb{R}$) is λ-measurable, $I_a \in \Sigma_\lambda$. Hence $\mathcal{B} \subset \Sigma_\lambda$.*

Proof Let $E \subset \mathbb{R}$. Since λ^* is subadditive, it is sufficient to show that $\lambda^*(E) \geq \lambda^*(E \cap I_a) + \lambda^*(E \cap I_a^c)$. Let $\varepsilon > 0$ and choose a cover $\mathcal{I}(E)$ such that $\lambda^*(E) \geq \sum_{I \in \mathcal{I}(E)} \lambda^*(I) - \varepsilon$, which is possible by the definition of λ^* and Lemma 2.4. This lemma also yields $\lambda^*(I) = \lambda^*(I \cap I_a) + \lambda^*(I \cap I_a^c)$. But then we have $\lambda^*(E) \geq \sum_{I \in \mathcal{I}(E)} \lambda^*(I \cap I_a) + \lambda^*(I \cap I_a^c) - \varepsilon$, which is bigger than $\lambda^*(E \cap I_a) + \lambda^*(E \cap I_a^c) - \varepsilon$. Let $\varepsilon \downarrow 0$. □

Putting the previous results together, we obtain existence of the *Lebesgue measure* on $\mathcal{B}$.

Theorem 2.6 *The (restricted) function $\lambda : \mathcal{B} \to [0, \infty]$ is the unique measure on $\mathcal{B}$ that satisfies $\lambda(I) = \lambda_0(I)$.*

Proof By Theorem 2.3 and Lemma 2.4 λ is a measure on Σ_λ and by Lemma 2.5 its restriction to $\mathcal{B}$ is a measure as well. Moreover, Lemma 2.4 states that $\lambda(I) = \lambda_0(I)$. The only thing that remains to be shown is that λ is the unique measure with the latter property. Suppose that also a measure μ enjoys this property. Then, for any $a \in \mathbb{R}$ we have and $n \in \mathbb{N}$, we have that $(-\infty, a] \cap [-n, +n]$ is an interval, hence $\lambda((-\infty, a] \cap [-n, +n]) = \mu((-\infty, a] \cap [-n, +n])$. Since the intervals $(-\infty, a]$ form a π-system that generates $\mathcal{B}$, we also have

$$\lambda(B \cap [-n, +n]) = \mu(B \cap [-n, +n]),$$

for any $B \in \mathcal{B}$ and $n \in \mathbb{N}$. Since λ and μ are measures, we obtain for $n \to \infty$ that $\lambda(B) = \mu(B), \forall B \in \mathcal{B}$. □

The sets in Σ_λ are also called *Lebesgue-measurable* sets. A function $f : \mathbb{R} \to \mathbb{R}$ is called Lebesgue-measurable if the sets $\{f \leq c\}$ are in Σ_λ for all $c \in \mathbb{R}$. The question arises whether all subsets of $\mathbb{R}$ are in Σ_λ. The answer is no, but the Axiom of Choice is needed for this, see Exercise 2.6. Unlike showing that there exist sets that are not Borel-measurable, here a counting argument as in Section 1.1 is useless, since it holds that Σ_λ has the same cardinality as $2^{\mathbb{R}}$. This fact can be seen as follows.

Consider the Cantor set in $[0, 1]$. Let $C_1 = [0, \frac{1}{3}] \cup [\frac{2}{3}, 1]$, obtained from $C_0 = [0, 1]$ be deleting the 'middle third'. From each of the components of C_1 we leave out the 'middle thirds' again, resulting in $C_2 = [0, \frac{1}{9}] \cup [\frac{2}{9}, \frac{1}{3}] \cup [\frac{2}{3}, \frac{7}{9}] \cup [\frac{8}{9}, 1]$, and so on. The obtained sequence of sets C_n is decreasing and its limit $C := \bigcap_{n=1}^{\infty} C_n$ the Cantor set, is well defined. Moreover, we see that $\lambda(C) = 0$. On the other hand, C is uncountable, since every number in it can be described by its ternary expansion $\sum_{k=1}^{\infty} x_k 3^{-k}$, with the $x_k \in \{0, 2\}$. By completeness of $([0, 1], \Sigma_\lambda, \lambda)$, every subset of C has Lebesgue measure zero as well, and the cardinality of the power set of C equals that of the power set of $[0, 1]$.

An interesting fact is that the Lebesgue-measurable sets Σ_λ coincide with the σ-algebra $\mathcal{B}(\mathbb{R}) \vee \mathcal{N}$, where $\mathcal{N}$ is the collection of subsets of $[0, 1]$ with outer measure zero. This follows from Exercise 2.4.

2.2 A general extension theorem

Recall Theorem 2.6. Its content can be described by saying that there exists a measure on a σ-algebra (in this case on $\mathcal{B}$) that is such that its restriction to a suitable subclass of sets (the intervals) has a prescribed behavior. This is basically also valid in a more general situation. The proof of the main result of this section parallels to a large extent the development of the previous section. Let's state the theorem, known as Carathéodory's extension theorem.

Theorem 2.7 *Let Σ_0 be an algebra on a set S and let $\mu_0 : \Sigma_0 \to [0, \infty]$ be finitely additive and countably subadditive. Then there exists a measure μ defined on $\Sigma = \sigma(\Sigma_0)$ such that μ restricted to Σ_0 coincides with μ_0. The measure μ is thus an extension of μ_0, and this extension is unique if μ_0 is σ-finite on Σ_0.*

Proof We only sketch the main steps. First we define an outer measure on 2^S by putting

$$\mu^*(E) = \inf_{\Sigma_0(E)} \sum_{E_k \in \Sigma_0(E)} \mu_0(E_k),$$

where the infimum is taken over all $\Sigma_0(E)$, countable covers of E with elements E_k from Σ_0. Compare this to the definition in (2.3). It follows as in the proof of Lemma 2.4 that μ^* is an outer measure.

Let $E \in \Sigma_0$. Obviously, $\{E\}$ is a (finite) cover of E and so we have that $\mu^*(E) \leq \mu_0(E)$. Let $\{E_1, E_2, \ldots\}$ be a cover of E with the $E_k \in \Sigma_0$. Since μ_0 is countably subadditive and $E = \cup_k (E \cap E_k)$, we have $\mu_0(E) \leq \sum_k \mu_0(E \cap E_k)$ and since μ_0 is finitely additive, we also have $\mu_0(E \cap E_k) \leq \mu_0(E_k)$. Collecting these results we obtain

$$\mu_0(E) \leq \sum_k \mu_0(E_k).$$

Taking the infimum in the displayed inequality over all covers $\Sigma_0(E)$, we obtain $\mu_0(E) \leq \mu^*(E)$, for $E \in \Sigma_0$. Hence $\mu_0(E) = \mu^*(E)$ and μ^* is an extension of μ_0.

In order to show that μ^* restricted to Σ is a measure, it is by virtue of Theorem 2.3 sufficient to show that $\Sigma_0 \subset \Sigma_\mu$, because we then also have $\Sigma \subset \Sigma_\mu$. We proceed to prove the former inclusion. Let $F \in S$ be arbitrary, $\varepsilon > 0$. Then there exists a cover $\Sigma_0(F)$ such that $\mu^*(F) \geq \sum_{E_k \in \Sigma_0(F)} \mu_0(E_k) - \varepsilon$. Using the same kind of arguments as in the proof of Lemma 2.5, one obtains (using that μ_0 is additive on the algebra Σ_0, where it coincides with μ^*) for every $E \in \Sigma_0$

$$\begin{aligned}
\mu^*(F) + \varepsilon &\geq \sum_k \mu_0(E_k) \\
&= \sum_k \mu_0(E_k \cap E) + \sum_k \mu_0(E_k \cap E^c) \\
&= \sum_k \mu^*(E_k \cap E) + \sum_k \mu^*(E_k \cap E^c) \\
&\geq \mu^*(F \cap E) + \mu^*(F \cap E^c),
\end{aligned}$$

by subadditivity of μ^*. Letting $\varepsilon \to 0$, we arrive at $\mu^*(F) \geq \mu^*(F \cap E) + \mu^*(F \cap E^c)$, which is equivalent to $\mu^*(F) = \mu^*(F \cap E) + \mu^*(F \cap E^c)$. Below we denote the restriction of μ^* to Σ by μ.

We turn to the asserted unicity. Let ν be a measure on Σ that also coincides with μ_0 on Σ_0. The key result, which we will show below, is that μ and ν also coincide on the sets F in Σ for which $\mu(F) < \infty$. Indeed, assuming that this is the case, we can write for $E \in \Sigma$ and $S_1, S_2, \ldots$ disjoint sets in Σ_0 with $\mu(S_n) < \infty$ and $\cup_n S_n = S$, using that also $\mu(E \cap S_n) < \infty$,

$$\nu(E) = \sum_n \nu(E \cap S_n) = \sum_n \mu(E \cap S_n) = \mu(E).$$

Now we show the mentioned key result. Let $E \in \Sigma$. Consider a cover $\Sigma_0(E)$ of E. Then we have, since ν is a measure on Σ, $\nu(E) \leq \sum_k \nu(E_k) = \sum_k \mu_0(E_k)$. By taking the infimum over such covers, we obtain $\nu(E) \leq \mu^*(E) = \mu(E)$. We proceed to prove the converse inequality for sets E with $\mu(E) < \infty$.

Let $E \in \Sigma$ with $\mu(E) < \infty$. Given $\varepsilon > 0$, we can chose a cover $\Sigma_0(E)$ such that $\mu(E) > \sum_{E_k \in \Sigma_0(E)} \mu(E_k) - \varepsilon$. Let $U_n = \cup_{k=1}^n E_k$ and note that $U_n \in \Sigma_0$ and $U := \cup_{n=1}^\infty U_n = \cup_{k=1}^\infty E_k \in \Sigma$. Since $U \supset E$, we obtain $\mu(E) \leq \mu(U)$, whereas σ-additivity of μ yields $\mu(U) < \mu(E) + \varepsilon$, which implies $\mu(U \cap E^c) = \mu(U) - \mu(U \cap E) = \mu(U) - \mu(E) < \varepsilon$. Since it also follows that $\mu(U) < \infty$, there is $N \in \mathbb{N}$ such that $\mu(U) < \mu(U_N) + \varepsilon$. Below we use that $\mu(U_N) = \nu(U_N)$, the already established fact that $\mu \geq \nu$ on Σ and arrive at the following chain of (in)equalities.

$$\begin{aligned}
\nu(E) &= \nu(E \cap U) = \nu(U) - \nu(U \cap E^c) \\
&\geq \nu(U_N) - \mu(U \cap E^c) \\
&\geq \nu(U_N) - \varepsilon \\
&= \mu(U_N) - \varepsilon \\
&> \mu(U) - 2\varepsilon \\
&\geq \mu(E) - 2\varepsilon.
\end{aligned}$$

It follows that $\nu(E) \geq \mu(E)$. □

The assumption of countable subadditivity of μ_0 can in Theorem 2.7 be replaced with the *equivalent assumption* of countable additivity. See Exercise 2.7.

The assumption in Theorem 2.7 that the collection Σ_0 is an algebra can be weakened by only assuming that it is a *semiring*. This notion is beyond the scope of the present course.

Unicity of the extension fails to hold for μ_0 that are not σ-finite. Here is a counterexample. Let S be an infinite set and Σ_0 an arbitrary algebra consisting of the empty set and infinite subsets of S. Let $\mu_0(E) = \infty$, unless $E = \emptyset$, in which case we have $\mu_0(E) = 0$. Then $\mu(F)$ defined by $\mu(F) = \infty$, unless $F = \emptyset$, yields the extension of Theorem 2.7 on 2^S, whereas the counting measure on 2^S also extends μ_0.

2.3 Exercises

2.1 Let μ be an outer measure on some set S. Let $N \subset S$ be such that $\mu(N) = 0$. Show that $N \in \Sigma_\mu$.

2.2 Let (S, Σ, μ) be a measure space. A measurable covering of a subset A of S is a countable collection $\{E_i : i \in \mathbb{N}\} \subset \Sigma$ such that $A \subset \cup_{i=1}^\infty E_i$. Let $\mathcal{M}(A)$ be the collection of all measurable coverings of A. Put $\mu^*(A) = \inf\{\sum_{i=1}^\infty \mu(E_i) : \{E_1, E_2, \ldots\} \in \mathcal{M}(A)\}$. Show that μ^* is an outer measure on S and that $\mu^*(E) = \mu(E)$, if $E \in \Sigma$. Show also that $\mu^*(A) = \inf\{\mu(E) : E \supset A, E \in \Sigma\}$. We call μ^* the outer measure associated to μ.

2.3 Let (S, Σ, μ) be a measure space and let μ^* be the outer measure on S associated to μ. If $A \subset S$, then there exists $E \in \Sigma$ such that $A \subset E$ and $\mu^*(A) = \mu(E)$. Prove this.

2.4 Consider a measure space (S, Σ, μ) with σ-finite μ and let μ^* be the outer measure on S associated to μ. Show that $\Sigma_\mu \subset \Sigma \vee \mathcal{N}$, where $\mathcal{N}$ is the collection of all μ-null sets. *Hint:* Reduce the question to the case where μ is finite. Take then $A \in \Sigma_\mu$ and E as in Exercise 2.3 and show that $\mu^*(E \setminus A) = 0$. (By Exercise 2.1, we even have $\Sigma_\mu = \Sigma \vee \mathcal{N}$.)

2.5 Show that the Lebesgue measure λ is translation invariant, i.e. $\lambda(E + x) = \lambda(E)$ for all $E \in \Sigma_\lambda$, where $E + x = \{y + x : y \in E\}$.

2.6 This exercise aims at showing the existence of a set $E \notin \Sigma_\lambda$. First we define an equivalence relation $\sim$ on $\mathbb{R}$ by saying $x \sim y$ iff $x - y \in \mathbb{Q}$. By *the axiom of choice* there exists a set $E \subset (0, 1)$ that has *exactly one point* in each equivalence class induced by $\sim$. The set E is our candidate.

(a) Show the following two statements. If $x \in (0, 1)$, then $\exists q \in \mathbb{Q} \cap (-1, 1) : x \in E + q$. If $q, r \in \mathbb{Q}$ and $q \neq r$, then $(E + q) \cap (E + r) = \emptyset$.

(b) Assume that $E \in \Sigma_\lambda$. Put $S = \cup_{q \in \mathbb{Q} \cap (-1,1)} E + q$ and note that $S \subset (-1, 2)$. Use translation invariance of λ (Exercise 2.5) to show that $\lambda(S) = 0$, whereas at the same time one should have $\lambda(S) \geq \lambda(0, 1)$.

(c) Show that $\lambda^*(E) = 1$ and $\lambda^*((0, 1) \setminus E) = 1$.

2.7 Let μ_0 be finitely additive on an algebra Σ_0 and on that algebra also σ-subadditive. Show that μ_0 is σ-additive on Σ_0. *Hint: Show that* $\mu_0(E) \geq \sum_{k=1}^n \mu_0(E_k)$ *for every* $n \geq 1$, *where you are supposed to also give a meaning to the sets* E, E_k.

3 Measurable functions and random variables

In this chapter we define random variables as *measurable* functions on a probability space and derive some properties.

3.1 General setting

Let (S, Σ) be a measurable space. Recall that the elements of Σ are called measurable sets. Also recall that $\mathcal{B} = \mathcal{B}(\mathbb{R})$ is the collection of all the Borel sets of $\mathbb{R}$.

Definition 3.1 A mapping $h : S \to \mathbb{R}$ is called *measurable* if $h^{-1}[B] \in \Sigma$ for all $B \in \mathcal{B}$.

It is clear that this definition depends on $\mathcal{B}$ and Σ. When there are more σ-algebras in the picture, we sometimes speak of Σ-measurable functions, or $\Sigma/\mathcal{B}$-measurable functions, depending on the situation. If S is a topological space with a topology $\mathcal{T}$ and if $\Sigma = \sigma(\mathcal{T})$, a measurable function h is also called a *Borel* measurable function.

Remark 3.2 Consider $E \subset S$. Recall that the indicator function of E is defined by $\mathbf{1}_E(s) = 1$ if $s \in E$ and $\mathbf{1}_E(s) = 0$ if $s \notin E$. Check that $\mathbf{1}_E$ is a measurable function iff E is a measurable set.

Sometimes one wants to extend the range of the function h to $[-\infty, \infty]$. If this happens to be the case, we extend $\mathcal{B}$ with the singletons $\{-\infty\}$ and $\{\infty\}$, and work with $\bar{\mathcal{B}} = \sigma(\mathcal{B} \cup \{\{-\infty\}, \{\infty\}\})$. We call $h : S \to [-\infty, \infty]$ measurable if $h^{-1}[B] \in \Sigma$ for all $B \in \bar{\mathcal{B}}$.

Below we will often use the shorthand notation $\{h \in B\}$ for the set $\{s \in S : h(s) \in B\}$. Likewise we also write $\{h \leq c\}$ for the set $\{s \in S : h(s) \leq c\}$. Many variations on this theme are possible.

Proposition 3.3 *Let (S, Σ) be a measurable space and $h : S \to \mathbb{R}$.*

(i) If $\mathcal{C}$ is a collection of subsets of $\mathbb{R}$ such that $\sigma(\mathcal{C}) = \mathcal{B}$, and if $h^{-1}[C] \in \Sigma$ for all $C \in \mathcal{C}$, then h is measurable.

(ii) If $\{h \leq c\} \in \Sigma$ for all $c \in \mathbb{R}$, then h is measurable.

(iii) If S is topological and h continuous, then h is measurable with respect to the σ-algebra generated by the open sets. In particular any constant function is measurable.

(iv) If h is measurable and another function $f : \mathbb{R} \to \mathbb{R}$ is Borel measurable ($\mathcal{B}/\mathcal{B}$-measurable), then $f \circ h$ is measurable as well.

Proof (i) Put $\mathcal{D} = \{B \in \mathcal{B} : h^{-1}[B] \in \Sigma\}$. One easily verifies that $\mathcal{D}$ is a σ-algebra and it is evident that $\mathcal{C} \subset \mathcal{D} \subset \mathcal{B}$. It follows that $\mathcal{D} = \mathcal{B}$.

(ii) This is an application of the previous assertion. Take $\mathcal{C} = \{(-\infty, c] : c \in \mathbb{R}\}$.

(iii) Take as $\mathcal{C}$ the collection of open sets and apply (i).

(iv) Take $B \in \mathcal{B}$, then $f^{-1}[B] \in \mathcal{B}$ since f is Borel. Because h is measurable, we then also have $(f \circ h)^{-1}[B] = h^{-1}[f^{-1}[B]] \in \Sigma$. □

Remark 3.4 There are many variations on the assertions of Proposition 3.3 possible. For instance in (ii) we could also use $\{h < c\}$, or $\{h > c\}$. Furthermore, (ii) is true for $h : S \to [-\infty, \infty]$ as well. We proved (iv) by a simple composition argument, which also applies to a more general situation. Let (S_i, Σ_i) be measurable spaces $(i = 1, 2, 3)$, $h : S_1 \to S_2$ is Σ_1/Σ_2-measurable and $f : S_2 \to S_3$ is Σ_2/Σ_3-measurable. Then $f \circ h$ is Σ_1/Σ_3-measurable.

The set of measurable functions will also be denoted by Σ. This notation is of course a bit ambiguous, but it turns out that no confusion can arise. Remark 3.2, in a way justifies this notation. The remark can, with the present convention, be rephrased as $\mathbf{1}_E \in \Sigma$ iff $E \in \Sigma$. Later on we often need the set of nonnegative measurable functions, denoted Σ^+.

Fortunately, the set Σ of measurable functions is closed under elementary operations.

Proposition 3.5 *We have the following properties.*

(i) The collection Σ of Σ-measurable functions is a vector space and products of measurable functions are measurable as well.

(ii) Let (h_n) be a sequence in Σ. Then also $\inf h_n, \sup h_n, \liminf h_n, \limsup h_n$ are in Σ, where we extend the range of these functions to $[-\infty, \infty]$. The set L, consisting of all $s \in S$ for which $\lim_n h_n(s)$ exists as a finite limit, is measurable.

Proof (i) If $h \in \Sigma$ and $\lambda \in \mathbb{R}$, then λh is also measurable (use (ii) of the previous proposition for $\lambda \neq 0$). To show that the sum of two measurable functions is measurable, we first note that $\{(x_1, x_2) \in \mathbb{R}^2 : x_1 + x_2 > c\} = \cup_{q \in \mathbb{Q}} \{(x_1, x_2) \in \mathbb{R}^2 : x_1 > q, x_2 > c - q\}$ (draw a picture!). But then we also have $\{h_1 + h_2 > c\} = \cup_{q \in \mathbb{Q}} (\{h_1 > q\} \cap \{h_2 > c - q\})$, a countable union. To show that products of measurable functions are measurable is left as Exercise 3.1.

(ii) Since $\{\inf h_n \geq c\} = \cap_n \{h_n \geq c\}$, it follows that $\inf h_n \in \Sigma$. To $\sup h_n$ a similar argument applies, that then also yield measurability of $\liminf h_n = \sup_n \inf_{m \geq n} h_m$ and $\limsup h_n$. To show the last assertion we consider $h := \limsup h_n - \liminf h_n$. Then $h : S \to [-\infty, \infty]$ is measurable. The assertion follows from $L = \{\limsup h_n < \infty\} \cap \{\liminf h_n > -\infty\} \cap \{h = 0\}$. □

For later use we present the *Monotone Class Theorem*.

Theorem 3.6 *Let $\mathcal{H}$ be a vector space of bounded functions, with the following properties.*

(i) $1 \in \mathcal{H}$.

(ii) If (f_n) is a nonnegative sequence in $\mathcal{H}$ such that $f_{n+1} \geq f_n$ for all n, and $f := \lim f_n$ is bounded as well, then $f \in \mathcal{H}$.

If, in addition, $\mathcal{H}$ contains the indicator functions of sets in a π-system $\mathcal{I}$, then $\mathcal{H}$ contains all bounded $\sigma(\mathcal{I})$-measurable functions.

Proof Put $\mathcal{D} = \{F \subset S : \mathbf{1}_F \in \mathcal{H}\}$. One easily verifies that $\mathcal{D}$ is a d-system, and that it contains $\mathcal{I}$. Hence, by Corollary 1.14, we have $\Sigma := \sigma(\mathcal{I}) \subset \mathcal{D}$. We will use this fact later in the proof.

Let f be a bounded, $\sigma(\mathcal{I})$-measurable function. Without loss of generality, we may assume that $f \geq 0$ (add a constant otherwise), and $f < K$ for some real constant K. Introduce the functions f_n defined by $f_n = 2^{-n}\lfloor 2^n f \rfloor$. In explicit terms, the f_n are given by

$$f_n(s) = \sum_{i=0}^{K2^n-1} i2^{-n}\mathbf{1}_{\{i2^{-n} \leq f < (i+1)2^{-n}\}}(s).$$

Then we have for all n that f_n is a bounded measurable function, $f_n \leq f$, and $f_n \uparrow f$ (check this!). Moreover, each f_n lies in $\mathcal{H}$. To see this, observe that $\{i2^{-n} \leq f < (i+1)2^{-n}\} \in \Sigma$, since f is measurable. But then this set is also an element of $\mathcal{D}$, since $\Sigma \subset \mathcal{D}$ (see above) and hence $\mathbf{1}_{\{i2^{-n} \leq f < (i+1)2^{-n}\}} \in \mathcal{H}$. Since $\mathcal{H}$ is a vector space, linear combinations remain in $\mathcal{H}$ and therefore $f_n \in \mathcal{H}$. Property (ii) of $\mathcal{H}$ yields $f \in \mathcal{H}$. □

3.2 Random variables

We return to the setting of Section 1.5 and so we consider a set (of outcomes) Ω and $\mathcal{F}$ a σ-algebra (of events) defined on it. In this setting Definition 3.1 takes the following form.

Definition 3.7 A function $X : \Omega \to \mathbb{R}$ is called a *random variable* if it is ($\mathcal{F}$-)measurable.

Following the tradition, we denote random variables by X (or other capital letters), rather than by h, as in the previous sections. By definition, random variables are nothing else but measurable functions with respect to a given σ-algebra $\mathcal{F}$. Given $X : \Omega \to \mathbb{R}$, let $\sigma(X) = \{X^{-1}[B] : B \in \mathcal{B}\}$. Then $\sigma(X)$ is a σ-algebra, and X is a random variable in the sense of Definition 3.7 iff $\sigma(X) \subset \mathcal{F}$. It follows that $\sigma(X)$ is the smallest σ-algebra on Ω such that X is a random variable. See also Exercise 3.2.

If we have a collection of mappings $X := \{X_i : \Omega \to \mathbb{R} | i \in I\}$, then we denote by $\sigma(X)$ the smallest σ-algebra on Ω such that all the X_i become measurable. See Exercise 3.3.

Having a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, a random variable X, and the measurable space $(\mathbb{R}, \mathcal{B})$, we will use these ingredients to endow the latter space with a probability measure. Define $\mu : \mathcal{B} \to [0, 1]$ by

$$\mu(B) := \mathbb{P}(X \in B) = \mathbb{P}(X^{-1}[B]). \tag{3.1}$$

It is straightforward to check that μ is a probability measure on $\mathcal{B}$. Commonly used alternative notations for μ are $\mathbb{P}^X$, or $\mathcal{L}_X$, $\mathcal{L}^X$. This probability measure is referred to as the *distribution* of X or the *law* of X. Along with the distribution

of X, we introduce its *distribution function*, usually denoted by F (or F_X, or F^X). By definition it is the function $F : \mathbb{R} \to [0,1]$, given by $F(x) = \mu((-\infty, x]) = \mathbb{P}(X \leq x)$.

Proposition 3.8 *The distribution function of a random variable is right continuous, non-decreasing and satisfies* $\lim_{x\to\infty} F(x) = 1$ *and* $\lim_{x\to-\infty} F(x) = 0$. *The set of points where F is discontinuous is at most countable.*

Proof Exercise 3.4. □

The fundamental importance of distribution functions in probability is based on the following proposition.

Proposition 3.9 *Let μ_1 and μ_2 be two probability measures on $\mathcal{B}$. Let F_1 and F_2 be the corresponding distribution functions. If $F_1(x) = F_2(x)$ for all x, then $\mu_1 = \mu_2$.*

Proof Consider the π-system $\mathcal{I} = \{(-\infty, x] : x \in \mathbb{R}\}$ and apply Theorem 1.15. □

This proposition thus states, in a different wording, that for a random variable X, its distribution, the collection of all probabilities $\mathbb{P}(X \in B)$ with $B \in \mathcal{B}$, is determined by the distribution function F_X.

We call *any* function on $\mathbb{R}$ that has the properties of Proposition 3.8 a distribution function. Note that any distribution function is Borel measurable (sets $\{F \geq c\}$ are intervals and thus in $\mathcal{B}$). Below, in Theorem 3.10, we justify this terminology. We will see that for any distribution function F, it is possible to construct a random variable on some $(\Omega, \mathcal{F}, \mathbb{P})$, whose distribution function equals F. This theorem is founded on the existence of the Lebesgue measure λ on the Borel sets $\mathcal{B}[0,1]$ of $[0,1]$, see Theorem 1.5. We now give a probabilistic translation of this theorem. Consider $(\Omega, \mathcal{F}, \mathbb{P}) = ([0,1], \mathcal{B}[0,1], \lambda)$. Let $U : \Omega \to [0,1]$ be the identity map. The distribution of U on $[0,1]$ is trivially the Lebesgue measure again, in particular the distribution function F^U of U satisfies $F^U(x) = x$ for $x \in [0,1]$ and so $\mathbb{P}(a < U \leq b) = F^U(b) - F^U(a) = b - a$ for $a, b \in [0,1]$ with $a \leq b$. Hence, to the distribution function F^U corresponds a probability measure on $([0,1], \mathcal{B}[0,1])$ and there exists a random variable U on this space, such that U has F^U as its distribution function. The random variable U is said to have the standard uniform distribution.

The proof of Theorem 3.10 (Skorokhod's representation of a random variable with a given distribution function) below is easy in the case that F is continuous and strictly increasing (Exercise 3.6), given the just presented fact that a random variable with a uniform distribution exists. The proof that we give below for the general case just follows a more careful line of arguments, but is in spirit quite similar.

Theorem 3.10 *Let F be a distribution function on $\mathbb{R}$. Then there exists a probability space and a random variable $X : \Omega \to \mathbb{R}$ such that F is the distribution function of X.*

Proof Let $(\Omega, \mathcal{F}, \mathbb{P}) = ((0,1), \mathcal{B}(0,1), \lambda)$. We define $X^-(\omega) = \inf\{z \in \mathbb{R} : F(z) \geq \omega\}$. Then $X^-(\omega)$ is finite for all ω and X^- is Borel measurable function, so a random variable, as this follows from the relation to be proven below, valid for all $c \in \mathbb{R}$ and $\omega \in (0,1)$,

$$X^-(\omega) \leq c \Leftrightarrow F(c) \geq \omega. \tag{3.2}$$

This equivalence can be represented as $\{X^- \leq c\} = [0, F(c)]$. It also shows that X^- serves in a sense as an inverse function of F. We now show that (3.2) holds. The implication $F(c) \geq \omega \Rightarrow X^-(\omega) \leq c$ is immediate from the definition of X^-. Conversely, let $z > X^-(\omega)$. Then $F(z) \geq \omega$, by definition of X^-. We now take a sequence of $z_n > X^-(\omega)$ and $z_n \downarrow X^-(\omega)$. Since F is right continuous, we obtain $F(X^-(\omega)) \geq \omega$. It trivially holds that $F(X^-(\omega)) \leq F(c)$ if $X^-(\omega) \leq c$, because F is non-decreasing. Combination with the previous inequality yields $F(c) \geq \omega$. This proves (3.2). In order to find the distribution function of X^-, we compute $\mathbb{P}(X^- \leq c) = \mathbb{P}([0, F(c)]) = \lambda([0, F(c)]) = F(c)$. □

3.3 Independence

Recall the definition of independent events. Two events $E, F \in \mathcal{F}$ are called independent if the product rule $\mathbb{P}(E \cap F) = \mathbb{P}(E)\mathbb{P}(F)$ holds. In the present section we generalize this notion of independence to independence of a sequence of events and to independence of a sequence of σ-algebras. It is even convenient and elegant to start with the latter.

Definition 3.11

(i) A sequence of σ-algebras $\mathcal{F}_1, \mathcal{F}_2, \ldots$ is called independent, if for every n it holds that $\mathbb{P}(E_1 \cap \cdots \cap E_n) = \prod_{i=1}^n \mathbb{P}(E_i)$, for all choices of $E_i \in \mathcal{F}_i$ $(i = 1, \ldots, n)$.

(ii) A sequence of random variables $X_1, X_2, \ldots$ is called independent if the σ-algebras $\sigma(X_1), \sigma(X_2), \ldots$ are independent.

(iii) A sequence of events $E_1, E_2, \ldots$ is called independent if the random variables $\mathbf{1}_{E_1}, \mathbf{1}_{E_2}, \ldots$ are independent.

The above definition also applies to finite sequences. For instance, a finite sequence of σ-algebras $\mathcal{F}_1, \ldots, \mathcal{F}_n$ is called independent if the infinite sequence $\mathcal{F}_1, \mathcal{F}_2, \ldots$ is independent in the sense of part (ii) of the above definition, where $\mathcal{F}_m = \{\emptyset, \Omega\}$ for $m > n$. It follows that two σ-algebras $\mathcal{F}_1$ and $\mathcal{F}_2$ are independent, if $\mathbb{P}(E_1 \cap E_2) = \mathbb{P}(E_1)\mathbb{P}(E_2)$ for all $E_1 \in \mathcal{F}_1$ and $E_2 \in \mathcal{F}_2$. To check independence of two σ-algebras, Theorem 1.15 is again helpful.

Proposition 3.12 *Let $\mathcal{I}$ and $\mathcal{J}$ be π-systems and suppose that for all $I \in \mathcal{I}$ and $J \in \mathcal{J}$ the product rule $\mathbb{P}(I \cap J) = \mathbb{P}(I)\mathbb{P}(J)$ holds. Then the σ-algebras $\sigma(\mathcal{I})$ and $\sigma(\mathcal{J})$ are independent.*

Proof Put $\mathcal{G} = \sigma(\mathcal{I})$ and $\mathcal{H} = \sigma(\mathcal{J})$. We define for each $I \in \mathcal{I}$ the *finite measures* μ_I and ν_I on $\mathcal{H}$ by $\mu_I(H) = \mathbb{P}(H \cap I)$ and $\nu_I(H) = \mathbb{P}(H)\mathbb{P}(I)$ $(H \in \mathcal{H})$.

Notice that μ_I and ν_I coincide on $\mathcal{J}$ by assumption and that $\mu_I(\Omega) = \mathbb{P}(I) = \nu_I(\Omega)$. Theorem 1.15 yields that $\mu_I(H) = \nu_I(H)$ for all $H \in \mathcal{H}$.

Now we consider for each $H \in \mathcal{H}$ the finite measures μ^H and ν^H on $\mathcal{G}$ defined by $\mu^H(G) = \mathbb{P}(G \cap H)$ and $\nu^H(G) = \mathbb{P}(G)\mathbb{P}(H)$. By the previous step, we see that μ^H and ν^H coincide on $\mathcal{I}$. Invoking Theorem 1.15 again, we obtain $\mathbb{P}(G \cap H) = \mathbb{P}(G)\mathbb{P}(H)$ for all $G \in \mathcal{G}$ and $H \in \mathcal{H}$. □

Here is an important consequence.

Corollary 3.13 *Let X_1, X_2 be random variables defined on some $(\Omega, \mathcal{F}, \mathbb{P})$. Then X_1 and X_2 are independent iff $\mathbb{P}(\{X_1 \leq x_1\} \cap \{X_2 \leq x_2\}) = \mathbb{P}(X_1 \leq x_1)\mathbb{P}(X_2 \leq x_2)$ for all $x_1, x_2 \in \mathbb{R}$.*

Proof Combine Proposition 3.12 and Exercise 3.2. □

Lemma 3.14 (Borel-Cantelli) *Let $E_1, E_2, \ldots$ be a sequence of events.*

(i) If it has the property that $\sum_{n\geq 1} \mathbb{P}(E_n) < \infty$, then $\mathbb{P}(\limsup E_n) = 0$.

(ii) If $\sum_{n\geq 1} \mathbb{P}(E_n) = \infty$ and if, moreover, the sequence is independent, then $\mathbb{P}(\limsup E_n) = 1$.

Proof (i) Let $U_n = \bigcup_{m\geq n} E_m$. Notice that the sequence (U_n) decreases to $U = \limsup E_n$. Hence we have $\mathbb{P}(U) \leq \mathbb{P}(U_n) \leq \sum_{m\geq n} \mathbb{P}(E_m)$, which converges to zero by assumption.

(ii) We prove that $\mathbb{P}(\liminf E_n^c) = 0$. Let $D_n^N = \bigcap_{m=n}^N E_m^c$ $(N \geq n)$. Notice that for fixed n the sequence $(D_n^N)_{N\geq n}$ decreases to $D_n^\infty := \bigcap_{m=n}^\infty E_m^c$. By independence we obtain $\mathbb{P}(D_n^N) = \prod_{m=n}^N (1 - \mathbb{P}(E_m))$, which is less than $\exp(-\sum_{m=n}^N \mathbb{P}(E_m))$. Hence by taking limits for $N \to \infty$, we obtain for every n that $\mathbb{P}(D_n^\infty) \leq \exp(-\sum_{m=n}^\infty \mathbb{P}(E_m)) = 0$. Finally, we observe that $\liminf E_n^c = \bigcup_{n=1}^\infty D_n^\infty$ and hence $\mathbb{P}(\liminf E_n^c) \leq \sum_{n=1}^\infty \mathbb{P}(D_n^\infty) = 0$. □

We close this section by presenting a nice construction of a probability space on which a sequence of *independent* random variables is defined, whereas at the same time the marginal distributions of each member is prescribed. This is the content of Theorem 3.16 below. It turns out that the probability space on which we can realize this construction is $([0,1), \mathcal{B}, \lambda)$. This must have something to do with Skorokhod's theorem 3.10!

Let's start with some preparations. Consider the set $[0,1)$ endowed with its Borel σ-algebra and let for each x the sequence $b_1(x), b_2(x), \ldots$ be its unique binary expansion. Uniqueness can be obtained in many ways, for instance $b_1(x) = 1$ iff $x \in [\frac{1}{2}, 1)$, $b_2(x) = 1$ iff $x \in [\frac{1}{4}, \frac{1}{2}) \cup [\frac{3}{4}, 1)$, etc. Then the functions $x \mapsto b_k(x)$, $k = 1, 2, \ldots$ are Borel-measurable and $x = \sum_{k=1}^\infty 2^{-k} b_k(x)$.

Lemma 3.15

(i) Let U be a random variable defined on some $(\Omega, \mathcal{F}, \mathbb{P})$ with values in $[0,1)$ and let $X_k = b_k \circ U$. Then U is uniformly distributed on $[0,1)$ iff the X_k are iid *with $\mathbb{P}(X_k = 1) = \frac{1}{2}$.*

(ii) If U is uniformly distributed on $[0,1)$, then there are Borel measurable functions $f_k : [0,1) \to [0,1)$ such that $Z_k = f_k \circ U$ defines an iid *sequence, with all Z_k uniformly distributed on $[0,1)$ as well.*

Proof (i) Let U have the uniform distribution on $[0,1)$. For $x_1, \ldots, x_n \in \{0,1\}$, one easily computes the joint probability $\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = 2^{-n}$. It follows that $\mathbb{P}(X_k = x_k) = \frac{1}{2}$ for all k and that the X_k are independent.

Conversely, let the X_k be distributed as assumed. Let V be a random variable having a uniform distribution on $[0,1)$. Then by the above part of the proof the sequence of $Y_k := b_k \circ V$ is distributed as the X_k and therefore $\sum_{k=1}^{\infty} 2^{-k} X_k$ has the same distribution as $\sum_{k=1}^{\infty} 2^{-k} Y_k$, which means that U and V have the same distribution. Hence U is uniformly distributed on $[0,1)$.

(ii) Take the functions b_k and relabel them in a rectangular array as b_{kj}, $j,k = 1,2,\ldots$ by using any bijective mapping from $\mathbb{N}$ onto $\mathbb{N}^2$. Put $f_k(x) := \sum_{j=1}^{\infty} 2^{-j} b_{kj}(x)$. The functions f_k are Borel measurable. Since for fixed k the $b_{kj} \circ U$ are *iid*, we have by the first part of the lemma that Z_k is uniform on $[0,1)$. Moreover, for different k and k' the sequences (b_{kj}) and $(b_{k'j})$ are disjoint and therefore Z_k and $Z_{k'}$ are independent. By extension of this argument the whole sequence (Z_k) becomes independent (think about this!). □

Here is the result we are after.

Theorem 3.16 *Let $\mu_1, \mu_2, \ldots$ be a sequence of probability measures on $(\mathbb{R}, \mathcal{B})$. Then there exists a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and random variables Y_k defined on it such that the law of each Y_k is μ_k and such that the sequence (Y_k) is an independent one.*

Proof Let $(\Omega, \mathcal{F}, \mathbb{P}) = ([0,1), \mathcal{B}[0,1), \lambda)$. Choose for each k a random variable X_k according to Theorem 3.10. Then certainly, the X_k have law μ_k. Let U be the identity mapping on $[0,1)$, then U is uniformly distributed on $[0,1)$. Choose the Z_k as in part (ii) of Lemma 3.15 and define $Y_k = X_k \circ Z_k$, $k \geq 1$. These are easily seen to have the desired properties. □

3.4 Exercises

3.1 If h_1 and h_2 are Σ-measurable functions on (S, Σ, μ), then $h_1 h_2$ is Σ-measurable too. Show this.

3.2 Let X be a random variable. Show that $\Pi(X) := \{\{X \leq x\} : x \in \mathbb{R}\}$ is a π-system and that it generates $\sigma(X)$.

3.3 Let $\{Y_\gamma : \gamma \in C\}$ be an arbitrary collection of random variables and $\{X_n : n \in \mathbb{N}\}$ be a countable collection of random variables, all defined on the same probability space.

(a) Show that $\sigma\{Y_\gamma : \gamma \in C\} = \sigma\{Y_\gamma^{-1}(B) : \gamma \in C, B \in \mathcal{B}\}$.

(b) Let $\mathcal{X}_n = \sigma\{X_1, \ldots, X_n\}$ $(n \in \mathbb{N})$ and $\mathcal{A} = \bigcup_{n=1}^{\infty} \mathcal{X}_n$. Show that $\mathcal{A}$ is an algebra and that $\sigma(\mathcal{A}) = \sigma\{X_n : n \in \mathbb{N}\}$.

3.4 Prove Proposition 3.8.

3.5 Let $\mathcal{F}$ be a σ-algebra on Ω with the property that for all $F \in \mathcal{F}$ it holds that $\mathbb{P}(F) \in \{0,1\}$. Let $X : \Omega \to \mathbb{R}$ be $\mathcal{F}$-measurable. Show that for some $c \in \mathbb{R}$ one has $\mathbb{P}(X = c) = 1$. (*Hint:* $\mathbb{P}(X \leq x) \in \{0,1\}$ for all x.)

3.6 Let F be a strictly increasing and continuous distribution function. Let U be a random variable defined on some $(\Omega, \mathcal{F}, \mathbb{P})$ having a uniform distribution on $[0,1]$ and put $X = F^{-1}(U)$. Show that X is $\mathcal{F}$-measurable and that it has distribution function F.

3.7 Let F be a distribution function and put $X^+(\omega) = \inf\{x \in \mathbb{R} : F(x) > \omega\}$. Show that (next to X^-) also X^+ has distribution function F and that $\mathbb{P}(X^+ = X^-) = 1$ (*Hint:* $\mathbb{P}(X^- \leq q < X^+) = 0$ for all $q \in \mathbb{Q}$). Show also that X^+ is a right continuous function and Borel-measurable.

3.8 Consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let $\mathcal{I}_1, \mathcal{I}_2, \mathcal{I}_3$ be π-systems on Ω with the properties $\Omega \in \mathcal{I}_k$ and $\mathcal{I}_k \subset \mathcal{F}$, for all k. Assume that for all $I_k \in \mathcal{I}_k$ $(k = 1,2,3)$

$$\mathbb{P}(I_1 \cap I_2 \cap I_3) = \mathbb{P}(I_1)\mathbb{P}(I_2)\mathbb{P}(I_3).$$

Show that $\sigma(\mathcal{I}_1), \sigma(\mathcal{I}_2), \sigma(\mathcal{I}_3)$ are independent.

3.9 Let $\mathcal{G}_1, \mathcal{G}_2, \ldots$ be sub-σ-algebras of a σ-algebra $\mathcal{F}$ on a set Ω and let $\mathcal{G} = \sigma(\mathcal{G}_1 \cup \mathcal{G}_2 \cup \ldots)$.

(a) Show that $\Pi = \{G_{i_1} \cap G_{i_2} \cap \ldots \cap G_{i_k} : k \in \mathbb{N}, i_k \in \mathbb{N}, G_{i_j} \in \mathcal{G}_{i_j}\}$ is a π-system that generates $\mathcal{G}$.

(b) Assume that $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space and that $\mathcal{G}_1, \mathcal{G}_2, \ldots$ is an independent sequence. Let M and N be disjoint subsets of $\mathbb{N}$ and put $\mathcal{M} = \sigma(\mathcal{G}_i, i \in M)$ and $\mathcal{N} = \sigma(\mathcal{G}_i, i \in N)$. Show that $\mathcal{M}$ and $\mathcal{N}$ are independent σ-algebras.

3.10 Consider an independent sequence $X_1, X_2, \ldots$. Let $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$ and $\mathcal{T}_n = \sigma(X_{n+1}, X_{n+2}, \ldots)$, $n \geq 1$. Let $\mathcal{I}$ be the collection of events of the type $\{X_1 \in B_1, \ldots, X_n \in B_n\}$, with the B_i Borel sets in $\mathbb{R}$. Show that $\mathcal{I}$ is a π-system that generates $\mathcal{F}_n$. Find a π-system that generates $\mathcal{T}_n$ and show that $\mathcal{F}_n$ and $\mathcal{T}_n$ are independent. (Use Proposition 3.12.)

3.11 Consider an infinite sequence of coin tossing. We take $\Omega = \{H,T\}^\infty$, a typical element ω is an infinite sequence $(\omega_1, \omega_2, \ldots)$ with each $\omega_n \in \{H,T\}$, and $\mathcal{F} = \sigma(\{\omega \in \Omega : \omega_n = w\}, w \in \{H,T\}, n \in \mathbb{N})$. Define functions X_n by $X_n(\omega) = 1$ if $\omega_n = H$ and $X_n(\omega) = 0$ if $\omega_n = T$.

(a) Show that all X_n are random variables, i.e. everyone of them is measurable.

(b) Let $S_n = \sum_{i=1}^n X_i$. Show that also S_n is a random variable.

(c) Let $p \in [0,1]$ and $E_p = \{\omega \in \Omega : \lim_{n\to\infty} \frac{1}{n} S_n(\omega) = p\}$. Show that E_p is an $\mathcal{F}$-measurable set.

4 Integration

In elementary courses on Probability Theory, there is usually a distinction between random variables X having a discrete distribution, on $\mathbb{N}$ say, and those having a density. In the former case we have for the expectation $\mathbb{E}X$ the expression $\sum_k k\,\mathbb{P}(X=k)$, whereas in the latter case one has $\mathbb{E}X = \int x f(x)\,\mathrm{d}x$. This distinction is annoying and not satisfactory from a mathematical point of view. Moreover, there exist random variables whose distributions are neither discrete, nor do they admit a density. Here is an example. Suppose Y and Z, defined on the same $(\Omega, \mathcal{F}, \mathbb{P})$, are independent random variables. Assume that $\mathbb{P}(Y=0) = \mathbb{P}(Y=1) = \frac{1}{2}$ and that Z has a standard normal distribution. Let $X = YZ$ and F the distribution function of X. Easy computations (do them!) yield $F(x) = \frac{1}{2}(\mathbf{1}_{[0,\infty)}(x) + \Phi(x))$. We see that F has a jump at $x=0$ and is differentiable on $\mathbb{R}\setminus\{0\}$, a distribution function of mixed type. How to compute $\mathbb{E}X$ in this case?

In this section we will see that expectations are special cases of the unifying concept of *Lebesgue integral*, a sophisticated way of addition. Lebesgue integrals have many advantages. It turns out that Riemann integrable functions (on a compact interval) are always Lebesgue integrable w.r.t. Lebesgue measure and that the two integrals are the same. Also sums are examples of Lebesgue integral. Furthermore, the theory of Lebesgue integrals allows for very powerful limit theorems. Below we work with a measurable space (S, Σ, μ).

4.1 Integration of simple functions

Bearing in mind the elementary formula for the area of a rectangle and the interpretation of the Riemann integral of a positive function as the area under its graph, it is natural to define the integral of a multiple of an indicator function $a \cdot \mathbf{1}_E$ as $a \cdot \mu(E)$, for $E \in \Sigma$. We extend this definition to the class of *simple functions*.

Definition 4.1 A function $f : S \to [0,\infty)$ is called a nonnegative simple function, if it has a representation as a finite sum

$$f = \sum_{i=1}^{n} a_i \mathbf{1}_{A_i}, \tag{4.1}$$

where $a_i \in [0,\infty)$ and $A_i \in \Sigma$. The class of all nonnegative simple functions is denoted by $\mathfrak{S}^+$.

Notice that a simple function is measurable. Since we remember that Riemann integrals are linear operators and knowing the definition of integral for an indicator function, we now present the definition of the integral of $f \in \mathfrak{S}^+$.

Definition 4.2 Let $f \in \mathfrak{S}^+$. The (*Lebesgue*) *integral* of f with respect to the measure μ is defined as

$$\int f\,\mathrm{d}\mu := \sum_{i=1}^{n} a_i \mu(A_i), \tag{4.2}$$

when f has representation (4.1).

Other notations that we often use for this integral are $\int f(s)\,\mu(\mathrm{d}s)$ and $\mu(f)$. Note that if $f = \mathbf{1}_A$, then $\mu(f) = \mu(\mathbf{1}_A) = \mu(A)$, so there is a bit of ambiguity in the notation, but also a reasonable level of consistency. Note that $\mu(f) \in [0, \infty]$ and also that the above summation is well defined, since all quantities involved are nonnegative, although possibly infinite. For products ab for $a, b \in [0, \infty]$, we use the convention $ab = 0$, when $a = 0$.

It should be clear that this definition of integral is, at first sight, troublesome. The representation of a simple function is not unique, and one might wonder if the just defined integral takes on different values for different representations. This would be very bad, and fortunately it is not the case.

Proposition 4.3 *Let f be a nonnegative simple function. Then the value of the integral $\mu(f)$ is independent of the chosen representation.*

Proof Step 1. Let f be given by (4.1) and define $\phi : S \to \{0,1\}^n$ by $\phi(s) = (\mathbf{1}_{A_1}(s), \ldots, \mathbf{1}_{A_n}(s))$. Let $\{0,1\}^n = \{u_1, \ldots, u_m\}$ where $m = 2^n$ and put $U_k = \phi^{-1}(u_k)$. Then the collection $\{U_1, \ldots, U_m\}$ is a measurable partition of S (the sets U_k are measurable). We will also need the sets $S_i = \{k : U_k \subset A_i\}$ and $T_k = \{i : U_k \subset A_i\}$. Note that these sets are dual in the sense that $k \in S_i$ iff $i \in T_k$.

Below we will use the fact $A_i = \cup_{k \in S_i} U_k$, when we rewrite (4.1). We obtain by interchanging the summation order

$$\begin{aligned} f = \sum_i a_i \mathbf{1}_{A_i} &= \sum_i a_i \Big(\sum_{k \in S_i} \mathbf{1}_{U_k} \Big) \\ &= \sum_k \Big(\sum_{i \in T_k} a_i \Big) \mathbf{1}_{U_k}. \end{aligned} \tag{4.3}$$

Now apply the definition of $\mu(f)$ by using the representation of f given by (4.3). This gives $\mu(f) = \sum_k (\sum_{i \in T_k} a_i) \mu(U_k)$. Interchanging the summation order, we see that this is equal to $\sum_i a_i (\sum_{k \in S_i} \mu(U_k)) = \sum_i a_i \mu(A_i)$, which coincides with (4.2). We conclude that if f is given by (4.1), we can also represent f in a similar fashion by using a partition, and that both representations give the same value for the integral.

Step 2: Suppose that we have two representations of a simple function f, one is as in (4.1) with the collection of A_i a measurable partition of S. The other one is

$$f = \sum_{j=1}^{m} b_j \mathbf{1}_{B_j}, \tag{4.4}$$

where the B_j form a measurable partition of S as well. We obtain a third measurable partition of S by taking the collection of all intersections $A_i \cap B_j$. Notice that if $s \in A_i \cap B_j$, then $f(s) = a_i = b_j$ and so we have the implication

$A_i \cap B_j \neq \emptyset \Rightarrow a_i = b_j$. We compute the integral of f according to the definition. Of course, this yields (4.2) by using the representation (4.1) of f, but $\sum_j b_j \mu(B_j)$ if we use (4.4). Rewrite

$$\begin{aligned}\sum_j b_j \mu(B_j) &= \sum_j b_j \mu(\cup_i (A_i \cap B_j)) = \sum_j b_j \sum_i \mu(A_i \cap B_j) \\ &= \sum_i \sum_j b_j \mu(A_i \cap B_j) = \sum_i \sum_j a_i \mu(A_i \cap B_j) \\ &= \sum_i a_i \sum_j \mu(A_i \cap B_j) = \sum_i a_i \mu(\cup_j (A_i \cap B_j)) \\ &= \sum_i a_i \mu(A_i),\end{aligned}$$

which shows that the two formulas for the integral are the same.

Step 3: Take now two arbitrary representations of f of the form (4.1) and (4.4). According to step 1, we can replace each of them with a representation in terms of a measurable partition, without changing the value of the integral. According to step 2, each of the representations in terms of the partitions also gives the same value of the integral. This proves the proposition. □

Corollary 4.4 *Let $f \in \mathfrak{S}^+$ and suppose that f assumes the* different *values $0 \leq a_1, \ldots, a_n < \infty$. Then $\mu(f) = \sum_{i=1}^n a_i \mu(\{f = a_i\})$. If f is indentically zero, then $\mu(f) = 0$.*

Proof We have the representation $f = \sum_{i=1}^n a_i \mathbf{1}_{\{f=a_i\}}$. The expression for $\mu(f)$ follows from Definition 4.2, which is unambiguous by Proposition 4.3. The result for the zero function follows by the representation $f = 0 \times \mathbf{1}_S$ and the convention $ab = 0$ for $a = 0$ and $b \in [0, \infty]$. □

Example 4.5 Here is an instructive example. Let $(S, \Sigma, \mu) = (\mathbb{N}, 2^{\mathbb{N}}, \tau)$, with counting measure τ. A function f on $\mathbb{N}$ can be identified with a sequence (f_i). Then f can be represented in a somewhat cumbersome way (it just means that $f(k) = f_k$) by

$$f(k) = \sum_{i=1}^{\infty} f_i \mathbf{1}_{\{i\}}(k).$$

For now, we assume $f_i = 0$ for $i > n$ and $f_i \geq 0$ for $i \leq n$; obviously, f is a simple function. Since $\tau(\{i\}) = 1$, we get $\tau(f) = \sum_{i=1}^n f_i$, nothing else but the finite sum of the f_i. In this case, integration is just summation. Of course, a different representation would yield the same answer. A generalization occurs when the set of values $\{f_i : i \in \mathbb{N}\}$ is finite. Then f is still a simple function if the f_i are nonnegative, as in Corollary 4.4. But note that now it may happen that $\tau(f) = \infty$.

Example 4.6 Let $(S, \Sigma, \mu) = ([0,1], \mathcal{B}([0,1]), \lambda)$ and f the indicator of the rational numbers in $[0,1]$, $f = \mathbf{1}_{\mathbb{Q}\cap[0,1]}$. We know that $\lambda(\mathbb{Q} \cap [0,1]) = 0$ and it follows that $\lambda(f) = 0$. This f is a nice example of a function that is not Riemann integrable, whereas its Lebesgue integral trivially exists and has a very sensible value.

We say that a property of elements of S holds almost everywhere (usually abbreviated by a.e. or by μ-a.e.), if the set for which this property does not hold, has measure zero. For instance, we say that two measurable functions are almost everywhere equal, if $\mu(\{f \neq g\}) = 0$. Elementary properties of the integral are listed below.

Proposition 4.7 *Let $f, g \in \mathfrak{S}^+$ and $c \in [0, \infty)$.*

(i) If $f \leq g$ a.e., then $\mu(f) \leq \mu(g)$.

(ii) If $f = g$ a.e., then $\mu(f) = \mu(g)$.

(iii) $\mu(f + g) = \mu(f) + \mu(g)$ and $\mu(cf) = c\mu(f)$.

Proof (i) Represent f and g by means of measurable partitions, $f = \sum_i a_i \mathbf{1}_{A_i}$ and $g = \sum_j b_j \mathbf{1}_{B_j}$. We have $\{f > g\} = \cup_{i,j:a_i>b_j} A_i \cap B_j$, and since $\mu(\{f > g\}) = 0$, we have that $\mu(A_i \cap B_j) = 0$ if $a_i > b_j$. It follows that for all i and j, the inequality $a_i\mu(A_i \cap B_j) \leq b_j\mu(A_i \cap B_j)$ holds. We use this in the computations below.

$$\begin{aligned}
\mu(f) &= \sum_i a_i \mu(A_i) \\
&= \sum_i \sum_j a_i \mu(A_i \cap B_j) \\
&\leq \sum_i \sum_j b_j \mu(A_i \cap B_j) \\
&= \sum_j b_j \mu(B_j).
\end{aligned}$$

Assertion (ii) follows by a double application of (i), whereas (iii) can also be proved by using partitions and intersections $A_i \cap B_j$. □

4.2 A general definition of integral

We start with a definition, in which we use that we already know how to integrate simple functions.

Definition 4.8 Let f be a nonnegative measurable function. The integral of f is defined as $\mu(f) := \sup\{\mu(h) : h \leq f, h \in \mathfrak{S}^+\}$, where $\mu(h)$ is as in Definition 4.2.

Notice that for functions $f \in \mathfrak{S}^+$, Definition 4.8 yields for $\mu(f)$ the same as Definition 4.2 in the previous section. Thus there is no ambiguity in notation by using the same symbol μ. We immediately have some extensions of results in the previous section.

Proposition 4.9 *Let $f, g \in \Sigma^+$. If $f = 0$ a.e., then $\mu(f) = 0$. If $f \leq g$ a.e., then $\mu(f) \leq \mu(g)$, and if $f = g$ a.e., then $\mu(f) = \mu(g)$.*

Proof Let $f \in \Sigma^+$, $f = 0$ a.e. Take $h \in \mathfrak{S}^+$ with $h \leq f$. From this inequality we obtain $\{h > 0\} \subset \{f > 0\}$, and hence $\mu(\{h > 0\}) \leq \mu(\{f > 0\})$, but the latter measure is zero and hence $h = 0$ a.e. By Corollary 4.4 and Proposition 4.7(ii), $\mu(h) = 0$. Therefore, $\mu(f)$, being the supremum of those $\mu(h)$, is also zero.

Let $f, g \in \Sigma^+$ and $N = \{f > g\}$. Take $h \in \mathfrak{S}^+$ with $h \leq f$. Then also $h\mathbf{1}_N, h\mathbf{1}_{N^c} \in \mathfrak{S}^+$ and by Proposition 4.7(iii) and the fact that $h\mathbf{1}_N = 0$ a.e., we then have $\mu(h) = \mu(h\mathbf{1}_N) + \mu(h\mathbf{1}_{N^c}) = \mu(h\mathbf{1}_{N^c})$. Moreover,

$$h\mathbf{1}_{N^c} \leq f\mathbf{1}_{N^c} \leq g\mathbf{1}_{N^c} \leq g.$$

By definition of $\mu(g)$ (as a supremum), we obtain $\mu(h) \leq \mu(g)$. By taking the supremum in this inequality over all h, we get $\mu(f) \leq \mu(g)$, which gives the first assertion. The other one immediately follows. □

Example 4.10 We extend the situation of Example 4.5, by allowing infinitely many f_i to be positive. The result will be $\tau(f) = \sum_{i=1}^{\infty} f_i$, classically defined as $\lim_{n\to\infty} \sum_{i=1}^{n} f_i$. Check that this is in agreement with Definition 4.8. See also Exercise 4.1.

The following will frequently be used.

Lemma 4.11 *Let $f \in \Sigma^+$ and suppose that $\mu(f) = 0$. Then $f = 0$ a.e.*

Proof Because $\mu(f) = 0$, it holds that $\mu(h) = 0$ for all nonnegative simple functions with $h \leq f$. Take $h_n = \frac{1}{n}\mathbf{1}_{\{f \geq 1/n\}}$, then $h_n \in \mathfrak{S}^+$ and $h_n \leq f$. The equality $\mu(h_n) = 0$ implies $\mu(\{f \geq 1/n\}) = 0$. The result follows from $\{f > 0\} = \cup_n \{f \geq 1/n\}$ and Corollary 1.8. □

We now present the first important limit theorem, the *Monotone Convergence Theorem*.

Theorem 4.12 *Let (f_n) be a sequence in Σ^+, such that $f_{n+1} \geq f_n$ a.e. for each n. Let $f = \limsup f_n$. Then $\mu(f_n) \uparrow \mu(f) \leq \infty$.*

Proof We first consider the case where $f_{n+1}(s) \geq f_n(s)$ for all $s \in S$, so (f_n) is increasing *everywhere*. Then $f(s) = \lim f_n(s)$ for all $s \in S$, possibly with value infinity. It follows from Proposition 4.9, that $\mu(f_n)$ is an increasing sequence, bounded by $\mu(f)$. Hence we have $\ell := \lim \mu(f_n) \leq \mu(f)$.

We show that we actually have an equality. Take $h \in \mathfrak{S}^+$ with $h \leq f$, $c \in (0, 1)$ and put $E_n = \{f_n \geq ch\}$. The sequence (E_n) is obviously increasing and we show that its limit is S. Let $s \in S$ and suppose that $f(s) = 0$. Then also $h(s) = 0$ and $s \in E_n$ for every n. If $f(s) > 0$, then eventually $f_n(s) \geq cf(s) \geq ch(s)$, and so $s \in E_n$. This shows that $\cup_n E_n = S$. Consider the chain of inequalities

$$\ell \geq \mu(f_n) \geq \mu(f_n\mathbf{1}_{E_n}) \geq c\mu(h\mathbf{1}_{E_n}). \tag{4.5}$$

Suppose that h has representation (4.1). Then $\mu(h\mathbf{1}_{E_n}) = \sum_i a_i \mu(A_i \cap E_n)$. This is a finite sum of nonnegative numbers and hence the limit of it for $n \to \infty$ can be taken inside the sum and thus equals $\mu(h)$, since $E_n \uparrow S$ and the continuity of the measure (Proposition 1.7). From (4.5) we then conclude $\ell \geq c\mu(h)$, for all $c \in (0,1)$, and thus $\ell \geq \mu(h)$. Since this holds for all our h, we get $\ell \geq \mu(f)$ by taking the supremum over h. This proves the first case.

Next we turn to the *almost everywhere* version. Let $N_n = \{f_n > f_{n+1}\}$, by assumption $\mu(N_n) = 0$. Put $N = \cup_n N_n$, then also $\mu(N) = 0$. It follows that $\mu(f_n) = \mu(f_n \mathbf{1}_{N^c})$. But on N^c we have that $f = f\mathbf{1}_{N^c}$ and similarly $\mu(f) = \mu(f\mathbf{1}_{N^c})$. The previous case can be applied to get $\mu(f_n\mathbf{1}_{N^c}) \uparrow \mu(f\mathbf{1}_{N^c})$, from which the result follows. □

Example 4.13 Here is a nice application of Theorem 4.12. Let $f \in \Sigma^+$ and, for each $n \in \mathbb{N}$, put $E_{n,i} = \{i2^{-n} \leq f < (i+1)2^{-n}\}$ $(i \in I_n := \{0, \ldots, n2^n - 1\})$, similar to the sets in the proof of Theorem 3.6. Put also $E_n = \{f \geq n\}$. Note that the sets $E_{n,i}$ and E_n are in Σ. Define

$$f_n = \sum_{i \in I_n} i2^{-n} \mathbf{1}_{E_{n,i}} + n\mathbf{1}_{E_n}.$$

These f_n form an increasing sequence in Σ^+, even in $\mathfrak{S}^+$, with limit f. Theorem 4.12 yields $\mu(f_n) \uparrow \mu(f)$. We have exhibited a sequence of simple functions with limit f, that can be used to approximate $\mu(f)$.

Proposition 4.14 *Let $f, g \in \Sigma^+$ and $\alpha, \beta > 0$. Then $\mu(\alpha f + \beta g) = \alpha\mu(f) + \beta\mu(g) \leq \infty$.*

Proof Exercise 4.2. □

We proceed with the next limit result, known as *Fatou's lemma*.

Lemma 4.15 *Let (f_n) be an arbitrary sequence in Σ^+. Then $\liminf \mu(f_n) \geq \mu(\liminf f_n)$. If there exists a function $h \in \Sigma^+$ such that $f_n \leq h$ a.e., and $\mu(h) < \infty$, then $\limsup \mu(f_n) \leq \mu(\limsup f_n)$.*

Proof Put $g_n = \inf_{m \geq n} f_m$. We have for all $m \geq n$ the inequality $g_n \leq f_m$. Then also $\mu(g_n) \leq \mu(f_m)$ for $m \geq n$, and even $\mu(g_n) \leq \inf_{m \geq n} \mu(f_m)$. We want to take limits on both side of this inequality. On the right hand side we get $\liminf \mu(f_n)$. The sequence (g_n) is increasing, with limit $g = \liminf f_n$, and by Theorem 4.12, $\mu(g_n) \uparrow \mu(\liminf f_n)$ on the left hand side. This proves the first assertion. The second assertion follows by considering $\bar{f}_n = h - f_n \geq 0$. Check where it is used that $\mu(h) < \infty$. □

Remark 4.16 Let (E_n) be a sequence of sets in Σ, and let $f_n = \mathbf{1}_{E_n}$ and $h = 1$. The statements of Exercise 1.4 follow from Lemma 4.15.

We now extend the notion of integral to (almost) arbitrary measurable functions. Let $f \in \Sigma$. For (extended) real numbers x one defines $x^+ = \max\{x, 0\}$ and $x^- = \max\{-x, 0\}$. Then, for $f : S \to [-\infty, \infty]$, one defines the functions f^+ and f^- by $f^+(s) = f(s)^+$ and $f^-(s) = f(s)^-$. Notice that $f = f^+ - f^-$ and $|f| = f^+ + f^-$. If $f \in \Sigma$, then $f^+, f^- \in \Sigma^+$.

Definition 4.17 Let $f \in \Sigma$ and assume that $\mu(f^+) < \infty$ or $\mu(f^-) < \infty$. Then we define $\mu(f) := \mu(f^+) - \mu(f^-)$. If both $\mu(f^+) < \infty$ and $\mu(f^-) < \infty$, we say that f is *integrable*. The collection of all integrable functions is denoted by $\mathcal{L}^1(S, \Sigma, \mu)$. Note that $f \in \mathcal{L}^1(S, \Sigma, \mu)$ implies that $|f| < \infty$ μ-a.e.

Proposition 4.18 *The following natural properties hold.*

(i) Let $f, g \in \mathcal{L}^1(S, \Sigma, \mu)$ and $\alpha, \beta \in \mathbb{R}$. Then $\alpha f + \beta g \in \mathcal{L}^1(S, \Sigma, \mu)$ and $\mu(\alpha f + \beta g) = \alpha\mu(f) + \beta\mu(g)$. Hence μ can be seen as a linear operator on $\mathcal{L}^1(S, \Sigma, \mu)$.

(ii) If $f, g \in \mathcal{L}^1(S, \Sigma, \mu)$ and $f \leq g$ a.e., then $\mu(f) \leq \mu(g)$.

(iii) Triangle inequality: If $f \in \mathcal{L}^1(S, \Sigma, \mu)$, then $|\mu(f)| \leq \mu(|f|)$.

Proof Exercise 4.3. □

The next theorem is known as the Dominated Convergence Theorem, also called Lebesgue's Convergence Theorem.

Theorem 4.19 *Let $(f_n) \subset \Sigma$ and $f \in \Sigma$. Assume that $f_n(s) \to f(s)$ for all s outside a set of measure zero. Assume also that there exists a function $g \in \Sigma^+$ such that $\sup_n |f_n| \leq g$ a.e. and that $\mu(g) < \infty$. Then $\mu(|f_n - f|) \to 0$, and hence $\mu(f_n) \to \mu(f)$.*

Proof The second assertion easily follows from the first one, which we prove now for the case that $f_n \to f$ *everywhere*. One has the inequality $|f| \leq g$, whence $|f_n - f| \leq 2g$. The second assertion of Fatou's lemma immediately yields $\limsup \mu(|f_n - f|) \leq 0$, which is what we wanted. The *almost everywhere* version is left as Exercise 4.4. □

The convergence $\mu(|f_n - f|) \to 0$ is often denoted by $f_n \xrightarrow{\mathcal{L}^1} f$. The following result is known as Scheffé's lemma.

Lemma 4.20 *Let $(f_n) \subset \Sigma^+$ and assume that $f_n \to f$ a.e. Assume that $\mu(f_n)$ is finite for all n and $\mu(f) < \infty$ as well. Then $\mu(|f_n - f|) \to 0$ iff $\mu(f_n) \to \mu(f)$.*

Proof The 'only if' part follows from Theorem 4.19. Assume then that $\mu(f_n) \to \mu(f)$. We have the elementary equality $|f_n - f| = (f_n - f) + 2(f_n - f)^-$, and hence $\mu(|f_n - f|) = (\mu(f_n) - \mu(f)) + 2\mu((f_n - f)^-)$. The first term on the right hand side of the last expression tends to zero by assumption. The second one we treat as follows. Since $f - f_n \leq f$ and $f \geq 0$, it follows that $(f_n - f)^- \leq f$. Hence $\mu((f_n - f)^-) \to 0$, by virtue of Theorem 4.19. □

Example 4.21 Let $(S, \Sigma, \mu) = ([0, 1], \mathcal{B}([0, 1]), \lambda)$, where λ is Lebesgue measure. Assume that $f \in C[0, 1]$. Exercise 4.6 yields that $f \in \mathcal{L}^1([0, 1], \mathcal{B}([0, 1]), \lambda)$ and that $\lambda(f)$ is equal to the Riemann integral $\int_0^1 f(x)\,\mathrm{d}x$. This implication fails to hold if we replace $[0, 1]$ with an unbounded interval, see Exercise 4.7.

On the other hand, one can even show that *every function that is Riemann integrable* over $[0, 1]$, not only a continuous function, is Lebesgue integrable too. Knowledge of Chapter 2 is required for a precise statement and its proof, see Exercise 4.14.

Many results in integration theory can be proved by what is sometimes called the *standard machine*. This 'machine' works along the following steps. First one shows that results hold true for an indicator function, then one extends this by a linearity argument to nonnegative simple functions. Invoking the Monotone Convergence Theorem, one can then prove the results for nonnegative measurable functions. In the final step one shows the result to be true for functions in $\mathcal{L}^1(S, \Sigma, \mu)$ by splitting into positive and negative parts.

4.3 Integrals over subsets

This section is in a sense a prelude to the theorem of Radon-Nikodym, Theorem 6.10. Let $f \in \Sigma^+$ and $E \in \Sigma$. Then we may define

$$\int_E f \, \mathrm{d}\mu := \mu(\mathbf{1}_E f). \tag{4.6}$$

An alternative approach is to look at the *measurable* space (E, Σ_E), where $\Sigma_E = \{E \cap F : F \in \Sigma\}$ (check that this a σ-algebra on E). Denote the restriction of μ to Σ_E by μ_E. Then (E, Σ_E, μ_E) is a measure space. We consider integration on this space.

Proposition 4.22 *Let $f \in \Sigma$ and denote by f_E its restriction to E. Then $f_E \in \mathcal{L}^1(E, \Sigma_E, \mu_E)$ iff $\mathbf{1}_E f \in \mathcal{L}^1(S, \Sigma, \mu)$, in which case the identity $\mu_E(f_E) = \mu(\mathbf{1}_E f)$ holds.*

Proof Exercise 4.8. □

Let $f \in \Sigma^+$. Define for all $E \in \Sigma$

$$\nu(E) = \int_E f \, \mathrm{d}\mu \, (= \mu(\mathbf{1}_E f)). \tag{4.7}$$

One verifies (Exercise 4.9) that ν is a measure on (S, Σ). We want to compute $\nu(h)$ for $h \in \Sigma^+$. For measurable indicator functions we have by definition that the *integral* $\nu(\mathbf{1}_E)$ equals $\nu(E)$, which is equal to $\mu(\mathbf{1}_E f)$ by (4.7). More generally we have

Proposition 4.23 *Let $f \in \Sigma^+$ and $h \in \Sigma$. Then $h \in \mathcal{L}^1(S, \Sigma, \nu)$ iff $hf \in \mathcal{L}^1(S, \Sigma, \mu)$, in which case one has $\nu(h) = \mu(hf)$.*

Proof Exercise 4.10. □

For the measure ν above, Proposition 4.23 states that $\int h \, \mathrm{d}\nu = \int hf \, \mathrm{d}\mu$, valid for all $h \in \mathcal{L}^1(S, \Sigma, \nu)$. The notation $f = \frac{\mathrm{d}\nu}{\mathrm{d}\mu}$ is often used and looks like a derivative. We will return to this in Chapter 6, where we discuss Radon-Nikodym derivatives. The equality $\int h \, \mathrm{d}\nu = \int hf \, \mathrm{d}\mu$ now takes the appealing form

$$\int h \, \mathrm{d}\nu = \int h \frac{\mathrm{d}\nu}{\mathrm{d}\mu} \, \mathrm{d}\mu.$$

Example 4.24 Let $(S, \Sigma, \mu) = (\mathbb{R}, \mathcal{B}, \lambda)$, $f \geq 0$, Borel measurable, $\nu(E) = \int \mathbf{1}_E f \, \mathrm{d}\lambda$ and $hf \in \mathcal{L}^1(\mathbb{R}, \mathcal{B}, \lambda)$. Then

$$\nu(h) = \int_{-\infty}^{\infty} h(x) f(x) \, \mathrm{d}x,$$

where the equality is valid under conditions as for instance in Example 4.21 or, in general, as in Exercise 4.14.

Remark 4.25 If f is continuous, see Example 4.21, then $x \mapsto F(x) = \int_{[0,x]} f \, \mathrm{d}\lambda$ defines a differentiable function on $(0, 1)$, with $F'(x) = f(x)$. This follows from the theory of Riemann integrals. We adopt the conventional notation $F(x) = \int_0^x f(u) \, \mathrm{d}u$. This case can be generalized as follows. If $f \in \mathcal{L}^1(\mathbb{R}, \mathcal{B}, \lambda)$, then (using a similar notational convention) $x \mapsto F(x) = \int_{-\infty}^x f(u) \, \mathrm{d}u$ is well defined for all $x \in \mathbb{R}$. Moreover, F is at (Lebesgue) almost all points x of $\mathbb{R}$ differentiable with derivative $F'(x) = f(x)$. The proof of this result, the *fundamental theorem of calculus for the Lebesgue integral*, will be given in Section 6.7.

4.4 Expectation and integral

The whole point of this section is that the expectation of a random variable is a Lebesgue integral. Indeed, consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and let X be a (real) random variable defined on it. Recall that $X : \Omega \to \mathbb{R}$ is by definition a measurable function. Making the switch between the notations (S, Σ, μ) and $(\Omega, \mathcal{F}, \mathbb{P})$, one has the following notation for the integral of X w.r.t. $\mathbb{P}$

$$\mathbb{P}(X) = \int_\Omega X \, \mathrm{d}\mathbb{P}, \tag{4.8}$$

provided that the integral is well defined, which is certainly the case if $\mathbb{P}(|X|) < \infty$. Other often used notations for this integral are $\mathbb{P}X$ and $\mathbb{E}X$. The latter is the favorite one among probabilists and one speaks of the Expectation of X. Note also that $\mathbb{E}X$ is always defined when $X \geq 0$ *almost surely*. The latter concept meaning almost everywhere w.r.t. the probability measure $\mathbb{P}$. We abbreviate almost surely by a.s.

Example 4.26 Let $(\Omega, \mathcal{F}, \mathbb{P}) = (\mathbb{N}, 2^{\mathbb{N}}, \mathbb{P})$, where $\mathbb{P}$ is defined by $\mathbb{P}(n) = p_n$, where all $p_n \geq 0$ and $\sum p_n = 1$. Let (x_n) be a sequence of nonnegative real numbers and define the random variable X by $X(n) = x_n$. In a spirit similar to what we have seen in Examples 4.5 and 4.10, we get $\mathbb{E}X = \sum_{n=1}^\infty x_n p_n$. Let us switch to a different approach. Let $\xi_1, \xi_2, \ldots$ be the different elements of the set $\{x_1, x_2, \ldots\}$ and put $E_i = \{j : x_j = \xi_i\}$, $i \in \mathbb{N}$. Notice that $\{X = \xi_i\} = E_i$ and that the E_i form a partition of $\mathbb{N}$ with $\mathbb{P}(E_i) = \sum_{j \in E_i} p_j$. It follows that $\mathbb{E}X = \sum_i \xi_i \mathbb{P}(E_i)$, or $\mathbb{E}X = \sum_i \xi_i \mathbb{P}(X = \xi_i)$, the familiar expression for the expectation.

If $h : \mathbb{R} \to \mathbb{R}$ is Borel measurable, then $Y := h \circ X$ (we also write $Y = h(X)$) is a random variable as well. We give two recipes to compute $\mathbb{E}Y$. One is of course the direct application of the definition of expectation to Y, $\mathbb{E}Y = \int_\Omega Y \, \mathrm{d}\mathbb{P} = \int_\Omega h(X) \, \mathrm{d}\mathbb{P}$, but we also have the next Proposition 4.27 that tells us that one can also compute $\mathbb{E}Y$ as $\mathbb{E}Y = \int_{\mathbb{R}} y \, \mathbb{P}^Y(\mathrm{d}y) = \int_{\mathbb{R}} h(x) \, \mathbb{P}^X(\mathrm{d}x)$.

Proposition 4.27 *Let X be a random variable, and $h : \mathbb{R} \to \mathbb{R}$ Borel measurable. Let $\mathbb{P}^X$ be the distribution of X. Then $h \circ X \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ iff $h \in \mathcal{L}^1(\mathbb{R}, \mathcal{B}, \mathbb{P}^X)$, in which case*

$$\mathbb{E}h(X) = \int_{\mathbb{R}} h \, \mathrm{d}\mathbb{P}^X. \tag{4.9}$$

Proof Exercise 4.11. □

Remark 4.28 Proposition 4.27 is in terms of equality of two integrals defined on different probability spaces. There is nothing essential here in working with probability measures. Indeed, consider a more general set up, where we have a measure space (S, Σ, μ) and another measurable space (S', Σ'). Let $f : S \to S'$ be a measurable mapping and $h : S' \to \mathbb{R}$ be Borel-measurable, then also the composition $h(f) = h \circ f : S \to \mathbb{R}$ is Borel-measurable. Moreover, we can define the measure μ^f (also called push-forward measure) on Σ' by $\mu^f(E') := \mu(f^{-1}(E'))$, $E' \in \Sigma'$. Then $h \circ f \in \mathcal{L}^1(S, \Sigma, \mu)$ iff $h \in (S', \Sigma', \mu^f)$, in which case

$$\int_S h(f) \, \mathrm{d}\mu = \int_{S'} h \, \mathrm{d}\mu^f. \tag{4.10}$$

Note that the distribution of a random variable is a push-forward measure. Equation (4.9) is known as the transformation formula for expectations, and its generalized version (4.10) for arbitrary measures as the transformation formula for integrals.

Example 4.29 Suppose there exists $f \geq 0$, Borel-measurable such that for all $B \in \mathcal{B}$ one has $\mathbb{P}^X(B) = \lambda(\mathbf{1}_B f)$, in which case it is said that X has a density f. Then, provided that the expectation is well defined, Example 4.24 and Proposition 4.27 yield

$$\mathbb{E}h(X) = \int_{\mathbb{R}} h(x) f(x) \, \mathrm{d}x,$$

another familiar formula for the expectation of $h(X)$.

We conclude that the definition of expectation as a Lebesgue integral w.r.t. a probability measure as in (4.8) yields the familiar formulas, sums for discrete random variables and Riemann integrals for random variables having an ordinary density function, as special cases. So, we see that the Lebesgue integral serves as a unifying concept for expectation. At least as important is that we can use the powerful convergence theorems (obtained for integrals) of Section 4.2

for expectations as well. Notice that every real constant (function) has a well defined, and trivially finite, expectation. Therefore one can in pertaining cases apply the Dominated Convergence Theorem (Theorem 4.19) with the function g equal to a constant. Here is a simple example of the application of the Monotone Convergence Theorem.

Example 4.30 Let (X_n) be a sequence of nonnegative random variables, so all $\mathbb{E}X_n \leq \infty$ are well defined. Then $\sum X_n$ is a well defined random variable as well, nonnegative, and we have $\mathbb{E}(\sum X_n) = \sum \mathbb{E}X_n$. Moreover if $\sum \mathbb{E}X_n < \infty$, then $\sum X_n < \infty$ a.s. Verification of these assertions is straightforward and left as Exercise 4.12.

The next two propositions have proven to be very useful in proofs of results in Probability Theory. We use the fact that $\mathbb{P}$ is a *probability* measure in an essential way.

Proposition 4.31 *Let X be a real valued random variable and $g : \mathbb{R} \to [0, \infty]$ an increasing function. Then $\mathbb{E}g(X) \geq g(c)\mathbb{P}(X \geq c)$.*

Proof This follows from the inequality $g(X)\mathbf{1}_{\{X \geq c\}} \geq g(c)\mathbf{1}_{\{X \geq c\}}$. □

The inequality in Proposition 4.31 is known as Markov's inequality. An example is obtained by taking $g(x) = x^+$ and by replacing X with $|X|$. One gets $\mathbb{E}|X| \geq c\mathbb{P}(|X| \geq c)$. For the special case where $g(x) = (x^+)^2$, it is known as Chebychev's inequality. This name is especially used, if we apply it with $|X - \mathbb{E}X|$ instead of X. For $c \geq 0$ we then obtain $\operatorname{Var} X \geq c^2\mathbb{P}(|X - \mathbb{E}X| \geq c)$.

We now turn to a result that is known as *Jensen's inequality*, Proposition 4.32 below. Recall that a function $g : G \to \mathbb{R}$ is convex, if G is a convex set and if for all $x, y \in G$ and $\alpha \in [0, 1]$ one has

$$g(\alpha x + (1 - \alpha)y) \leq \alpha g(x) + (1 - \alpha)g(y).$$

We consider only the case where G is an interval. Let us first give some properties of convex functions. Let $x, y, z \in G$ and $x < y < z$. Then $y = \alpha x + (1-\alpha)z$, with $\alpha = \frac{z-y}{z-x}$, and from convexity of g we get

$$(g(y) - g(x))(z - x) \leq (g(z) - g(x))(y - x).$$

By taking the appropriate limits, one obtains that g is continuous on $\operatorname{Int} G$, the interior of G. Take $x \in \operatorname{Int} G$ and rewrite the above inequality as

$$\frac{g(y) - g(x)}{y - x} \leq \frac{g(z) - g(x)}{z - x}. \tag{4.11}$$

It follows that the right derivative $D_+g(x) := \lim_{y \downarrow x} \frac{g(y)-g(x)}{y-x}$ exists and is finite. In a similar way one can show that the left derivative $D_-g(x) := \lim_{y \uparrow x} \frac{g(y)-g(x)}{y-x}$ exists and is finite. Moreover, one has $D_+g(x) \geq D_-g(x)$ for all $x \in \operatorname{Int} G$ and

both onesided derivatives are increasing. If one takes in (4.11) the limit when $y \downarrow x$, one gets the inequality $(z-x)D_+g(x) \leq g(z) - g(x)$, valid for $z > x$. Likewise one has the inequality $(x-z)D_-g(x) \geq g(x) - g(z)$, valid for $z < x$. It follows that for any $d(x) \in [D_-g(x), D_+g(x)]$, it holds that

$$g(z) - g(x) \geq d(x)(z-x). \tag{4.12}$$

The $d(x)$ are also called subgradients of g. The following proposition (Jensen's inequality) is now easy to prove.

Proposition 4.32 *Let $g : G \to \mathbb{R}$ be convex and X a random variable with $\mathbb{P}(X \in G) = 1$. Assume that $\mathbb{E}|X| < \infty$ and $\mathbb{E}|g(X)| < \infty$. Then*

$$\mathbb{E}g(X) \geq g(\mathbb{E}X).$$

Proof We exclude the trivial case $\mathbb{P}(X = x_0) = 1$ for some $x_0 \in G$. Since $\mathbb{P}(X \in G) = 1$, we have $\mathbb{E}X \in \operatorname{Int} G$ (Exercise 4.18) and (4.12) with $x = \mathbb{E}X$ and z replaced with X holds a.s. So, in view of (4.12),

$$g(X) - g(\mathbb{E}X) \geq d(\mathbb{E}X)(X - \mathbb{E}X).$$

Take expectations to get $\mathbb{E}g(X) - g(\mathbb{E}X) \geq 0$. □

4.5 Functions of bounded variation and Stieltjes integrals

In this section we define functions of bounded variation and review some basic properties. Stieltjes integrals will be discussed subsequently. We consider functions defined on an interval $[a, b]$. Next to these we consider partitions Π of $[a, b]$, finite subsets $\{t_0, \ldots, t_n\}$ of $[a, b]$ with the convention $t_0 \leq \cdots \leq t_n$, and $\mu(\Pi)$ denotes the mesh of Π. Extended partitions, denoted Π^*, are partitions Π, together with additional points τ_i, with $t_{i-1} \leq \tau_i \leq t_i$. By definition $\mu(\Pi^*) = \mu(\Pi)$. Along with a function α, a partition Π, we define

$$V^1(\alpha; \Pi) := \sum_{i=1}^{n} |\alpha(t_i) - \alpha(t_{i-1})|,$$

the variation of α over the partition Π.

Definition 4.33 A function α is said to be of bounded variation if $V^1(\alpha) := \sup_\Pi V^1(\alpha; \Pi) < \infty$, the supremum taken over all partitions Π. The variation function $v_\alpha : [a, b] \to \mathbb{R}$ is defined by $v_\alpha(t) = V^1(\alpha \mathbf{1}_{[a,t]})$.

A refinement Π' of a partition Π satisfies by definition the inclusion $\Pi \subset \Pi'$. In such a case, one has $\mu(\Pi') \leq \mu(\Pi)$ and $V^1(\alpha; \Pi') \geq V^1(\alpha; \Pi)$. It follows from the definition of $V^1(\alpha)$, that there exists a sequence (Π_n) of partitions (which can be taken as successive refinements) such that $V^1(\alpha; \Pi_n) \to V^1(\alpha)$.

Example 4.34 Let α be continuously differentiable and assume $\int_a^b |\alpha'(t)|\,\mathrm{d}t$ finite. Then $V^1(\alpha) = \int_a^b |\alpha'(t)|\,\mathrm{d}t$. This follows, since $V^1(\alpha;\Pi)$ can be written as a Riemann sum

$$\sum_{i=1}^{n} |\alpha'(\tau_i)|\,(t_i - t_{i-1}),$$

where the τ_i satisfy $t_{i-1} \le \tau_i \le t_i$ and $\alpha'(\tau_i) = \frac{\alpha(t_i)-\alpha(t_{i-1})}{t_i - t_{i-1}}$.

Note that v_α is an increasing function with $v_\alpha(a) = 0$ and $v_\alpha(b) = V^1(\alpha)$. Any monotone function α is of bounded variation and in this case $V^1(\alpha) = |\alpha(b) - \alpha(a)|$ and $v_\alpha(t) = |\alpha(t) - \alpha(a)|$. Also the difference of two increasing functions is of bounded variation. This fact has a converse.

Proposition 4.35 *Let α be of bounded variation. Then there exists increasing functions v_α^+ and v_α^- such that $v_\alpha^+(a) = v_\alpha^-(a) = 0$, $\alpha(t) - \alpha(a) = v_\alpha^+(t) - v_\alpha^-(t)$. Moreover, one can choose them such that $v_\alpha^+ + v_\alpha^- = v_\alpha$.*

Proof Define

$$v_\alpha^+(t) = \frac{1}{2}(v_\alpha(t) + \alpha(t) - \alpha(a))$$
$$v_\alpha^-(t) = \frac{1}{2}(v_\alpha(t) - \alpha(t) + \alpha(a)).$$

We only have to check that these functions are increasing, since the other statements are obvious. Let $t' > t$. Then $v_\alpha^+(t') - v_\alpha^+(t) = \frac{1}{2}(v_\alpha(t') - v_\alpha(t) + \alpha(t') - \alpha(t))$. The difference $v_\alpha^+(t') - v_\alpha^+(t)$ is the variation of α over the interval $[t, t']$, which is greater than or equal to $|\alpha(t') - \alpha(t)|$. Hence $v_\alpha^+(t') - v_\alpha^+(t) \ge 0$, and the same holds for $v_\alpha^-(t') - v_\alpha^-(t)$. □

The decomposition in this proposition enjoys a minimality property. If w^+ and w_- are increasing functions, $w^+(a) = w^-(a) = 0$ and $\alpha(t) - \alpha(a) = w^+(t) - w^-(t)$, then for all $t' > t$ one has $w^+(t') - w^+(t) \ge v_\alpha^+(t') - v_\alpha^+(t)$ and $w^-(t') - w^-(t) \ge v_\alpha^-(t') - v_\alpha^-(t)$. This property is basically the counterpart of the Jordan decomposition (6.4) of signed measures.

The following definition generalizes the concept of Riemann integral.

Definition 4.36 Let $f, \alpha : [a, b] \to \mathbb{R}$ and Π^* be an extended partition of $[a, b]$. Write

$$S(f, \alpha; \Pi^*) = \sum_{i=1}^{n} f(\tau_i)\,(\alpha(t_i) - \alpha(t_{i-1})).$$

We say that $S(f, \alpha) = \lim_{\mu(\Pi^*)\to 0} S(f, a; \Pi^*)$, if for all $\varepsilon > 0$, there exists $\delta > 0$ such that $\mu(\Pi^*) < \delta$ implies $|S(f, \alpha) - S(f, \alpha; \Pi^*)| < \varepsilon$. If this happens, we say that f is integrable w.r.t. α and we commonly write $\int f\,\mathrm{d}\alpha$ for $S(f, \alpha)$, and call it the Stieltjes integral of f w.r.t. α.

Proposition 4.37 *Let $f, \alpha : [a,b] \to \mathbb{R}$, f continuous and α of bounded variation. Then f is integrable w.r.t. α. Moreover, the triangle inequality $|\int f \,\mathrm{d}\alpha| \leq \int |f| \,\mathrm{d}v_\alpha$ holds.*

Proof To show integrability of f w.r.t. α, the idea is to compare $S(f,\alpha;\Pi_1^*)$ and $S(f,\alpha;\Pi_2^*)$ for two extended partitions Π_1^* and Π_2^*. By constructing another extended partition Π^* that is a *refinement* of Π_1^* and Π_2^* in the sense that all t_i and t_i from Π_1^* and Π_2^* belong to Π^*, it follows from

$$|S(f,\alpha;\Pi_1^*) - S(f,\alpha;\Pi_2^*)| \leq |S(f,\alpha;\Pi_1^*) - S(f,\alpha;\Pi^*)| + |S(f,\alpha;\Pi^*) - S(f,\alpha;\Pi_2^*)|$$

that it suffices to show that $|S(f,\alpha;\Pi_1^*) - S(f,\alpha;\Pi_2^*)|$ can be made small for Π_2 a refinement of Π_1. Let $\varepsilon > 0$ and choose $\delta > 0$ such that $|f(t) - f(s)| < \varepsilon$ whenever $|t-s| < \delta$ (possible by uniform continuity of f). Assume that $\mu(\Pi_1) < \delta$, then also $\mu(\Pi_2) < \delta$. Consider first two extended partitions of Π_1, Π_1^* and Π_1', the latter with intermediate points τ_i'. Then

$$\begin{aligned} |S(f,\alpha;\Pi_1^*) - S(f,\alpha;\Pi_1')| &\leq \sum_i |f(\tau_i) - f(\tau_i')||\alpha(t_i) - \alpha(t_{i-1})| \\ &\leq \varepsilon \, V^1(\alpha;\Pi_1) \leq \varepsilon \, V^1(\alpha). \end{aligned}$$

In the next step we assume that Π_2 is obtained from Π_1 by adding one point, namely τ_j for some τ_j from the extended partition Π_1^*. Further we assume that Π_2^* contains all the intermediate points τ_i from Π_1^*, whereas we also take the intermediate points from the intervals $[t_{j-1}, \tau_j]$ and $[\tau_j, t_j]$ both equal to τ_j. It follows that $S(f,\alpha;\Pi_1^*) = S(f,\alpha;\Pi_2^*)$. A combination of the two steps finishes the proof. The triangle inequality for the integrals holds almost trivially. □

Proposition 4.38 *Let $f, \alpha : [a,b] \to \mathbb{R}$, be continuous and of bounded variation. Then the following integration by parts formula holds.*

$$\int f \,\mathrm{d}\alpha + \int \alpha \,\mathrm{d}f = f(b)\alpha(b) - f(a)\alpha(a).$$

Proof Choose points $t_a \leq \cdots \leq t_n$ in $[a,b]$ and $\tau_i \in [t_{i-1}, t_i]$, $i = 1, \ldots, n$, to which we add $\tau_0 = a$ and $\tau_{n+1} = b$. By Abel's summation formula, we have

$$\begin{aligned} &\sum_{i=1}^{n} f(\tau_i)\,(\alpha(t_i) - \alpha(t_{i-1})) = \\ &\quad f(b)\alpha(b) - f(a)\alpha(a) - \sum_{i=1}^{n+1} \alpha(t_{i-1})\,(f(\tau_i) - f(\tau_{i-1})). \end{aligned}$$

The result follows by application of Proposition 4.37. □

This proposition can be used to *define* $\int \alpha \,\mathrm{d}f$ for functions α of bounded variation and continuous functions f, simply by putting

$$\int \alpha \,\mathrm{d}f := f(b)\alpha(b) - f(a)\alpha(a) - \int f \,\mathrm{d}\alpha.$$

Next we give an example that illustrates that the continuity assumption on f in Proposition 4.37 cannot be omitted in general.

Example 4.39 Let $\alpha : [-1,1] \to \mathbb{R}$ be given by $\alpha(t) = \mathbf{1}_{[0,1]}(t)$. Let $f : [-1,1] \to \mathbb{R}$ be discontinuous at zero, for instance $f = \alpha$; note that f and α share a point of discontinuity. Let Π be a partition of $[-1,1]$ whose elements are numbered in such a way that $t_{m-1} \leq 0 \leq t_m$, let $\tau = \tau_m \in [t_{m-1}, t_m]$. Then $S(f, \alpha; \Pi^*) = f(\tau)$, and this doesn't converge for $\tau \to 0$.

In the annoying case of the above example, the Stieltjes integral is not defined. By adjusting the concept of Stieltjes integral into the direction of a Lebesgue integral, we obtain that also in this case the integral is well defined.

It follows from Proposition 4.35 that α admits finite left and right limits at all t in $[a,b]$. By α_+ we denote the function given by $\alpha_+(t) = \lim_{u \downarrow t} \alpha(u)$ for $t \in [a,b)$ and $\alpha_+(b) = \alpha(b)$. Note that α_+ is right-continuous. Let $\mu = \mu_\alpha$ be the *signed measure* (see also Section 6.3) on $([a,b], \mathcal{B}([a,b]))$ that is uniquely defined by $\mu((s,t]) = \alpha_+(t) - \alpha_+(s)$ for all $a \leq s < t \leq b$. For measurable f one can consider the Lebesgue integral $\int f \, \mathrm{d}\mu_\alpha$. We now present a result relating Stieltjes and Lebesgue integrals. The proposition below gives an example of such a connection, but can be substantially generalized.

Proposition 4.40 *Let $f, \alpha : [a,b] \to \mathbb{R}$, f continuous and α of bounded variation. Then the Lebesgue integral $\int f \, \mathrm{d}\mu_\alpha$ and the Stieltjes integral $\int f \, \mathrm{d}\alpha$ are equal, $\int f \, \mathrm{d}\mu_\alpha = \int f \, \mathrm{d}\alpha$.*

Proof Exercise 4.17. □

For $\mu = \mu_\alpha$ as above, we call the integral $\int f \, \mathrm{d}\alpha$, defined as $\int f \, \mathrm{d}\mu_\alpha$, the Lebesgue-Stieltjes integral of f w.r.t. α. Consider again Example 4.39. Now the Lebesgue-Stieltjes integral $\int \alpha \, \mathrm{d}\alpha$ is well defined and $\int \alpha \, \mathrm{d}\alpha = 1$.

4.6 $\mathcal{L}^p$-spaces of random variables

In this section we introduce the p-norms and the spaces of random variables with finite p-norm. We start with a definition.

Definition 4.41 Let $1 \leq p < \infty$ and X a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$. If $\mathbb{E}|X|^p < \infty$, we write $X \in \mathcal{L}^p(\Omega, \mathcal{F}, \mathbb{P})$ and $||X||_p = (\mathbb{E}|X|^p)^{1/p}$.

The notation $||\cdot||$ suggests that we deal with a norm. In a sense, this is correct, but we will not prove until the end of this section. It is however obvious that $\mathcal{L}^p := \mathcal{L}^p(\Omega, \mathcal{F}, \mathbb{P})$ is a vector space, since $|X+Y|^p \leq (|X| + |Y|)^p \leq 2^p(|X|^p + |Y|^p)$.

In the special case $p = 2$, we have for $X, Y \in \mathcal{L}^2$, that $|XY| = \frac{1}{2}((|X|+|Y|)^2 - X^2 - Y^2)$ has finite expectation and is thus in $\mathcal{L}^1$. Of course we have $|\mathbb{E}(XY)| \leq \mathbb{E}|XY|$. For the latter we have the famous Cauchy-Schwarz inequality.

Proposition 4.42 *Let $X, Y \in \mathcal{L}^2$. Then $XY \in \mathcal{L}^1$ and $\mathbb{E}|XY| \leq ||X||_2 \, ||Y||_2$.*

Proof If $\mathbb{E}Y^2 = 0$, then $Y = 0$ a.s. (Lemma 4.11), so also $XY = 0$ a.s. and there is nothing to prove. Assume then that $\mathbb{E}Y^2 > 0$ and let $c = \mathbb{E}|XY|/\mathbb{E}Y^2$. One trivially has $\mathbb{E}(|X| - c|Y|)^2 \geq 0$. But the left hand side equals $\mathbb{E}X^2 - \frac{(\mathbb{E}|XY|)^2}{\mathbb{E}Y^2}$. □

Proposition 4.42 tells us that $X, Y \in \mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$ is sufficient to guarantee that the product XY is integrable. For independent X and Y weaker integrability assumptions suffice and the product rule for probabilities of intersections extends to a product rule for expectations.

Proposition 4.43 *Let $X, Y \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ be independent random variables. Then $XY \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ and $\mathbb{E}(XY) = \mathbb{E}X \cdot \mathbb{E}Y$.*

Proof The standard machine easily gives $\mathbb{E}(\mathbf{1}_A Y) = \mathbb{P}(A) \cdot \mathbb{E}Y$ for A an event independent of Y. Assume that $X \in \mathfrak{S}^+$. Since X is integrable we can assume that it is finite, and thus bounded by a constant c. Since then $|XY| \leq c|Y|$, we obtain $\mathbb{E}|XY| < \infty$. If we represent X as $\sum_{i=1}^n a_i \mathbf{1}_{A_i}$, then $\mathbb{E}(XY) = \sum_{i=1}^n a_i \mathbb{P}(A_i)\mathbb{E}Y$ readily follows and thus $\mathbb{E}(XY) = \mathbb{E}X \cdot \mathbb{E}Y$. The proof may be finished by letting the standard machine operate on X. □

We continue with some properties of $\mathcal{L}^p$-spaces. First we have monotonicity of norms.

Proposition 4.44 *Let $1 \leq p \leq r$ and $X \in \mathcal{L}^r(\Omega, \mathcal{F}, \mathbb{P})$, then $X \in \mathcal{L}^p(\Omega, \mathcal{F}, \mathbb{P})$ and $||X||_p \leq ||X||_r$.*

Proof It follows from the trivial inequality $|u| \leq 1 + |u|^a$, valid for $u \in \mathbb{R}$ and $a \geq 1$, that $|X|^p \leq 1 + |X|^r$, by taking $a = r/p$, and hence $X \in \mathcal{L}^p(\Omega, \mathcal{F}, \mathbb{P})$. Observe that $x \to |x|^a$ is convex. We apply Jensen's inequality to get $(\mathbb{E}|X|^p)^a \leq \mathbb{E}(|X|^{pa})$, from which the result follows. □

4.7 $\mathcal{L}^p$-spaces of functions

In the previous section we have introduced the $\mathcal{L}^p$-spaces for random variables defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. In the present section, we consider in some more generality the spaces $\mathcal{L}^p(S, \Sigma, \mu)$. For completeness, we give the definition, which is of course completely analogous to Definition 4.41.

Definition 4.45 Let $1 \leq p < \infty$ and f a measurable function on (S, Σ, μ). If $\mu(|f|^p) < \infty$, we write $f \in \mathcal{L}^p(S, \Sigma, \mu)$ and $||f||_p = (\mu(|f|^p))^{1/p}$.

Occasionally, it is useful to work with $||f||_p$ for $p = \infty$. It is defined as follows. For $f \in \Sigma$ we put

$$||f||_\infty := \inf\{m \in \mathbb{R} : \mu(\{|f| > m\}) = 0\},$$

with the convention $\inf \emptyset = \infty$. It is clear that $|f| \leq ||f||_\infty$ a.e. We write $f \in \mathcal{L}^\infty(S, \Sigma, \mu)$ if $||f||_\infty < \infty$.

Here is the first of two fundamental inequalities, known as Hölder's inequality.

Theorem 4.46 *Let $p, q \in [1, \infty]$, $f \in \mathcal{L}^p(S, \Sigma, \mu)$ and $g \in \mathcal{L}^q(S, \Sigma, \mu)$. If $\frac{1}{p} + \frac{1}{q} = 1$, then $fg \in \mathcal{L}^1(S, \Sigma, \mu)$ and $||fg||_1 \leq ||f||_p ||g||_q$.*

Proof Notice first that for $p = 1$ or $p = \infty$ there is basically nothing to prove. So we assume $p, q \in (1, \infty)$. We give a probabilistic proof by introducing a conveniently chosen probability measure and by using Jensen's inequality. We assume without loss of generality that $f, g \geq 0$ a.e. If $||f||_p = 0$, then $f = 0$ a.e. in view of Lemma 4.11 and we have a trivial inequality. Let then $0 < ||f||_p < \infty$. We now define a probability measure $\mathbb{P}$ on Σ by

$$\mathbb{P}(E) = \frac{\mu(\mathbf{1}_E f^p)}{\mu(f^p)}.$$

Put $h(s) = g(s)/f(s)^{p-1}$ if $f(s) > 0$ and $h(s) = 0$ otherwise. Jensen's inequality gives $(\mathbb{P}(h))^q \leq \mathbb{P}(h^q)$. We compute

$$\mathbb{P}(h) = \frac{\mu(fg)}{\mu(f^p)},$$

and

$$\mathbb{P}(h^q) = \frac{\mu(\mathbf{1}_{\{f>0\}} g^q)}{\mu(f^p)} \leq \frac{\mu(g^q)}{\mu(f^p)}.$$

Insertion of these expressions into the above version of Jensen's inequality yields

$$\frac{(\mu(fg))^q}{(\mu(f^p))^q} \leq \frac{\mu(g^q)}{\mu(f^p)},$$

whence $(\mu(fg))^q \leq \mu(g^q)\mu(f^p)^{q-1}$. Take q-th roots on both sides and the result follows. □

Remark 4.47 For $p = 2$ Theorem 4.46 yields the Cauchy-Schwarz inequality $||fg||_1 \leq ||f||_2 ||g||_2$ for square integrable functions. Compare to Proposition 4.42.

We now give the second fundamental inequality, *Minkowski's inequality.*

Theorem 4.48 *Let $f, g \in \mathcal{L}^p(S, \Sigma, \mu)$ and $p \in [1, \infty]$. Then $||f + g||_p \leq ||f||_p + ||g||_p$.*

Proof The case $p = \infty$ is almost trivial, so we assume $p \in [1, \infty)$. To exclude another triviality, we suppose $||f + g||_p > 0$. Note the following elementary relations.

$$|f + g|^p = |f + g|^{p-1}|f + g| \leq |f + g|^{p-1}|f| + |f + g|^{p-1}|g|.$$

Now we take integrals and apply Hölder's inequality to obtain

$$\begin{aligned}
\mu(|f + g|^p) &\leq \mu(|f + g|^{p-1}|f|) + \mu(|f + g|^{p-1}|g|) \\
&\leq (||f||_p + ||g||_p)(\mu(|f + g|^{(p-1)q})^{1/q} \\
&= (||f||_p + ||g||_p)(\mu(|f + g|^p)^{1/q},
\end{aligned}$$

because $(p-1)q = p$. After dividing by $(\mu(|f+g|^p)^{1/q}$, we obtain the result, because $1 - 1/q = 1/p$. □

Recall the definition of a norm on a (real) vector space X. One should have $||x|| = 0$ iff $x = 0$, $||\alpha x|| = |\alpha|\,||x||$ for $\alpha \in \mathbb{R}$ (homogeneity) and $||x+y|| \leq ||x|| + ||y||$ (triangle inequality). For $||\cdot||_p$ homogeneity is obvious, the triangle inequality has just been proved under the name Minkowski's inequality and we also trivially have $f = 0 \Rightarrow ||f||_p = 0$. But, conversely $||f||_p = 0$ only implies $f = 0$ a.e. This annoying fact disturbs $||\cdot||_p$ being called a genuine norm. This problem can be circumvented by identifying a function f that is zero a.e. with the zero function. The proper mathematical way of doing this is by defining the *equivalence relation* $f \sim g$ iff $\mu(\{f \neq g\}) = 0$. By considering the equivalence classes induced by this equivalence relation one gets the quotient space $L^p(S, \Sigma, \mu) := \mathcal{L}^p(S, \Sigma, \mu)/\sim$. One can show that $||\cdot||_p$ induces a norm on this space in the obvious way. We don't care too much about these details and just call $||\cdot||_p$ a norm and $\mathcal{L}^p(S, \Sigma, \mu)$ a normed space, thereby violating a bit the standard mathematical language.

A desirable property of a normed space, (a version of) completeness, holds for $\mathcal{L}^p$ spaces. We give the proof for $\mathcal{L}^p(\Omega, \mathcal{F}, \mathbb{P})$.

Theorem 4.49 *Let $p \in [1, \infty]$. The space $\mathcal{L}^p(\Omega, \mathcal{F}, \mathbb{P})$ is* complete *in the following sense. Let (X_n) be a Cauchy-sequence in $\mathcal{L}^p$: $||X_n - X_m||_p \to 0$ for $n, m \to \infty$. Then there exists a* limit *$X \in \mathcal{L}^p$ such that $||X_n - X||_p \to 0$. The limit is unique in the sense that any other limit X' satisfies $||X - X'||_p = 0$.*

Proof Assume that $p \in [1, \infty)$. Since (X_n) is Cauchy, for all $n \in \mathbb{N}$ there exists $k_n \in \mathbb{N}$ such that $||X_l - X_m||_p \leq 2^{-n}$ if $l, m \geq k_n$. By monotonicity of the p-norms, we then have $||X_{k_{n+1}} - X_{k_n}||_1 \leq ||X_{k_{n+1}} - X_{k_n}||_p \leq 2^{-n}$. It follows (see Example 4.30) that $\mathbb{E}\sum_n |X_{k_{n+1}} - X_{k_n}| = \sum_n \mathbb{E}|X_{k_{n+1}} - X_{k_n}| < \infty$. But then $\sum_n |X_{k_{n+1}} - X_{k_n}| < \infty$ a.s., which implies $\sum_n (X_{k_{n+1}} - X_{k_n}) < \infty$ a.s. This is a telescopic sum, so we obtain that $\lim X_{k_n}$ exists and is finite a.s. To have a proper random variable, we define $X := \limsup X_{k_n}$ and we have $X_{k_n} \to X$ a.s.

We have to show that X is also a limit in $\mathcal{L}^p$. Recall the definition of the k_n. Take $m \geq n$ and $l \geq k_n$. Then $||X_l - X_{k_m}||_p \leq 2^{-n}$, or $\mathbb{E}|X_l - X_{k_m}|^p \leq 2^{-np}$. We use Fatou's lemma for $m \to \infty$ and get

$$\mathbb{E}|X_l - X|^p = \mathbb{E}\liminf |X_l - X_{k_m}|^p \leq \liminf \mathbb{E}|X_l - X_{k_m}|^p \leq 2^{-np}.$$

This shows two things. First $X_l - X \in \mathcal{L}^p$ and then, since $X_l \in \mathcal{L}^p$ also $X \in \mathcal{L}^p$. Secondly, $\limsup_{l\to\infty} \mathbb{E}|X_l - X|^p \leq 2^{-np}$, for all n. Hence $||X_l - X||_p \to 0$.

The proof for $p = \infty$ is left as Exercise 4.13. □

Remark 4.50 Notice that it follows from Theorem 4.49 and the discussion preceding it, that $L^p(\Omega, \mathcal{F}, \mathbb{P})$ is a truly complete normed space, a Banach space. The same is true for $L^p(S, \Sigma, \mu)$ ($p \in [0, \infty]$), for which you need Exercise 4.13. For the special case $p = 2$ we obtain that $L^2(S, \Sigma, \mu)$ is a Hilbert space with the inner product $\langle f, g\rangle := \int fg\,\mathrm{d}\mu$. Likewise $L^2(\Omega, \mathcal{F}, \mathbb{P})$ is a Hilbert space with inner product $\langle X, Y\rangle := \mathbb{E}XY$.

Let S be a vector space with inner product $\langle \cdot, \cdot \rangle$ and norm $|| \cdot ||$ defined by $||x|| = \langle x, x \rangle^{1/2}$. Recall that an orthogonal projection on a subspace L is a linear mapping $\pi : S \to L$ with the property that $\pi x = \arg\inf\{||x-y|| : y \in L\}$. The argument of the infimum exists if L is complete, which is partly contained in the following theorem.

Theorem 4.51 *Let S be a Hilbert space and L a closed subspace. Let $x \in S$. Then there exists a unique $\hat{x} \in L$ such that*

(i) $||x - \hat{x}|| = \inf\{||x - y|| : y \in L\}$.

Moreover, the element $\hat{x}$ satisfying (i) is characterized by each the following two properties.

(ii) $||x - y||^2 = ||x - \hat{x}||^2 + ||\hat{x} - y||^2$, $\forall y \in L$.

(iii) $x - \hat{x}$ *is orthogonal to* L, *i.e.* $\langle x - \hat{x}, y \rangle = 0$, $\forall y \in L$.

Proof Let α be the infimum in (i). For every $n \in \mathbb{N}$ there exists $x_n \in L$ such that $||x - x_n||^2 < \alpha^2 + 1/n$. Add up the two equalities

$$\begin{aligned} &||x - x_m + \frac{1}{2}(x_m - x_n)||^2 \\ &= ||x - x_m||^2 + ||\frac{1}{2}(x_m - x_n)||^2 + 2\langle x - x_m, \frac{1}{2}(x_m - x_n)\rangle \end{aligned}$$

and

$$\begin{aligned} &||x - x_n - \frac{1}{2}(x_m - x_n)||^2 \\ &= ||x - x_n||^2 + ||\frac{1}{2}(x_m - x_n)||^2 - 2\langle x - x_n, \frac{1}{2}(x_m - x_n)\rangle \end{aligned}$$

to get

$$2||x - \frac{1}{2}(x_m + x_n)||^2 = ||x - x_m||^2 + ||x - x_n||^2 - \frac{1}{2}||x_m - x_n||^2. \quad (4.13)$$

Since the left hand side of (4.13) is at least equal to $2\alpha^2$, one obtains by definition of x_n and x_m that $\frac{1}{2}||x_m - x_n||^2 \leq \frac{1}{m} + \frac{1}{n}$. Therefore (x_n) is a Cauchy sequence in L which by completeness has a limit, we call it $\hat{x}$. We now show that $\hat{x}$ attains the infimum. Since $||x - \hat{x}|| \leq ||x - x_n|| + ||x_n - \hat{x}||$, we get for $n \to \infty$ that $||x - \hat{x}|| \leq \alpha$ and hence we must have equality.

Suppose that there are two $\hat{x}_1, \hat{x}_2 \in L$ that attain the infimum. Let $\hat{x} = \frac{1}{2}(\hat{x}_1 + \hat{x}_2)$. Then $||x - \hat{x}|| \leq \frac{1}{2}(||x - \hat{x}_1|| + ||x - \hat{x}_2||) = \alpha$, so also $\hat{x}$ attains the infimum. Replace in (4.13) x_m with $\hat{x}_1$ and x_n with $\hat{x}_2$ to conclude that $||\hat{x}_1 - \hat{x}_2|| = 0$. Hence $\hat{x}_1 = \hat{x}_2$.

We now show that the three characterizations of $\hat{x}$ are equivalent. First we prove (i) $\Rightarrow$ (iii). Consider the quadratic function $f(t) = ||x - \hat{x} + ty||^2, t \in \mathbb{R}$, where $y \in L$ is arbitrary. The function f has minimum at $t = 0$. Computing $f(t)$ explicitly gives $f(t) = ||x - \hat{x}||^2 + 2t\langle x - \hat{x}, y \rangle + t^2||y||$. A minimum at $t = 0$ can only happen if the coefficient of t vanishes, so $\langle x - \hat{x}, y \rangle = 0$.

The implication (iii) $\Rightarrow$ (ii) follows from $||x-y||^2 = ||x-\hat{x}||^2 + 2\langle x-\hat{x}, \hat{x}-y\rangle + ||y-\hat{x}||^2$, since the cross term vanishes by assumption.

The implication (ii) $\Rightarrow$ (i) is obvious. $\square$

Corollary 4.52 *Consider $\mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$ and let $\mathcal{G}$ be a sub-sigma-algebra of $\mathcal{F}$. Let $X \in \mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$. Then there exists $\hat{X} \in \mathcal{L}^2(\Omega, \mathcal{G}, \mathbb{P})$ such that $\mathbb{E}(X-\hat{X})^2 \leq \mathbb{E}(X-Y)^2$ for all $Y \in \mathcal{L}^2(\Omega, \mathcal{G}, \mathbb{P})$. Any other X' with this property is a.s. equal to $\hat{X}$.*

Proof The space $\mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$ is complete in the sense of Theorem 4.49 and $\mathcal{L}^2(\Omega, \mathcal{G}, \mathbb{P})$ is a closed subspace. The proof of this theorem then becomes just a copy of the proof of Theorem 4.51, the only difference is that for two minimizers one only gets $||\hat{X}_1 - \hat{X}_2||_2 = 0$, which is equivalent to $\hat{X}_1 = \hat{X}_2$ a.s. $\square$

4.8 Exercises

4.1 Let $(x_1, x_2, \ldots)$ be a sequence of nonnegative real numbers, let $\ell : \mathbb{N} \to \mathbb{N}$ be a bijection and define the sequence $(y_1, y_2, \ldots)$ by $y_k = x_{\ell(k)}$. Let for each n the n-vector y^n be given by $y^n = (y_1, \ldots, y_n)$. Consider then for each n a sequence of numbers x^n defined by $x^n_k = x_k$ if x_k is a coordinate of y^n. Otherwise put $x^n_k = 0$. Show that $x^n_k \uparrow x_k$ for every k as $n \to \infty$. Show that $\sum_{k=1}^{\infty} y_k = \sum_{k=1}^{\infty} x_k$.

4.2 Prove Proposition 4.14.

4.3 Prove Proposition 4.18 (assume Proposition 4.14). Show also that $|\mu(f)| \leq \mu(|f|)$, if $f \in \mathcal{L}^1(S, \Sigma, \mu)$.

4.4 Prove the 'almost everywhere' version of Theorem 4.19 by using the 'everywhere' version.

4.5 Here is another version of Scheffé's lemma (Lemma 4.20). Let $(f_n) \subset \mathcal{L}^1(S, \Sigma, \mu)$ and assume that $f_n \to f$ a.e., where $f \in \mathcal{L}^1(S, \Sigma, \mu)$ too. Then $\mu(|f_n - f|) \to 0$ iff $\mu(|f_n|) \to \mu(|f|)$. Show this. (*Hint:* apply Lemma 4.20 to f_n^+ and f_n^-.)

4.6 In this exercise λ denotes Lebesgue measure on the Borel sets of $[0,1]$. Let $f : [0,1] \to \mathbb{R}$ be continuous. Then the Riemann integral $I := \int_0^1 f(x)\,\mathrm{d}x$ exists (this is standard Analysis). But also the Lebesgue integral of f exists. (Explain why.). Construct (use the definition of the Riemann integral) an increasing sequence of simple functions h_n with limit f satisfying $h_n \leq f$ and $\lambda(h_n) \uparrow I$. Prove that $\lambda(f) = I$.

4.7 Let $f : [0, \infty) \to \mathbb{R}$ be given by $f(x) = \frac{\sin x}{x}$ for $x > 0$ and $f(0) = 1$. Show that $I := \int_0^\infty f(x)\,\mathrm{d}x$ exists as an improper Riemann integral (i.e. the limit $\lim_{T\to\infty} \int_0^T f(x)\,\mathrm{d}x$ exists and is finite), but that $f \notin \mathcal{L}^1([0,\infty), \mathcal{B}([0,\infty)), \lambda)$. In Exercise 5.9 you compute that $I = \frac{\pi}{2}$.

4.8 Prove Proposition 4.22 by means of the standard machinery.

4.9 Verify that ν defined in (4.7) is a measure.

4.10 Prove Proposition 4.23.

4.11 Prove Proposition 4.27. *Hint*: Use the standard machinery for h.

4.12 Give the details for Example 4.30.

4.13 Give the proof of Theorem 4.49 for an arbitrary measure space $\mathcal{L}^p(S, \Sigma, \mu)$ and $p \in [0, \infty)$ (it requires minor modifications). Give also the proof of completeness of $\mathcal{L}^\infty(S, \Sigma, \mu)$.

4.14 *This exercise requires knowledge of Chapter 2.* Let $f : [0, 1] \to \mathbb{R}$ be Riemann integrable and let $I = \int_0^1 f(x)\,\mathrm{d}x$. The aim is to show that $f \in \mathcal{L}^1([0, 1], \Sigma_\lambda, \lambda)$. Without loss of generality we may assume that $f \geq 0$.

(a) Exploit Riemann integrability to show that there exist a decreasing sequence of simple functions (u_n) in $\mathfrak{S}^+$ and an increasing sequence (ℓ_n) in $\mathfrak{S}^+$ such that $\ell_n \leq f \leq u_n$, for all n, and $\lambda(\ell_n) \uparrow I$, $\lambda(u_n) \downarrow I$.

(b) Let $u = \lim u_n$ and $\ell = \lim \ell_n$ and put $\hat{f} = \mathbf{1}_{\{u=\ell\}} u$. Show that $\lambda(\{u \neq \ell\}) = 0$ and $\{f \neq \hat{f}\} \subset \{u \neq \ell\}$.

(c) Conclude that $f \in \Sigma_\lambda$ and that $\mu(f) = I$.

4.15 *This exercise concerns a more general version of Theorem 4.19.* Let $(f_n) \subset \Sigma$ and $f = \limsup f_n$ and assume that $f_n(s) \to f(s)$ for all s outside a set of measure zero. Assume also there exist functions $g, g_n \in \Sigma^+$ such that $|f_n| \leq g_n$ a.e. with $g_n(s) \to g(s)$ for all s outside a set of measure zero and that $\mu(g_n) \to \mu(g) < \infty$. Show that $\mu(|f_n - f|) \to 0$.

4.16 Let (S, Σ, μ) be a measurable space, Σ' a sub-σ-algebra of Σ and μ' be the restriction of μ to Σ'. Then also (S, Σ', μ') is a measurable space and integrals of Σ'-measurable functions can be defined according to the usual procedure. Show that $\mu'(f) = \mu(f)$, if $f \geq 0$ and Σ'-measurable. Show also that $\mathcal{L}^1(S, \Sigma, \mu) \cap \Sigma' = \mathcal{L}^1(S, \Sigma', \mu')$.

4.17 Prove Proposition 4.40.

4.18 Let G be an interval, X a random variable. Assume $\mathbb{P}(X \in G) = 1$ and $\mathbb{E}|X| < \infty$. If X is not degenerate ($\mathbb{P}(X = x) < 1$ for all $x \in G$), show that $\mathbb{E}X \in \operatorname{Int} G$.

4.19 Let $f \in \mathcal{L}^\infty(S, \Sigma, \mu)$ and suppose that $\mu(\{f \neq 0\}) < \infty$. We will see that $\lim_{p\to\infty} \|f\|_p = \|f\|_\infty$.

(a) Show that $\limsup_{p\to\infty} \|f\|_p \leq \|f\|_\infty$.

(b) Show that $\liminf_{p\to\infty} \|f\|_p \geq \|f\|_\infty$. *Hint: for $\varepsilon > 0$ it holds that* $\mu(\{f > \|f\|_\infty - \varepsilon\}) > 0$.

(c) Show also that $\|f\|_p$ converges monotonically to $\|f\|_\infty$ if μ is a probability measure.

5 Product measures

So far we have considered measure spaces (S, Σ, μ) and we have looked at integrals of the type $\mu(f) = \int f \, \mathrm{d}\mu$. Here f is a function of 'one' variable (depends on how you count and what the underlying set S is). Suppose that we have two measure spaces (S_1, Σ_1, μ_1) and (S_2, Σ_2, μ_2) and a function $f : S_1 \times S_2 \to \mathbb{R}$. Is it possible to integrate such a function of two variables w.r.t. some measure, that has to be defined on some Σ-algebra of $S_1 \times S_2$. There is a natural way of constructing this σ-algebra and a natural construction of a measure on this σ-algebra. Here is a setup with some informal thoughts.

Take $f : S_1 \times S_2 \to \mathbb{R}$ and assume any good notion of measurability and integrability. Then $\mu(f(\cdot, s_2)) := \int f(\cdot, s_2) \, \mathrm{d}\mu_1$ defines a function of s_2 and so we'd like to take the integral w.r.t. μ_2. We could as well have gone the other way round (integrate first w.r.t. μ_2), and the questions are whether these integrals are well defined and whether both approaches yield the same result.

Here is a simple special case, where the latter question has a negative answer. We have seen that integration w.r.t. counting measure is nothing else but addition. What we have outlined above is in this context just interchanging the summation order. So if $(a_{n,m})$ is a double array of real numbers, the above is about whether $\sum_n \sum_m a_{n,m} = \sum_m \sum_n a_{n,m}$. This is obviously true if n and m run through a finite set, but things can go wrong for indices from infinite sets. Consider for example

$$a_{n,m} = \begin{cases} 1 & \text{if } n = m+1 \\ -1 & \text{if } m = n+1 \\ 0 & \text{else.} \end{cases}$$

One easily verifies $\sum_m a_{1,m} = -1$, $\sum_m a_{n,m} = 0$, if $n \geq 2$ and hence we find $\sum_n \sum_m a_{n,m} = -1$. Similarly one shows that $\sum_m \sum_n a_{n,m} = +1$. In order that interchanging of the summation order yields the same result, additional conditions have to be imposed. We will see that $\sum_m \sum_n |a_{n,m}| < \infty$ is a sufficient condition. As a side remark we note that this case has everything to do with a well known theorem by Riemann that says that a series of real numbers is absolutely convergent iff it is unconditionally convergent.

5.1 Product of two measure spaces

Our aim is to construct a measure space (S, Σ, μ) with $S = S_1 \times S_2$. First we construct Σ. It is natural that 'measurable rectangles' are in Σ. Let $\mathcal{R} = \{E_1 \times E_2 : E_1 \in \Sigma_1, E_2 \in \Sigma_2\}$. Obviously $\mathcal{R}$ is a π-system, but in general not a σ-algebra on S. Therefore we define $\Sigma := \sigma(\mathcal{R})$, the *product σ-algebra* of Σ_1 and Σ_2. A common notation for this product σ-algebra, also used below for similar cases, is $\Sigma = \Sigma_1 \times \Sigma_2$.

Alternatively, one can consider the projections $\pi_i : S \to S_i$, defined by $\pi_i(s_1, s_2) = s_i$. It is easy to show that Σ coincides with the smallest σ-algebra that makes these projections measurable.

Next to the projections, we now consider *embeddings*. For fixed $s_1 \in S_1$ we define $e_{s_1} : S_2 \to S$ by $e_{s_1}(s_2) = (s_1, s_2)$. Similarly we define $e^{s_2}(s_1) = (s_1, s_2)$. One easily checks that the embeddings e_{s_1} are Σ_2/Σ-measurable and that the e^{s_2} are Σ_1/Σ-measurable (Exercise 5.1). As a consequence we have the following proposition.

Proposition 5.1 *Let $f : S \to \mathbb{R}$ be Σ-measurable. Then the* marginal *mappings $s_1 \mapsto f(s_1, s_2)$ and $s_2 \mapsto f(s_1, s_2)$ are Σ_1-, respectively Σ_2-measurable, for any $s_2 \in S_2$, respectively $s_1 \in S_1$.*

Proof This follows from the fact that a composition of measurable functions is also measurable. □

Remark 5.2 The converse statement of Proposition 5.1 is in general not true. There are functions $f : S \to \mathbb{R}$ that are not measurable w.r.t. the product σ-algebra Σ, although the mappings $s_1 \mapsto f(s_1, s_2)$ and $s_2 \mapsto f(s_1, s_2)$ are Σ_1-, respectively Σ_2-measurable. Counterexamples are not obvious, see below for a specific one. Fortunately, there are also conditions that are sufficient to have measurability of f w.r.t. Σ, when measurability of the marginal functions is given. See Exercise 5.8.

Here is a sketch of a counterexample, based on the Continuum Hypothesis, with $(S_1, \Sigma_1, \mu_1) = (S_2, \Sigma_2, \mu_2) = ([0,1], \mathcal{B}([0,1]), \lambda)$. Consider $V := \{0,1\}^{\mathbb{N}}$ and the set W of (countable) ordinal numbers smaller than ω_1, the first uncountable ordinal number. The cardinality of W is equal to $\aleph_1$, whereas $\{0,1\}^{\mathbb{N}}$ has cardinality $2^{\aleph_0}$. The Continuum hypothesis states that there exists a bijective mapping between V and W. Hence there also exists a bijective mapping ϕ between the interval $[0,1]$ and W. The set W has the property that for every $x \in [0,1]$ the element $\phi(x)$ of W has countable many predecessors. Consider $Q := \{(x,y) \in [0,1]^2 : \phi(x) < \phi(y)\}$. For fixed y we let $Q^y = \{x \in [0,1] : (x,y) \in Q\}$ and for fixed x we let $Q_x = \{y \in [0,1] : (x,y) \in Q\}$. It follows that for every y, the set Q^y is countable and thus Borel-measurable and it has Lebesgue measure zero. For every x, the complement of Q_x is countable and has Lebesgue measure zero, hence Q_x has Lebesgue measure one. For $f = \mathbf{1}_Q$, we thus have that $x \mapsto f(x,y)$ and $y \mapsto f(x,y)$ are Borel-measurable and that $I_1^f(x) = 1$ and $I_2^f(y) = 0$. We see that Lemma 5.3 doesn't hold and therefore conclude that f cannot be measurable w.r.t. the product σ-algebra $\mathcal{B}([0,1]) \times \mathcal{B}([0,1])$.

Having constructed the product σ-algebra Σ, we now draw our attention to the construction of the *product measure* μ on Σ, denoted by $\mu_1 \times \mu_2$. We will construct μ such that the property $\mu(E_1 \times E_2) = \mu_1(E_1)\mu_2(E_2)$ holds. This justifies the name product measure.

Until later notice we assume that the measures μ_1 and μ_2 are finite.

Consider a bounded Σ-measurable function f. We know that the mappings $s_i \mapsto f(s_1, s_2)$ are Σ_i-measurable and therefore the integrals w.r.t. μ_i are well

defined (why?). Let then

$$I_1^f(s_1) = \int f(s_1, s_2)\mu_2(\mathrm{d}s_2)$$
$$I_2^f(s_2) = \int f(s_1, s_2)\mu_1(\mathrm{d}s_1).$$

Lemma 5.3 *Let f be a bounded Σ-measurable function. Then the mappings $I_i^f : S_i \to \mathbb{R}$ are Σ_i-measurable ($i = 1, 2$). Moreover we have the identity*

$$\mu_1(I_1^f) = \mu_2(I_2^f), \tag{5.1}$$

or, in a more appealing notation,

$$\int_{S_1} \big(\int_{S_2} f(s_1, s_2)\mu_2(\mathrm{d}s_2)\big)\mu_1(\mathrm{d}s_1) = \int_{S_2} \big(\int_{S_1} f(s_1, s_2)\mu_1(\mathrm{d}s_1)\big)\mu_2(\mathrm{d}s_2). \tag{5.2}$$

Proof We use the Monotone Class Theorem, Theorem 3.6, and so we have to find a good vector space $\mathcal{H}$. The obvious candidate is the collection of all bounded Σ-measurable functions f that satisfy the assertions of the lemma.

First we notice that $\mathcal{H}$ is indeed a vector space, since sums of measurable functions are measurable and by linearity of the integral. Obviously, the constant functions belong to $\mathcal{H}$. Then we have to show that if $f_n \in \mathcal{H}$, $f_n \geq 0$ and $f_n \uparrow f$, where f is bounded, then also $f \in \mathcal{H}$. Of course here the Monotone Convergence Theorem comes into play. First we notice that measurability of the I_i^f follows from measurability of the $I_i^{f_n}$ for all n. Theorem 4.12 yields that the sequences $I_i^{f_n}(s_i)$ are increasing and converging to $I_i^f(s_i)$. Another application of this theorem yields that $\mu_1(I_1^{f_n})$ converges to $\mu_1(I_1^f)$ and that $\mu_2(I_2^{f_n})$ converges to $\mu_2(I_2^f)$. Since $\mu_1(I_1^{f_n}) = \mu_2(I_2^{f_n})$ for all n, we conclude that $\mu_1(I_1^f) = \mu_2(I_2^f)$, whence $f \in \mathcal{H}$.

Next we check that $\mathcal{H}$ contains the indicators of sets in $\mathcal{R}$. A quick computation shows that for $f = \mathbf{1}_{E_1 \times E_2}$ one has $I_1^f = \mathbf{1}_{E_1}\mu_2(E_2)$, which is Σ_1-measurable, $I_2^f = \mathbf{1}_{E_2}\mu_1(E_1)$, and $\mu_1(I_1^f) = \mu_2(I_2^f) = \mu_1(E_1)\mu_2(E_2)$. Hence $f \in \mathcal{H}$. By Theorem 3.6 we conclude that $\mathcal{H}$ coincides with the space of all bounded Σ-measurable functions. □

It follows from Lemma 5.3 that for all $E \in \Sigma$, the indicator function $\mathbf{1}_E$ satisfies the assertions of the lemma. This shows that the following definition is meaningful.

Definition 5.4 We define $\mu : \Sigma \to [0, \infty)$ by $\mu(E) = \mu_2(I_2^{\mathbf{1}_E})$ for $E \in \Sigma$.

In Theorem 5.5 below (known as Fubini's theorem) we assert that this defines a measure on (S, Σ) and it also tells us how to compute integrals w.r.t. this measure in terms of iterated integrals w.r.t. μ_1 and μ_2.

Theorem 5.5 *The mapping μ of Definition 5.4 has the following properties.*

(i) It is a measure on (S, Σ). Moreover, it is the only measure on (S, Σ) with the property that $\mu(E_1 \times E_2) = \mu_1(E_1)\mu_1(E_2)$. It is therefore called the product measure *of μ_1 and μ_2 and often written as $\mu_1 \times \mu_2$.*

(ii) If $f \in \Sigma^+$, then

$$\mu(f) = \mu_2(I_2^f) = \mu_1(I_1^f) \leq \infty. \tag{5.3}$$

(iii) If $f \in \mathcal{L}^1(S, \Sigma, \mu)$, then Equation (5.3) is still valid and $\mu(f) \in \mathbb{R}$.

Proof (i) It is obvious that $\mu(\emptyset) = 0$. If (E_n) is a disjoint sequence in Σ with union E, then we have $\mathbf{1}_E = \lim_n \sum_{i=1}^n \mathbf{1}_{E_i}$. Linearity of the integral and Monotone Convergence (applied two times) show that μ is σ-additive. Uniqueness of μ follows from Theorem 1.15 applied to the π-system $\mathcal{R}$.

(ii) We use the standard machine. The two equalities in (5.3) are by definition of μ valid for $f = \mathbf{1}_E$, when $E \in \Sigma$. Linearity of the integrals involved show that it is true for nonnegative simple functions f and Monotone Convergence yields the assertion for $f \in \Sigma^+$.

(iii) Of course, here we have to use the decomposition $f = f^+ - f^-$. The tricky details are left as Exercise 5.2. □

Theorem 5.5 has been proved under the standing assumption that the initial measures μ_1 and μ_2 are finite. The results extend to the case where both these measures are σ-finite. The approach is as follows. Write $S_1 = \cup_{i=1}^\infty S_1^i$ with the $S_1^i \in \Sigma_1$ and $\mu_1(S_1^i) < \infty$. Without loss of generality, we can take the S_1^i disjoint. Take a similar partition (S_2^j) of S_2. Then $S = \cup_{i,j} S_{ij}$, where the $S_{ij} := S_1^i \times S_2^j$, form a countable disjoint union as well. Let $\Sigma_{ij} = \{E \cap S_{ij} : E \in \Sigma\}$. On each measurable space (S_{ij}, Σ_{ij}) the above results apply and one has e.g. identity of the involved integrals by splitting the integration over the sets S_{ij} and adding up the results.

We note that if one goes beyond σ-finite measures (often a good thing to do if one wants to have counterexamples), the assertion may no longer be true. Let $S_1 = S_2 = [0, 1]$ and $\Sigma_1 = \Sigma_2 = \mathcal{B}[0, 1]$. Take μ_1 equal to Lebesgue measure and μ_2 the counting measure, the latter is not σ-finite. It is a nice exercise to show that $\Delta := \{(x, y) \in S : x = y\} \in \Sigma$. Let $f = \mathbf{1}_\Delta$. Obviously $I_1^f(s_1) \equiv 1$ and $I_2^f(s_2) \equiv 0$ and the two iterated integrals in (5.3) are 1 and 0. So, more or less everything above concerning product measures fails in this example.

We proceed with a few remarks on products with more than two factors. The construction of a product measure space carries over, without any problem, to products of more than two factors, as long as there are finitely many. This results in product spaces of the form $(S_1 \times \ldots \times S_n, \Sigma_1 \times \ldots \times \Sigma_n, \mu_1 \times \ldots \times \mu_n)$ under conditions similar to those of Theorem 5.5. The product σ-algebra is again defined as the smallest σ-algebra that makes all projections measurable. Existence of product measures is proved in just the same way as before, using an induction argument. Note that there will be many possibilities to extend (5.1) and (5.2), since there are $n!$ different integration orders. We leave the details to the reader.

Things become however more complicated, when we work with infinite products of measure spaces (S_i, Σ_i, μ_i). In Section 5.3 we treat the construction of an infinite product of probability spaces.

The measure μ in Fubini's theorem is a product measure of two base measures, but it is possible to go beyond the setting of product measures. We present this extension here. We consider a probability space $(S, \Sigma, \mathbb{Q})$ and a measurable space $(\Omega, \mathcal{F})$. On the latter space we have a family of probability measures $\{\mathbb{P}_x : x \in S\}$ that have the property that for every $F \in \mathcal{F}$ the function $x \mapsto \mathbb{P}_x(F)$ is measurable as a function on (S, Σ). Such a family is known under various names (depending on the context), we mention Markov kernel, or transition kernel. Let $f : S \times \Omega \to \mathbb{R}$ be measurable w.r.t. to the product σ-algebra $\Sigma \times \mathcal{F}$. We already know from Proposition 5.1 that the marginal mappings of f are measurable on the factor spaces. A version of Fubini's theorem in the present situation is as follows.

Theorem 5.6 *There exists a unique probability measure Π on $(S \times \Omega, \Sigma \times \mathcal{F})$ such that $\Pi(E \times F) = \int_E \mathbb{P}_x(F)\mathbb{Q}(\,\mathrm{d}x)$. Moreover, if $f : S \times \Omega \to [0, \infty]$ is $\Sigma \times \mathcal{F}$-measurable, then $x \mapsto \int_\Omega f(x, \omega)\mathbb{P}_x(\,\mathrm{d}\omega)$ is Σ-measurable, and*

$$\int_{S\times\Omega} f \,\mathrm{d}\Pi = \int_S \int_\Omega f(x, \omega)\mathbb{P}_x(\,\mathrm{d}\omega)\,\mathbb{Q}(\,\mathrm{d}x).$$

Proof The proof can be given completely along the lines of the proofs of Lemma 5.3 and Theorem 5.5. This is Exercise 5.14. □

Here is a corollary that one easily derives from the theorem.

Corollary 5.7 *Consider the measure Π in Theorem 5.6. Then $\bar{\Pi}(F) := \Pi(S \times F) = \int_S \mathbb{P}_x(F)\,\mathbb{Q}(\,\mathrm{d}x)$ defines a probability measure on $(\Omega, \mathcal{F})$. Let $X : \Omega \to \mathbb{R}$ be a random variable, nonnegative or integrable w.r.t. $\bar{\Pi}$. Then the expectation of X under $\bar{\Pi}$, $\mathbb{E}_{\bar{\Pi}}X$ satisfies $\mathbb{E}_{\bar{\Pi}}X = \int_S \mathbb{E}_x X\,\mathbb{Q}(\,\mathrm{d}x)$, where $\mathbb{E}_x$ denotes expectation under $\mathbb{P}_x$.*

Proof Note that the measure $\bar{\Pi}$ on $(\Omega, \mathcal{F})$ can be considered as the image measure of Π under the projection of $S \times \Omega$ onto Ω. A combination of Proposition 4.27 and Theorem 5.6 then gives the result. □

5.2 Further applications in Probability theory

In this section we consider real valued random variables, as well as real *random vectors*. The latter require a definition. Consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a map $X : \Omega \to E$, where E is some other set. Let $\mathcal{E}$ be a σ-algebra on E. If the map X is $\mathcal{F}/\mathcal{E}$ measurable, X is also called a random element of E. If E is a vector space, we call X in such a case a random vector. Notice that this definition depends on the σ-algebras at hand, which we don't immediately recognize in the term random vector.

An obvious example of a vector space is $\mathbb{R}^2$. Suppose we have two random variables $X_1, X_2 : \Omega \to \mathbb{R}$. We can consider the map $X = (X_1, X_2) : \Omega \to \mathbb{R}^2$, defined by $X(\omega) = (X_1(\omega), X_2(\omega))$ and it is natural to call X a random vector. To justify this terminology, we need a σ-algebra on $\mathbb{R}^2$ and there are two obvious candidates, the Borel σ-algebra $\mathcal{B}(\mathbb{R}^2)$ generated by the ordinary open sets (as in Section 1.1), and, continuing our discussion of the previous section, the product σ-algebra $\mathcal{B}(\mathbb{R}) \times \mathcal{B}(\mathbb{R})$.

Proposition 5.8 *It holds that* $\mathcal{B}(\mathbb{R}^2) = \mathcal{B}(\mathbb{R}) \times \mathcal{B}(\mathbb{R})$.

Proof The projections $\pi_i : \mathbb{R}^2 \to \mathbb{R}$ are continuous and thus $\mathcal{B}(\mathbb{R}^2)$-measurable. Since $\mathcal{B}(\mathbb{R}) \times \mathcal{B}(\mathbb{R})$ is the smallest σ-algebra for which the projections are measurable, we have $\mathcal{B}(\mathbb{R}) \times \mathcal{B}(\mathbb{R}) \subset \mathcal{B}(\mathbb{R}^2)$. Conversely, if G is open in $\mathbb{R}^2$, it is the countable union of (open) rectangles in $\mathcal{R}$ (similar to the proof of Proposition 1.3) and hence $G \in \mathcal{B}(\mathbb{R}) \times \mathcal{B}(\mathbb{R})$, which yields the other inclusion. □

Remark 5.9 Observe that the proof of $\mathcal{B}(\mathbb{R}) \times \mathcal{B}(\mathbb{R}) \subset \mathcal{B}(\mathbb{R}^2)$ generalizes to the situation, where one deals with two topological spaces with the Borel sets. For the proof of the other inclusion, we used (and needed) the fact that $\mathbb{R}$ is separable under the ordinary topology. In a general setting one might have the strict inclusion of the product σ-algebra in the Borel σ-algebra on the product space (with the product topology).

We now know that there is no difference between $\mathcal{B}(\mathbb{R}^2)$ and $\mathcal{B}(\mathbb{R}) \times \mathcal{B}(\mathbb{R})$. This facilitates the use of the term 2-dimensional random vector and we have the following easy to prove corollary.

Corollary 5.10 *Let* $X_1, X_2 : \Omega \to \mathbb{R}$ *be given. The vector mapping* $X = (X_1, X_2) : \Omega \to \mathbb{R}^2$ *is a random vector iff the* X_i *are random variables.*

Proof Exercise 5.3. □

Remark 5.11 Let X_1, X_2 be two random variables. We already knew that $X_1 + X_2$ is a random variable too. This also follows from the present results. Let $f : \mathbb{R}^2 \to \mathbb{R}$ be a continuous function. Then it is also $\mathcal{B}(\mathbb{R}^2)$-measurable, and by Corollary 5.10 and composition of measurable functions, $f(X_1, X_2)$ is a random variable as well. Apply this with $f(x_1, x_2) = x_1 + x_2$.

Recall that we defined in Section 3.2 the distribution, or the law, of a random variable. Suppose that $X = (X_1, X_2)$ is a random vector defined on $(\Omega, \mathcal{F}, \mathbb{P})$ with values in $\mathbb{R}^2$. Let $E \in \mathcal{B}(\mathbb{R}^2)$, then

$$\mathbb{P}^X(E) := \mathbb{P}(X \in E),$$

for $E \in \mathcal{B}(\mathbb{R}^2)$ defines a probability measure on $(\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2))$, the distribution of X, also called the *joint* distribution of (X_1, X_2). If we take $E = E_1 \times \mathbb{R}$, then we have $\mathbb{P}^X(E_1 \times \mathbb{R}) = \mathbb{P}(X_1 \in E_1) = \mathbb{P}^{X_1}(E_1)$. Common terminology is to call $\mathbb{P}^{X_1}$ the marginal distribution, or marginal law, of X_1.

Along with the (joint) distribution of X, we introduce the joint distribution function $F = F_X : \mathbb{R}^2 \to [0,1]$, given by

$$F(x_1, x_2) = \mathbb{P}^X((-\infty, x_1] \times (-\infty, x_2]) = \mathbb{P}(X_1 \leq x_1, X_2 \leq x_2).$$

Notice that, for instance, $F_{X_1}(x_1) = \lim_{x_2 \to \infty} F(x_1, x_2)$, also denoted $F(x_1, \infty)$.

It may happen that there exists a nonnegative $\mathcal{B}(\mathbb{R}^2)$-measurable function f such that $\mathbb{P}^X(E) = \int_E f \, \mathrm{d}(\lambda \times \lambda)$, for all $E \in \mathcal{B}(\mathbb{R}^2)$. In that case, f is called the (joint) density of X. The obvious marginal density f_{X_1} of X_1 is defined by $f_{X_1}(x_1) = \int f(x_1, x_2) \, \lambda(\mathrm{d}x_2)$. One similarly defines the marginal density of X_2. Check these are indeed densities in the sense of Example 4.29.

Independence (of random variables) had to do with multiplication of probabilities (see Definition 3.11), so it should in a natural way be connected to product measures.

Proposition 5.12 *Two random variables X_1, X_2 on $(\Omega, \mathcal{F}, \mathbb{P})$ are independent iff the joint distribution $\mathbb{P}^{(X_1, X_2)}$ is the product measure $\mathbb{P}^{X_1} \times \mathbb{P}^{X_2}$. This in turn happens iff $F(x_1, x_2) = F_{X_1}(x_1) F_{X_2}(x_2)$, for all $x_1, x_2 \in \mathbb{R}$. Assume further that (X_1, X_2) has a joint probability density function f. Let f_1 and f_2 be the (marginal) probability density functions of X_1 and X_2 respectively. Then X_1 and X_2 are independent iff $f(x_1, x_2) = f_1(x_1) f_2(x_2)$ for all (x_1, x_2) except in a set of $\lambda \times \lambda$-measure zero.*

Proof Exercise 5.4. □

The results of the present section (Proposition 5.8, Corollary 5.10, Proposition 5.12) have obvious extensions to higher dimensional situations. We leave the formulation to the reader.

Remark 5.13 Suppose one is given a random variable X, defined on a given $(\Omega, \mathcal{F}, \mathbb{P})$. Sometimes one needs an additional random variable Y having a specified distribution. It may happen that the given probability space is not rich enough to have such a random variable well defined. Suppose $\Omega = \{0, 1\}$ and $X(\omega) = \omega$, having a Bernoulli distribution for $\mathbb{P}$ defined on the power set of Ω with $\mathbb{P}(\{1\}) = p$. Clearly, it is impossible to define on this Ω a random variable having more than two different outcomes. Extending the probability space to a suitable product space offers a way out, see Exercise 5.13, from which it even follows that X and Y are independent.

5.3 Infinite products

The extension of product spaces from finite to infinite products is a different matter. Nevertheless, this extension is inevitable if one wants to construct a well defined independent infinite sequence of random variables. Just recall that we have seen that independence of two random variables has everything to do with product measures. Hence for an (infinite) sequence of independent random variables, one should use an infinite product of probability measures.

For real valued random variables we have already encountered a construction of a supporting probability space in Section 3.3. Here we continue with the construction of a *countable product of probability spaces*. See also Exercise 5.12 for products of arbitrarily many factors.

Assume that we have for every $n \in \mathbb{N}$ a probability space $(\Omega_n, \mathcal{F}_n, \mathbb{P}_n)$. Let $\Omega = \prod_{n=1}^{\infty} \Omega_n$ and denote by $\omega = (\omega_1, \omega_2, \ldots)$ a typical element of Ω. The construction of the product σ-algebra remains the same as before. The projection $\pi_n : \Omega \to \Omega_n$ is defined by $\pi_n(\omega) = \omega_n$. On the product set Ω we define the σ-algebra $\mathcal{F}$ as the smallest one that makes all projections π_n measurable mappings (this reminds you of a subbasis of the product topology). One can also define a multivariate projection $\pi_{(1,\ldots,n)} : \Omega \to \prod_{k=1}^{n} \Omega_k$ by $\pi_{(1,\ldots,n)}(\omega) = (\omega_1, \ldots, \omega_n)$. It follows that all multivariate projections are $\mathcal{F}$-measurable as well. A *cylinder*, or a *measurable rectangle*, is by definition the inverse image of a measurable set in some $\prod_{k=1}^{n} \Omega_k$, endowed with the product σ-algebra $\mathcal{F}_1 \times \cdots \times \mathcal{F}_n$ under the projection $\pi_{(1,\ldots,n)}$. It follows that a cylinder is of the type $B_n \times \prod_{k=n+1}^{\infty} \Omega_k$ (for some n), with $B_n \in \mathcal{F}_1 \times \cdots \times \mathcal{F}_n$. Such a cylinder will be denoted by C_n. Let $\mathcal{C}$ be the collection of all cylinders and note that $\mathcal{C}$ is an algebra. Define the mapping $\mathbb{P}_0 : \mathcal{C} \to [0,1]$ by $\mathbb{P}_0(C) = \prod_{i=1}^{n} \mathbb{P}_i(B_n)$, if $C = C_n$ for some n. Verify that if one writes $C = C_m$ for some $m \neq n$, it holds that $\prod_{i=1}^{n} \mathbb{P}_i(B_n) = \prod_{i=1}^{m} \mathbb{P}_i(B_m)$, which implies that $\mathbb{P}_0$ is unambiguously defined, i.e. not depending on the chosen representation of C. Verify too that $\mathbb{P}_0$ is finitely additive on $\mathcal{C}$. The next theorem states the existence of an infinite product probability measure $\mathbb{P}$, sometimes denoted by $\prod_{n=1}^{\infty} \mathbb{P}_n$. In the proof we use results from Section 2.2.

Theorem 5.14 *There exists a unique probability measure $\mathbb{P}$ on $(\Omega, \mathcal{F})$ such that $\mathbb{P}$ restricted to $\mathcal{C}$ is equal to $\mathbb{P}_0$. In particular,*

$$\mathbb{P}(E_1 \times \cdots \times E_n \times \prod_{k=n+1}^{\infty} \Omega_k) = \prod_{i=1}^{n} \mathbb{P}_i(E_i),$$

if $E_i \in \mathcal{F}_i$, $i = 1, \ldots, n$.

Proof The proof is based on an application of Theorem 2.7. We proceed by showing that $\mathbb{P}_0$ is countably additive on $\mathcal{C}$. To that end we invoke Exercise 1.9, from which we deduce that it is sufficient to show for a decreasing sequence of cylinders C^n with $\lim_{n\to\infty} \mathbb{P}_0(C^n) > 0$, one must have $C := \cap_{n=1}^{\infty} C^n \neq \emptyset$. Without loss of generality we may assume that the C^n are of the type C_n as above. If it happens that there is an $N \in \mathbb{N}$ such that all C_n can be written as $B_n \times \prod_{n=N+1}^{\infty} \Omega_n$, with $B_n \in \mathcal{F}_1 \times \ldots \times \mathcal{F}_N$, we are done, since in this case $\mathbb{P}_0(C_n) = \mathbb{P}_1 \times \ldots \times \mathbb{P}_N(B_n)$. We already know that $\mathbb{P}_1 \times \ldots \times \mathbb{P}_N$ is a measure and therefore countably additive. Henceforth we assume that such an N doesn't exist.

For simplicity we write $\Omega'_n = \prod_{k=n+1}^{\infty} \Omega_k$, so that $C_n = B_n \times \Omega'_n$. Typical elements of Ω'_n are ω'_n and we have $\omega = (\omega_1, \ldots, \omega_n, \omega'_n)$. On the cylinders in Ω'_n we can define set functions $\mathbb{P}'_n$ in the same way as $\mathbb{P}_0$ was defined on $\mathcal{C}$. Note

that, similar to the definition of $\mathbb{P}_0$, the action of $\mathbb{P}'_n$ on cylinders in Ω'_n only involves product measures with finitely many factors. For every cylinder C in $\mathcal{C}$, one defines $C(\omega_1, \ldots, \omega_n) = \{\omega'_n : (\omega_1, \ldots, \omega_n, \omega'_n) \in C\}$. Then the probabilities $\mathbb{P}'_n(C(\omega_1, \ldots, \omega_n))$ are well defined.

Since we have assumed that $\lim_{n\to\infty} \mathbb{P}_0(C_n) > 0$ for the decreasing sequence (C_n), there exists $\varepsilon > 0$ such that $\mathbb{P}_0(C_n) > \varepsilon$ for all n. Define

$$E_n^1 = \{\omega_1 : \mathbb{P}'_1(C_n(\omega_1)) > \frac{1}{2}\varepsilon\}.$$

It follows from Lemma 5.3 that $E_n^1 \in \mathcal{F}_1$. Then, using 'Fubini computations', we obtain

$$\begin{aligned}
\mathbb{P}_0(C_n) &= \int_{\Omega_1} \mathbb{P}'_1(C_n(\omega_1))\, \mathbb{P}_1(\mathrm{d}\omega_1) \\
&= \int_{E_n^1} \mathbb{P}'_1(C_n(\omega_1))\, \mathbb{P}_1(\mathrm{d}\omega_1) + \int_{\Omega_1 \setminus E_n^1} \mathbb{P}'_1(C_n(\omega_1))\, \mathbb{P}_1(\mathrm{d}\omega_1) \\
&\leq \mathbb{P}_1(E_n^1) + \frac{1}{2}\varepsilon \mathbb{P}_1(\Omega_1 \setminus E_n^1).
\end{aligned}$$

Since $\mathbb{P}_0(C_n) > \varepsilon$, it then follows that $\mathbb{P}_1(E_n^1) > \frac{1}{2}\varepsilon$. Since the C_n form a decreasing sequence, the same holds for the E_n^1. Letting $E^1 = \cap_n E_n^1$, we get by continuity of $\mathbb{P}_1$ that $\mathbb{P}_1(E^1) \geq \frac{1}{2}\varepsilon$. In particular E^1 is not empty and we can choose some $\omega_1^* \in E^1$ for which we have $\mathbb{P}'_1(C_n(\omega_1^*)) > \frac{\varepsilon}{2}$ for all n.

Then we repeat the above story applied to the sets $C_n(\omega_1^*)$ instead of the C_n. So we consider the sets $C_n(\omega_1^*, \omega_2)$ and $E_n^2(\omega_1^*) = \{\omega_2 : \mathbb{P}'_2(C_n(\omega_1^*, \omega_2)) > \frac{\varepsilon}{4}\}$. This results in a non-empty limit set $E^2(\omega_1^*) \subset \Omega_2$ from which we select some ω_2^* and we obtain $\mathbb{P}'_2(C_n(\omega_1^*, \omega_2^*)) > \frac{\varepsilon}{4}$ for all n. Continuing this way we construct a point $\omega^* = (\omega_1^*, \omega_2^*, \ldots)$ that belongs to all C_n and therefore the intersection $\cap_n C_n$ is not empty. □

5.4 Exercises

5.1 Show that the embeddings e_{s_1} are Σ_2/Σ-measurable and that the e^{s_2} are Σ_1/Σ-measurable. Also prove Proposition 5.1.

5.2 Prove part (iii) of Fubini's theorem (Theorem 5.5) for $f \in \mathcal{L}^1(S, \Sigma, \mu)$ (you already know it for $f \in \Sigma^+$). Explain why $s_1 \mapsto f(s_1, s_2)$ is in $\mathcal{L}^1(S_1, \Sigma_1, \mu_1)$ for all s_2 outside a set N of μ_2-measure zero and that I_2^f is well defined on N^c.

5.3 Prove Corollary 5.10.

5.4 Prove Proposition 5.12.

5.5 A two-dimensional random vector (X, Y) is said to have a density f w.r.t. the Lebesgue measure on $\mathcal{B}(\mathbb{R})^2$ is for every set $B \in \mathcal{B}(\mathbb{R}^2)$ one has

$$\mathbb{P}((X, Y) \in B) = \int\int_B f(x, y)\, \mathrm{d}x\, \mathrm{d}y.$$

Define

$$f_X(x) = \int_{\mathbb{R}} f(x,y)\ \mathrm{d}y.$$

Show that for all $B \in \mathcal{B}(\mathbb{R})$ one has

$$\mathbb{P}^X(B) = \int_B f_X(x)\ \mathrm{d}x.$$

5.6 Let X and Y be independent random variables on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let F_X and F_Y be their distribution functions and μ_X and μ_Y their laws. Put $Z = X + Y$ and F_Z its distribution function.

(a) Show that $F_Z(z) = \int_{\mathbb{R}} F_X(z-y)\,\mu_Y(\mathrm{d}y)$.

(b) Assume that F_X admits a density f_X (w.r.t. Lebesgue measure). Show that also F_Z admits a density, which can be taken to be

$$f_Z(z) := \int_{\mathbb{R}} f_X(z-y)\,\mu_Y(\mathrm{d}y).$$

5.7 If $Z_1, Z_2, \ldots$ is a sequence of nonnegative random variables, then

$$\mathbb{E}\sum_{k=1}^{\infty} Z_k = \sum_{k=1}^{\infty} \mathbb{E}Z_k. \tag{5.4}$$

(a) Show that this follows from Fubini's theorem (as an alternative to the arguments of Exercise 4.12). If $\sum_{k=1}^{\infty} \mathbb{E}Z_k < \infty$, what is $\mathbb{P}(\sum_{k=1}^{\infty} Z_k = \infty)$?

(b) Formulate a result similar to (5.4) for random variables Z_k that may assume negative values as well.

5.8 Let f be defined on $\mathbb{R}^2$ such that for all $a \in \mathbb{R}$ the function $y \mapsto f(a,y)$ is Borel measurable and such that for all $b \in \mathbb{R}$ the function $x \mapsto f(x,b)$ is continuous.

(a) Show that for all $a, b, c \in \mathbb{R}$ the function $(x,y) \mapsto bx + cf(a,y)$ is Borel-measurable on $\mathbb{R}^2$.

(b) Let $a_i^n = i/n$, $i \in \mathbb{Z}, n \in \mathbb{N}$. Define

$$f^n(x,y) = \sum_i \mathbf{1}_{(a_{i-1}^n, a_i^n]}(x)\Big(\frac{a_i^n - x}{a_i^n - a_{i-1}^n} f(a_{i-1}^n, y) + \frac{x - a_{i-1}^n}{a_i^n - a_{i-1}^n} f(a_i^n, y)\Big).$$

Show that the f^n are Borel-measurable on $\mathbb{R}^2$ and conclude that f is Borel-measurable on $\mathbb{R}^2$.

5.9 Show that for $t > 0$

$$\int_0^{\infty} \sin x\, e^{-tx}\ \mathrm{d}x = \frac{1}{1+t^2}.$$

Although $x \mapsto \frac{\sin x}{x}$ doesn't belong $\mathcal{L}^1([0,\infty), \mathcal{B}([0,\infty)), \lambda)$, show that one can use Fubini's theorem to compute the improper Riemann integral

$$\int_0^{\infty} \frac{\sin x}{x}\ \mathrm{d}x = \frac{\pi}{2}.$$

5.10 Let $F, G : \mathbb{R} \to \mathbb{R}$ be nondecreasing and right-continuous. Similar to the case of distribution functions, these generate measures μ_F and μ_G on the Borel sets satisfying e.g. $\mu_F((a,b]) = F(b) - F(a)$. Integrals w.r.t μ_F are commonly denoted by $\int f \,\mathrm{d}F$ instead of $\int f \,\mathrm{d}\mu_F$. See also Section 4.5.

(a) Use Fubini's theorem to show the integration by parts formula, valid for all $a < b$,

$$F(b)G(b) - F(a)G(a) = \int_{(a,b]} F(s-) \,\mathrm{d}G(s) + \int_{(a,b]} G(s) \,\mathrm{d}F(s),$$

where $F(s-) = \lim_{u \uparrow s} F(u)$. *Hint:* integrate $\mathbf{1}_{(a,b]^2}$ and split the square into a lower and an upper triangle.

(b) The above displayed formula is not symmetric in F and G. Show that it can be rewritten in the symmetric form

$$F(b)G(b) - F(a)G(a) = \\ \int_{(a,b]} F(s-) \,\mathrm{d}G(s) + \int_{(a,b]} G(s-) \,\mathrm{d}F(s) + [F,G](b) - [F,G](a),$$

where $[F,G](t) = \sum_{a < s \le t} \Delta F(s) \Delta G(s)$ (for $t \ge a$), with $\Delta F(s) = F(s) - F(s-)$. Note that this sum involves at most countably many terms and is finite.

5.11 Let F be the distribution function of a nonnegative random variable X and $\alpha > 0$. Show (use Exercise 5.10 for instance, or write $\mathbb{E}X^\alpha = \mathbb{E}f(X)$, with $f(x) = \int_0^x \alpha y^{\alpha-1} \,\mathrm{d}y$) that

$$\mathbb{E}X^\alpha = \alpha \int_0^\infty x^{\alpha-1}(1 - F(x)) \,\mathrm{d}x.$$

5.12 Let I be an arbitrary uncountable index set. For each i there is a probability space $(\Omega_i, \mathcal{F}_i, \mathbb{P}_i)$. Define the product σ-algebra $\mathcal{F}$ on $\prod_{i \in I} \Omega_i$ as for the case that I is countable. Call a set C a countable cylinder if it can be written as a product $\prod_{i \in I} C_i$, with $C_i \in \mathcal{F}_i$ and C_i a strict subset of Ω_i for at most countably many indices i.

(a) Show that the collection of countable cylinders is a σ-algebra, that it contains the measurable rectangles and that every set in $\mathcal{F}$ is in fact a countable cylinder.

(b) Let $F = \prod_{i \in I} C_i \in \mathcal{F}$ and let I_F be the set of indices i for which C_i is a strict subset of Ω_i. Define $\mathbb{P}(F) := \prod_{i \in I_F} \mathbb{P}_i(C_i)$. Show that this defines a probability measure on $\mathcal{F}$ with the property that $\mathbb{P}(\pi_i^{-1}[E]) = \mathbb{P}_i(E)$ for every $i \in I$ and $E \in \mathcal{F}_i$.

5.13 Let X be a random variable, defined on some $(\Omega, \mathcal{F}, \mathbb{P})$. Let Y be random variable defined on another probability space $(\Omega', \mathcal{F}', \mathbb{P}')$. Consider the product space, with the product σ-algebra and the product probability measure. Redefine X and Y on the product space by $X(\omega, \omega') = X(\omega)$ and $Y(\omega, \omega') = Y(\omega')$.

Show that the redefined X and Y are independent and that their marginal distributions are the same as they were originally.

5.14 Prove Theorem 5.6. It is sufficient to verify that the proofs hardly differ from those of Lemma 5.3 and Theorem 5.5.

6 Derivative of a measure

The topics of this chapter are absolute continuity and singularity of a pair of measures. The main result is a kind of converse of Proposition 4.23, known as the Radon-Nikodym theorem, Theorem 6.10. In the proof of it that we will give, we need that a continuous map on a Hilbert space can be represented by the inner product with a fixed element in that space. This will be shown in Section 6.2.

6.1 Linear functionals on $\mathbb{R}^n$

Let $H = \mathbb{R}^n$. It is well known that every linear map $T : H \to \mathbb{R}^m$ can uniquely be represented by an $m \times n$ matrix $M = M(T)$ via $T(x) = Mx$ (the usual product of matrix and a vector) , which we will reprove below for the case $m = 1$. Let $m = 1$ and $\langle \cdot, \cdot \rangle$ be the usual inner product on H, $\langle x, y \rangle = x^\top y$. For this case the matrix M becomes a row vector. For $y = M^\top \in \mathbb{R}^n$ one then has

$$T(x) = \langle x, y \rangle. \tag{6.1}$$

Hence we can identify the mapping T with the vector y. Let H^* be the set of all linear maps on H. Then we have for this case the identification of H^* with H itself via equation (6.1).

Suppose that we know that (6.1) holds. Then the kernel K of T is the space of vectors that are orthogonal to y and the orthogonal complement of K is the space of all vectors that are multiples of y. This last observation is the core of the following elementary proof of (6.1).

Let us first exclude the trivial situation in which $T = 0$. Let K be the kernel of T. Then K is a proper linear subspace of H. Take a nonzero vector z in the orthogonal complement of K, a one-dimensional linear subspace of H. Every vector x can be written as a sum $x = \lambda z + u$, with $\lambda \in \mathbb{R}$ and $u \in K$. Then we have

$$\lambda = \frac{\langle x, z \rangle}{\langle z, z \rangle} \text{ and } T(x) = \lambda\, T(z). \tag{6.2}$$

Let $y = \frac{T(z)}{\langle z,z \rangle} z$. Then $\langle x, y \rangle = \frac{T(z)}{\langle z,z \rangle} \langle x, z \rangle = \lambda T(z) = T(x)$, as follows from (6.2). Uniqueness of y is shown as follows. Let $y' \in H$ be such that $T(x) = \langle x, y' \rangle$. Then $\langle x, y - y' \rangle$ is zero for all $x \in H$, in particular for $x = y - y'$. But then $y - y'$ must be the zero vector.

The interesting observation is that this proof carries over to the case where one works with continuous linear functionals on a Hilbert space, which we treat in the next section. We will henceforth write Tx instead of $T(x)$.

6.2 Linear functionals on a Hilbert space

Let H be a (real) Hilbert space, a vector space over the real numbers, endowed with an inner product $\langle \cdot, \cdot \rangle$, that is complete w.r.t. the norm $|| \cdot ||$ generated by

this inner product. Let $T : H \to \mathbb{R}$ be a continuous linear functional on H, there is a $C > 0$ such that $|Tx| \leq C||x||$ for all $x \in H$. We denote by H^* the linear space of all continuous linear functionals on H, also called the dual space of H. We will prove the *Riesz-Fréchet* theorem, which states that every continuous linear functional on H is given by an inner product with a fixed element of H.

Theorem 6.1 *Let $T \in H^*$. Then there exists a unique element $y \in H$ such that $Tx = \langle x, y\rangle$.*

Proof Uniqueness of y follows by the same argument as in the finite dimensional case. We given an alternative proof for the existence and exclude the trivial case $T = 0$. Let K be the kernel of T. Since T is linear and continuous, K is a closed subspace of H and the orthogonal complement $K^\perp$ of K contains a non-zero vector u with $||u|| = 1$. Let $y = (Tu)u \in K^\perp$, then $Ty = (Tu)^2 = ||Ty||^2$. Put $z = (Tx)y - (Ty)x$. Then $Tz = 0$, so $z \in K$ and therefore $\langle z, y\rangle = 0$. This gives $0 = (Tx)\langle y, y\rangle - (Ty)\langle x, y\rangle = (Tu)^2(Tx - \langle x, y\rangle)$ and the result follows since $Tu \neq 0$. □

This theorem can be summarized as follows. The dual space H^* can be identified with H itself. Moreover, we can turn H^* into a Hilbert space itself by defining an inner product $\langle \cdot, \cdot \rangle^*$ on H^*, Let $T, T' \in H^*$ and let y, y' the elements in H that represent T and T' according to the theorem. Then we define $\langle T, T'\rangle^* = \langle y, y'\rangle$. One readily shows that this defines an inner product. Let $|| \cdot ||^*$ be the norm on H^* generated by this inner product. Then H^* is complete as well. Indeed, let (T_n) be a Cauchy sequence in H^* with corresponding elements (y_n) in H, satisfying $T_n x \equiv \langle x, y_n\rangle$. Then $||T_n - T_m||^* = ||y_n - y_m||$. The sequence (y_n) is thus Cauchy in H and has a limit y. Define $Tx = \langle x, y\rangle$. Then T is obviously linear and $||T_n - T||^* = ||y_n - y|| \to 0$. Concluding, we say that the normed spaces $(H^*, || \cdot ||^*)$ and $(H, || \cdot ||)$ are isomorphic.

The usual *operator norm* of a linear functional T on a normed space is defined as $||T||^* = \sup_{x \neq 0} \frac{|Tx|}{||x||}$. It is a simple consequence of the Cauchy-Schwarz inequality that this norm $|| \cdot ||^*$ is the same as the one in the previous paragraph. In fact one can show that continuity of T as a mapping (in the usual sense) is equivalent to finiteness of $||T||^*$. Hence the constant C at the beginning of this section satisfies $C \geq ||T||^*$.

Remark 6.2 Suppose $H = \mathcal{L}^2(S, \Sigma, \mu)$, then H is not a genuine Hilbert space, since $||f||_2 = 0$ only implies $f = 0$ a.e. The assertion of Theorem 6.1 still holds true, except that uniqueness has to be replaced with uniqueness a.e.

6.3 Real and complex measures

Consider a measurable space (S, Σ). A function $\mu : \Sigma \to \mathbb{C}$ is called a *complex* measure if it is countably additive. Such a μ is called a *real* or a *signed* measure if it has its values in $\mathbb{R}$. What we called a measure before, will in this and the next sections sometimes be called a *positive* measure. In the next sections of

this chapter a measure could be either a positive or a complex (or real) measure. Notice that a positive measure can assume the value infinity, unlike a complex measure, whose values lie in $\mathbb{C}$ (see also (6.3)).

Let μ be a complex measure and $E_1, E_2, \ldots$ be disjoint sets in Σ with $E = \bigcup_{i\geq 1} E_i$, then (by definition)

$$\mu(E) = \sum_{i\geq 1} \mu(E_i),$$

where the sum is convergent and the summation is independent of the order. Hence the series is absolutely convergent as well, and we also have

$$|\mu(E)| \leq \sum_{i\geq 1} |\mu(E_i)| < \infty. \tag{6.3}$$

For a given set $E \in \Sigma$ let $\Pi(E)$ be the collection of all *measurable* partitions of E, countable partitions of E with elements in Σ. If μ is a complex measure, then we define

$$|\mu|(E) = \sup\{\sum_i |\mu(E_i)| : E_i \in \pi(E) \text{ and } \pi(E) \in \Pi(E)\}.$$

We will show below that $|\mu|$ is a (positive) measure on (S, Σ) with $|\mu|(S) < \infty$; it is called the *total variation measure* (of μ). Notice that always $|\mu|(E) \geq |\mu(E)|$ and that in particular $\mu(E) = 0$ as soon as $|\mu|(E) = 0$.

In the special case where μ is real valued,

$$\mu^+ = \frac{1}{2}(|\mu| + \mu)$$

and

$$\mu^- = \frac{1}{2}(|\mu| - \mu)$$

define two bounded positive measures such that

$$\mu = \mu^+ - \mu^-. \tag{6.4}$$

This decomposition of the real measure μ is called the Jordan decomposition.

Theorem 6.3 *Let μ be a complex measure on (S, Σ), then $|\mu|$ is a positive measure on (S, Σ), with $|\mu|(S) < \infty$.*

Proof To show that $|\mu|$ is a measure, it suffices to prove countable additivity of $|\mu|$. So let (A_n) be a sequence of disjoint sets of Σ with union A. Choose $0 \leq b_n \leq |\mu|(A_n)$. Let $\varepsilon > 0$ and choose countable measurable partitions (A_{nj}) of the A_n such that for all n

$$b_n - \varepsilon\, 2^{-n} \leq \sum_j |\mu(A_{nj})|.$$

By summing over n and then letting $\varepsilon \to 0$ we get

$$\sum_n b_n \leq \sum_{n,j} |\mu(A_{n,j})| \leq |\mu|(A).$$

Taking now the supremum over all b_n satisfying the constraints, we obtain

$$\sum_n |\mu|(A_n) \leq |\mu|(A).$$

We proceed by showing the converse of this inequality. Let (B_j) be any other measurable partition of A. Then for each fixed n, $(A_n \cap B_j)$ is a partition of A_n. We have

$$\begin{aligned}\sum_j |\mu(B_j)| &= \sum_j |\sum_n \mu(A_n \cap B_j)| \\ &\leq \sum_n \sum_j |\mu(A_n \cap B_j)| \\ &\leq \sum_n |\mu|(A_n).\end{aligned}$$

Taking now the supremum over all partitions (B_j), we obtain

$$|\mu|(A) \leq \sum_n |\mu|(A_n).$$

Next we show that $|\mu|(S)$ is finite. It suffices to show that this is true for a *real* measure. We claim the following fact. If $E \in \Sigma$ is such that $|\mu(E)| = \infty$, then there exist disjoint $A, B \in \Sigma$ such that $A \cup B = E$, $|\mu(A)| > 1$ and $|\mu(B)| = \infty$.

This is proved as follows. Let $m \in \mathbb{N}$ and choose a measurable partition (E_n) of E such that $\sum_n |\mu(E_n)| > m$. Then there exists $N \in \mathbb{N}$ such that $\sum_{n=1}^N |\mu(E_n)| > m$. It follows that there exists a subset J of $\{1, \ldots, N\}$ such that either $\sum_{n\in J} \mu(E_n)^+ > m/2$ or $\sum_{n\in J} \mu(E_n)^- > m/2$. In either case, we have $|\sum_{n\in J} \mu(E_n)| > m/2$. Let $A = \bigcup_{n\in J} E_n$. Then $A \subset E$ and $|\mu(A)| > m/2$. Let then $B = E \setminus A$, for which we have by finiteness of μ, $|\mu(B)| = |\mu(E) - \mu(A)| \geq |\mu(A)| - |\mu(E)| \geq m/2 - |\mu(E)|$. Choose $m > 2(1 + |\mu(E)|)$ to get $|\mu(B)| > 1$ as well as $|\mu(A)| > 1$. Since $E = A \cup B$ and $|\mu(E)| = \infty$, we must have $|\mu(A)| = \infty$ or $|\mu(B)| = \infty$. This proves the claim (possibly after swapping the roles of A and B).

Reasoning by contradiction, we now assume that $|\mu|(S) = \infty$. Let $B_0 = S$, According to our claim, B_0 is the disjoint union of sets A_1 and B_1 with $|\mu(A_1)| > 1$ and $|\mu(B_1)| = \infty$. By induction we construct a sequence (A_n) with $|\mu(A_n)| > 1$ and $|\mu(B_n)| = \infty$. Let $U = \bigcup_{n=1}^\infty A_n$, which by countable additivity of μ should satisfy $\mu(U) = \sum_{n=1}^\infty \mu(A_n)$. But the construction of the A_n was such that this series cannot be convergent, yielding a contradiction. □

Integration w.r.t. a signed measure is defined as can be predicted. Let μ be a signed measure on (S, Σ) and $\mu = \mu^+ - \mu^-$ be its Jordan decomposition. If

$f : S \to \mathbb{R}$ is measurable and both $\int |f| \, \mathrm{d}\mu^+$ and $\int |f| \, \mathrm{d}\mu^-$ are finite, equivalent to $\int |f| \, \mathrm{d}|\mu| < \infty$, we put $\int f \, \mathrm{d}\mu := \int f \, \mathrm{d}\mu^+ - \int f \, \mathrm{d}\mu^-$. For f or μ complex valued, one splits both in their real and imaginary parts and assume that the corresponding real integrals are well defined. Adding up in the obvious way yields the definition of $\int f \, \mathrm{d}\mu$ for this case. If all defining real integrals are finite, one writes $f \in \mathcal{L}^1(S, \Sigma, \mu)$.

For signed measures μ the usual triangle inequality $|\int f \, \mathrm{d}\mu| \leq \int |f| \, \mathrm{d}\mu$ is not valid in general, think of a negative μ. Instead we have the following result.

Proposition 6.4 *Let μ be a real measure on a measurable space (S, Σ) and $|\mu|$ its total variation measure. Assume that $f \in \mathcal{L}^1(S, \Sigma, \mu)$. Then $|\int f \, \mathrm{d}\mu| \leq \int |f| \, \mathrm{d}|\mu|$.*

Proof Exercise 6.12. □

6.4 Absolute continuity and singularity

We start this section with the definition of absolute continuity and singularity for two measures. The former is connected to Section 4.3.

Definition 6.5 Let μ be a positive measure and ν a complex or positive measure on a measurable space (S, Σ). We say that ν is *absolutely continuous* w.r.t. μ (notation $\nu \ll \mu$), if $\nu(E) = 0$ for every $E \in \Sigma$ with $\mu(E) = 0$. Two arbitrary measures μ and ν on (S, Σ) are called *mutually singular* (notation $\nu \perp \mu$) if there exist disjoint sets E and F in Σ such that $\nu(A) = \nu(A \cap E)$ and $\mu(A) = \mu(A \cap F)$ for all $A \in \Sigma$.

An example of absolute continuity we have seen already in the previous section: $\mu \ll |\mu|$ for a complex measure μ. Another example is provided by the measures ν and μ of (4.7), $\nu \ll \mu$. See also Proposition 6.8 below. Note that for two mutually singular measures μ and ν one has $\nu(F) = \mu(E) = 0$, where E and F are as in Definition 6.5.

Proposition 6.6 *Let μ be a positive measure and ν_1, ν_2 arbitrary measures, all defined on the same measurable space. Then the following properties hold true.*

(i) If $\nu_1 \perp \mu$ and $\nu_2 \perp \mu$, then $\nu_1 + \nu_2 \perp \mu$.

(ii) If $\nu_1 \ll \mu$ and $\nu_2 \ll \mu$, then $\nu_1 + \nu_2 \ll \mu$.

(iii) If $\nu_1 \ll \mu$ and $\nu_2 \perp \mu$, then $\nu_1 \perp \nu_2$.

(iv) If $\nu_1 \ll \mu$ and $\nu_1 \perp \mu$, then $\nu_1 = 0$.

Proof Exercise 6.2. □

Proposition 6.7 *Let μ be a positive measure and ν_a and ν_s be arbitrary real or complex measures on (S, Σ). Assume that $\nu_a \ll \mu$ and $\nu_s \perp \mu$. Put*

$$\nu = \nu_a + \nu_s. \tag{6.5}$$

Suppose that ν also admits the decomposition $\nu = \nu'_a + \nu'_s$ with $\nu'_a \ll \mu$ and $\nu'_s \perp \mu$. Then $\nu'_a = \nu_a$ and $\nu'_s = \nu_s$.

Proof It follows that

$$\nu'_a - \nu_a = \nu_s - \nu'_s,$$

$\nu'_a - \nu_a \ll \mu$ and $\nu_s - \nu'_s \perp \mu$ (Proposition 6.6), and hence both are zero (Proposition 6.6 again). □

The content of Proposition 6.7 is that the decomposition (6.5) of ν, if it exists, is unique. We will see in Section 6.5 that, given a positive measure μ, such a decomposition exists for any measure ν and it is called the *Lebesgue decomposition* of ν w.r.t. μ. We extend the definition of the measure ν as given in (4.7) to the real and complex case.

Proposition 6.8 *Let μ be a positive measure on (S, Σ) and h a nonnegative measurable function on S. Then the map $\nu : \Sigma \to [0, \infty]$ defined by*

$$\nu(E) = \mu(\mathbf{1}_E h) \tag{6.6}$$

is a positive measure on (S, Σ) that is absolutely continuous w.r.t. μ. If h is complex valued and in $\mathcal{L}^1(S, \Sigma, \mu)$, then ν is a complex measure.

Proof See Exercise 4.9 for nonnegative h. The other case is Exercise 6.3. □

The Radon-Nikodym theorem of the next section states that every σ-finite measure ν that is absolutely continuous w.r.t. μ is of the form (6.6). We will use in that case the notation

$$h = \frac{\mathrm{d}\nu}{\mathrm{d}\mu}.$$

In the next section we use

Lemma 6.9 *Let μ be a finite positive measure and $f \in \mathcal{L}^1(S, \Sigma, \mu)$, possibly complex valued. Let A be the set of averages*

$$a_E = \frac{1}{\mu(E)} \int_E f \, \mathrm{d}\mu,$$

where E runs through the collection of sets $E \in \Sigma$ with $\mu(E) > 0$. Then $\mu(\{f \notin \bar{A}\}) = 0$.

Proof Assume that $\mathbb{C} \setminus \bar{A}$ is not the empty set (otherwise there is nothing to prove) and let B be a closed ball in $\mathbb{C} \setminus \bar{A}$ with center c and radius $r > 0$. Notice that $|c - a| > r$ for all $a \in \bar{A}$. It is sufficient to prove that $E = f^{-1}[B]$ has measure zero, since $\mathbb{C} \setminus \bar{A}$ is a countable union of such balls (argue as in the proof of Proposition 1.3). Suppose that $\mu(E) > 0$. Then we would have

$$|a_E - c| \leq \frac{1}{\mu(E)} \int_E |f - c| \, \mathrm{d}\mu \leq r.$$

But this is a contradiction since $a_E \in A$. □

6.5 The Radon-Nikodym theorem

As an appetizer for the Radon-Nikodym theorem (Theorem 6.10) we consider a special case. Let S be a finite or countable set and $\Sigma = 2^S$. Let μ be a positive σ-finite measure on (S, Σ) and ν another finite measure such that $\nu \ll \mu$. Define $h(x) = \frac{\nu(\{x\})}{\mu(\{x\})}$ if $\mu(\{x\}) > 0$ and zero otherwise. It is easy to verify that $h \in \mathcal{L}^1(S, \Sigma, \mu)$ and

$$\nu(E) = \mu(\mathbf{1}_E h), \forall E \subset S. \tag{6.7}$$

Observe that we have obtained an expression like (6.6), but now starting from the assumption $\nu \ll \mu$. The principal theorem on absolute continuity (and singularity) is the following.

Theorem 6.10 *Let μ be a positive σ-finite measure and ν a complex measure. Then there exists a unique decomposition $\nu = \nu_a + \nu_s$ and a function $h \in \mathcal{L}^1(S, \Sigma, \mu)$ such that $\nu_a(E) = \mu(\mathbf{1}_E h)$ for all $E \in \Sigma$ (so $\nu_a \ll \mu$) and $\nu_s \perp \mu$. Moreover, h is unique in the sense that any other h' with this property is such that $\mu(\{h \neq h'\}) = 0$. The function h is called the Radon-Nikodym derivative of ν_a w.r.t. μ and is often written as*

$$h = \frac{\mathrm{d}\nu_a}{\mathrm{d}\mu}.$$

Proof Uniqueness of the decomposition $\nu = \nu_a + \nu_s$ is the content of Proposition 6.7. Hence we proceed to show existence. Let us first assume that $\mu(S) < \infty$ and that ν is positive and finite.

Consider then the positive bounded measure $\phi = \nu + \mu$. Let $f \in \mathcal{L}^2(S, \Sigma, \phi)$. The Cauchy-Schwarz inequality (Remark 4.47) gives

$$|\nu(f)| \leq \nu(|f|) \leq \phi(|f|) \leq (\phi(f^2))^{1/2}(\phi(S))^{1/2}.$$

We see that the linear map $f \mapsto \nu(f)$ is bounded on $\mathcal{L}^2(S, \Sigma, \phi)$. Hence there exists, by virtue of the Riesz-Fréchet Theorem 6.1 and Remark 6.2, a $g \in \mathcal{L}^2(S, \Sigma, \phi)$ such that for all f

$$\nu(f) = \phi(fg). \tag{6.8}$$

Take $f = \mathbf{1}_E$ for any E with $\phi(E) > 0$. Then $\phi(E) \geq \nu(E) = \phi(\mathbf{1}_E g) \geq 0$ so that the average $\frac{1}{\phi(E)}\phi(\mathbf{1}_E g)$ lies in $\in [0, 1]$. From Lemma 6.9 we obtain that $\phi(\{g \notin [0, 1]\}) = 0$. Replacing g with $g\mathbf{1}_{\{0 \leq g \leq 1\}}$, we see that (6.8) still holds and hence we may assume that $0 \leq g \leq 1$.

Rewrite (6.8) as

$$\nu(f(1 - g)) = \mu(fg) \tag{6.9}$$

and take $f = \mathbf{1}_{\{g=1\}}$ to obtain $\mu(\{g = 1\}) = 0$. Define the positive measure ν_s on Σ by $\nu_s(E) = \nu(E \cap \{g = 1\})$. Then $\nu_s(\{g < 1\}) = \nu_s(\emptyset) = 0$. It follows that $\nu_s \perp \mu$. Define the measurable function

$$h = \frac{g}{1 - g}\mathbf{1}_{\{g<1\}}, \tag{6.10}$$

and the measure ν_a on Σ by $\nu_a(E) = \mu(\mathbf{1}_E h)$. By Proposition 6.8 this indeed defines a measure and obviously $\nu_a \ll \mu$.

Let $f = \frac{\mathbf{1}_{E\cap\{g<1\}}}{1-g}$ and note that $= \lim_n f_n$, where $f_n = \mathbf{1}_{E\cap\{g<1\}} \sum_{k=0}^{n-1} g^k$. Apply (6.9) to the f_n and apply monotone convergence to obtain

$$\nu(E \cap \{g < 1\}) = \nu_a(E).$$

It follows that $\nu = \nu_a + \nu_s$, which is the desired decomposition and the function h of (6.10) is as required. Since $\mu(h) = \nu_a(S) < \infty$, we also see that $h \in \mathcal{L}^1(S, \Sigma, \mu)$. Uniqueness of h is left as Exercise 6.6.

If μ is not bounded but merely σ-additive and ν bounded and positive we decompose S into a measurable partition $S = \bigcup_{n\geq 1} S_n$, with $\mu(S_n) < \infty$. Apply the previous part of the proof to each of the spaces (S_n, Σ_n) with Σ_n the trace σ-algebra of Σ on S_n. This yields measures $\nu_{a,n}$ and functions h_n defined on the S_n. Put then $\nu_a(E) = \sum_n \nu_{a,n}(E \cap S_n)$, $h = \sum_n \mathbf{1}_{S_n} h_n$. Then $\nu_a(E) = \mu(\mathbf{1}_E h)$ and $\mu(h) = \nu_a(S) < \infty$. For real measures ν we apply the results to ν^+ and ν^- and finally, if ν is complex we treat the real and imaginary part separately. The boring details are omitted. □

Remark 6.11 If ν is a positive σ-finite measure, then the Radon-Nikodym theorem is still true with the exception that we only have $\mu(h\mathbf{1}_{S_n}) < \infty$, where the S_n form a measurable partition of S such that $\nu(S_n) < \infty$ for all n. Notice that in this case (inspect the proof above) we may still take $h \geq 0$.

Remark 6.12 The function h of Theorem 6.10, the Radon-Nikodym derivative of ν_a w.r.t. μ, is also called the *density* of ν_a w.r.t. μ. If λ is Lebesgue measure on $(\mathbb{R}, \mathcal{B})$ and ν is the law of a random variable X that is absolutely continuous w.r.t. λ, we have that $F(x) := \nu((-\infty, x]) = \int_{(-\infty,x]} f \,\mathrm{d}\lambda$, where $f = \frac{\mathrm{d}\nu}{\mathrm{d}\lambda}$. Traditionally, the function f was called the density of X, and we see that calling a Radon-Nikodym derivative a density is in agreement with this tradition, but also extends it.

Theorem 6.10 is often used for probability measures $\mathbb{Q}$ and $\mathbb{P}$ with $\mathbb{Q} \ll \mathbb{P}$. Write $Z = \frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}$ and note that $\mathbb{E}Z = 1$. It is immediate from Proposition 4.23 that for $X \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{Q})$ one has

$$\mathbb{E}_{\mathbb{Q}} X = \mathbb{E}[XZ], \tag{6.11}$$

where $\mathbb{E}_{\mathbb{Q}}$ is used to denote expectation under the probability measure $\mathbb{Q}$.

6.6 Decomposition of a distribution function

In elementary probability one often distinguishes between distribution functions that are of pure jump type (for discrete random variables) and those that admit an ordinary density. These can both be recognized as examples of the following result.

Proposition 6.13 *Let F be a distribution function, $F : \mathbb{R} \to \mathbb{R}$. Then there exists a purely discontinuous right-continuous nondecreasing function F_d, with $\lim_{x\to-\infty} F_d(x) = 0$, a nonnegative Borel-measurable function f and a nondecreasing continuous function F_s with $\lim_{x\to-\infty} F_s(x) = 0$ such that the decomposition*

$$F = F_d + F_s + F_{ac},$$

holds true, with F_{ac} defined by $F_{ac}(x) = \int_{-\infty}^{x} f(y)\,\mathrm{d}y$. Such a decomposition is unique.

Proof Since F is increasing, it has at most countably many discontinuities, collected in a set D. Define

$$F_d(x) = \sum_{y\in D\cap(-\infty,x]} \Delta F(y).$$

One verifies that F_d has the asserted properties, the set of discontinuities of F_d is also D and that $F_c := F - F_d$ is continuous. Up to the normalization constant $1 - F_d(\infty)$, F_c is a distribution function if $F_d(\infty) < 1$ and equal to zero if $F_d(\infty) = 1$. Hence there exists a subprobability measure μ_c on $\mathcal{B}$ such that $\mu_c((-\infty, x]) = F_c(x)$, Theorem 3.10. According to the Radon-Nikodym theorem, we can split $\mu_c = \mu_{ac} + \mu_s$, where μ_{ac} is absolutely continuous w.r.t. Lebesgue measure λ. Hence, there exists a λ-a.e. unique function f in $\mathcal{L}^1_+(\mathbb{R}, \mathcal{B}, \lambda)$ such that $\mu_{ac}(B) = \int_B f\,\mathrm{d}\lambda$. □

We have already encountered two examples, where the above decomposition consists of a single term only. If a random variable X has a discrete distribution, there are x_k, $k = 1, 2, \ldots$ with $\sum_{k\geq 1} \mathbb{P}(X = x_k) = 1$, $F = F_d$, and if the distribution function admits a density f, then $F = F_{ac}$. Another extreme case occurs when F is the distribution function of Exercise 6.5, then $F = F_s$. Think of an example of a random variable for which all three terms in the decomposition of Proposition 6.13 are nontrivial.

6.7 The fundamental theorem of calculus

The classical fundamental theorem of calculus states the following. If $f : [0, \infty) \to \mathbb{R}$ is continuous and $F : [0, \infty) \to \mathbb{R}$ is defined by $F(x) = \int_0^x f(y)\,\mathrm{d}y$ (as a Riemann or a Lebesgue integral), then F is (continuously) differentiable on $(0, \infty)$ and $F' = f$. If the requirement that f is continuous is not satisfied, but $f \in \mathcal{L}([0, n], \mathcal{B}[0, n], \lambda)$ for all $n \in \mathbb{N}$, then F is still well defined, but not necessarily everywhere differentiable. For instance, if $f = \mathbf{1}_{[0,1]}$, then $F(x) = x \wedge 1$. In this case F is not differentiable in 1, but $F'(x) = f(x)$ for all $x \in (0, \infty) \setminus \{1\}$. More generally, if f is piecewise continuous, $f = \sum_{i=1}^{\infty} \mathbf{1}_{[x_{i-1},x_i)} f_i$, where the x_i are increasing to infinity, $x_1 = 0$ and the f_i are continuous on the $[x_{i-1}, x_i]$, then F is differentiable on $(0, \infty) \setminus \{x_0, x_1, \ldots\}$ with $F' = f$. In both cases the 'exceptional set' where F is not differentiable, has Lebesgue measure zero. Here

is another example. Let $f = \mathbf{1}_{\mathbb{Q}\cap[0,\infty)}$. Then $F = 0$, and trivially also $F'(x) = 0$ for all $x > 0$, equal to $f(x)$ outside $\mathbb{Q}$. In this example, F' exists everywhere on $(0, \infty)$, but only coincides with f outside a set of Lebesgue measure zero. This all suggests a generalization of the classical fundamental theorem of calculus, Theorem 6.18 below. We need some preparations, formulated as the next three lemmas, but first a definition.

Definition 6.14 Let (S, Σ, μ) be a measure space and suppose that S is endowed with a topology. A set $A \in \Sigma$ is called regular if

$$\mu(A) = \sup\{\mu(K) : K \in \Sigma, K \text{ is compact and } K \subset A\}.$$

Usually, if S is topological Σ is the collection of Borel sets in which case the compact sets K in this definition automatically belong to Σ.

Lemma 6.15 *Let μ be a finite measure on the Borel sets of a finite interval $I = [0, b]$. Then all sets in $\mathcal{B}([0, b])$ are regular. In particular, if B is Borel with $\mu(B) = 0$, for any $\varepsilon > 0$, there exists an open set U in $[0, b]$ such that $U \supset B$ and $\mu(U) < \varepsilon$.*

Proof First one shows (Exercise 6.13) that the collection of sets A in $\mathcal{B}([0, b])$ that are regular as well as their complements A^c form a σ-algebra, Σ say. Next, every closed set in $[0, b]$ is compact and thus automatically regular. Let then U be open. The sets $U_n := \{x : d(x, U^c) \geq \frac{1}{n}\}$ (d the ordinary metric) are closed and their union is U. It follows that U is regular. Hence Σ contains all open (and closed) sets in $[0, b]$ and thus $\Sigma = \mathcal{B}([0, b])$.

To prove the second assertion, we observe that $\mu(B^c) = \mu([0, b])$ and that B^c is regular. Hence there is a compact set $K \subset [0, b]$ such that $K \subset B$ and $\mu(B^c \setminus K) < \varepsilon$. But then the open set K^c contains B and $\mu(K) < \varepsilon$. □

Lemma 6.16 *Let $\mathcal{U}$ be a collection of open intervals in $\mathbb{R}$ with bounded union U. For all $t < \lambda(U)$, there exists $q = q(t) \in \mathbb{N}$ and disjoint intervals $V_1, \ldots, V_q \in \mathcal{U}$ such that $t < 3\sum_{i=1}^{q} \lambda(V_i)$.*

Proof Let $t < \lambda(U)$. Because U is an interval, by the continuity of λ there exists a compact interval K with $K \subset U$ and $\lambda(K) > t$. Since $\mathcal{U}$ is an open cover of K, there are finitely many $U_1, \ldots, U_n$ in $\mathcal{U}$ such that $\bigcup_{i=1}^{n} U_i \supset K$. Order the U_i such that the $\lambda(U_i)$ are non-increasing. Let $V_1 := U_1$. Suppose $U_i \cap V_1 \neq \emptyset$ for all $i \geq 2$. Write $U_i = (m_i - \delta_i, m_i + \delta_i)$ and $W_1 = (m_1 - 3\delta_1, m_1 + 3\delta_1)$. For $U_i \cap V_1 \neq \emptyset$, we have for instance $m_i - \delta_i < m_1 + \delta_1$, hence $m_i + \delta_i < m_i - \delta_i + 2\delta_i < m_1 + 3\delta_1$. Hence $m_i + \delta_i \in (m_1 - 3\delta_1, m_1 + 3\delta_1) =: W_1$. It follows that $U_i \subset W_1$, for all $i \geq 1$, and hence $\bigcup_{i=1}^{n} U_i \subset W_1$) and $t < \lambda(\bigcup_{i=1}^{n} U_i) \leq \lambda(W_1) = 3\lambda(V_1)$. In this case we set $q = 1$.

If there is U_i such that $V_1 \cap U_i = \emptyset$, we define $V_2 = U_{k_2}$, where k_2 is the smallest index in $\{2, \ldots, n\}$ such that $V_1 \cap U_k = \emptyset$. If $(V_1 \cup V_2) \cap U_i \neq \emptyset$ for all $i \geq 2$, we set $q = 2$ and next to W_1 above, we construct W_2 like W_1 above by 'blowing up' V_2 by a factor 3. For $i \neq 1, k_2$, we have $U_i \neq V_1, V_2$ and $U_i \cap V_1 \neq \emptyset$

or $U_i \cap V_2 \neq \emptyset$. It follows, by parallel reasoning as above that $U_i \subset W_1$ or $U_i \subset W_2$. Hence all $U_i \subset (W_1 \cup W_2)$ and the same holds for their union. So $\lambda(\bigcup_{i=1}^n U_i) \leq \lambda(W_1) + \lambda(W_2) = 3(\lambda(V_1) + \lambda(V_2))$.

Here is the general case. Choose recursively for $i \geq 2$, V_i such that $V_i = U_{k_i}$, where k_i is the smallest integer k in $\{i, \ldots, n\}$ for which $U_k \cap \bigcup_{j=1}^{i-1} V_j = \emptyset$. The largest i for which this is possible is q. Then any U_k is either one of the V_i (and thus contained in the corresponding W_i), or it intersects one of the V_i, in which case it is also contained in the corresponding W_i. It follows that $\bigcup_{k=1}^n U_k \subset \bigcup_{i=1}^q W_i$, from which the assertion follows. □

Lemma 6.17 *Let μ be a probability (or a finite) measure on the Borel sets of an interval $[a, b]$. Let $F(x) = \mu([a, x])$ be its distribution function. If A is a Borel subset of $[a, b]$ with $\mu(A) = 0$, then F is differentiable on a subset A_0 of A with $\lambda(A \setminus A_0) = 0$ and $F'(x) = 0$ for all x in A_0.*

Proof A little reflection shows that it is sufficient to show that

$$\lim_{h \downarrow 0} \frac{\mu((x-h, x+h))}{h} = 0, \lambda\text{-a.e. on } A. \tag{6.12}$$

To that end it we shall show that the sets A_j defined for $j = 1, 2, \ldots$ by

$$N_j = \{x \in A : \limsup_{h \downarrow 0} \frac{\mu((x-h, x+h))}{h} > \frac{1}{j}\}$$

have Lebesgue measure zero. Indeed on the set $N = \bigcup_{j=1}^{\infty}$, satisfying $\lambda(N) = 0$, the limit in (6.12) may not exist or may be different from zero. First we argue that the N_j are measurable. It is easy to verify that $(x, y, h) \to \mathbf{1}_{(x-h,x+h)}(y)$ is a measurable function of its three arguments. By the construction leading to Fubini's theorem (Lemma 5.3), the mapping $(x, h) \to \int \mathbf{1}_{(x-h,x+h)}\, \mathrm{d}\lambda = \mu((x-h, x+h))$ is measurable as well. Realizing that by the continuity of the measure, the map $h \to \mu((x-h, x+h))$ is left continuous, for computing the lim sup in the definition of the N_j we can restrict ourself to rational sequences, which makes the lim sup a measurable function of x. Consequently, P_j is Borel-measurable.

Let $\varepsilon > 0$. Since $\mu(A) = 0$, by virtue of Lemma 6.15 there exists an open $V \supset A$, with $\mu(V) < \varepsilon$. Let $x \in P_j$. Then there is $h > 0$ such that $(x - h, x + h) \subset V$ and $\mu((x-h, x+h)) > h/j$. (Note that the latter entails $2j\mu((x-h, x+h)) > \lambda((x-h, x+h))$.) All such intervals cover P_j and if $t < \lambda(P_j)$, we can choose from them according to Lemma 6.16 disjoint intervals $J_1, \ldots, J_q$ all contained in V such that (here it is essential that the intervals are disjoint)

$$t \leq 3 \sum_{i=1}^{q} \lambda(J_i) \leq 6j \sum_{i=1}^{q} \mu(J_i) = 6j\lambda(\bigcup_{i=1}^{q} J_i) \leq 6j\lambda(V) \leq 6j\varepsilon.$$

It follows that $\lambda(P_j) = 0$ as as claimed. □

Here is the result we aim at, the generalized fundamental theorem of calculus.

Theorem 6.18 *Let $f : \mathbb{R} \to \mathbb{R}$ be measurable such that $\int_K |f| \, \mathrm{d}\lambda < \infty$ for every compact subset K of $\mathbb{R}$. Define for any $a \in \mathbb{R}$ the function $F : [a, \infty) \to \mathbb{R}$ by $F(x) = \int_{(a,x]} f \, \mathrm{d}\lambda$. Then outside a set N of Lebesgue measure zero, F is differentiable and $F'(x) = f(x)$, for all x not belonging to N.*

Proof The proof involves the countably many nonnegative functions $f_r := (f - r)^+$, $r \in \mathbb{Q}$. Consider the measures μ_r defined by $\mu_r(E) = \int_E f_r \, \mathrm{d}\lambda$ and the function F_r defined by $F_r(x) = \int_{[a,x]} f_r \, \mathrm{d}\lambda$. It follows that $\mu_r(\{f \leq r\}) = 0$ and by Lemma 6.17, the sets $B_r := \{f \leq r\} \cap Z_r^c$ have Lebesgue measure zero, where Z_r is the set of x where $F_r'(x)$ exists and is equal to zero. It follows that $\lambda(B) = 0$, where $B = \bigcup_{r \in \mathbb{Q}} B_r$.

We investigate what happens on B^c. Let $x \in B^c$, then $x \in B_r^c$ for all $r \in \mathbb{Q}$. For any $r \in \mathbb{Q}$ with $r > f(x)$, we must then have $x \in Z_r$. For $h > 0$ we have

$$\int_x^{x+h} f \, \mathrm{d}\lambda \leq rh + \int_x^{x+h} f_r \, \mathrm{d}\lambda,$$

and we obtain

$$\limsup_{h \downarrow 0} \frac{1}{h} \int_x^{x+h} f \, \mathrm{d}\lambda \leq r.$$

Letting $r \downarrow f(x)$, we have

$$\limsup_{h \downarrow 0} \frac{1}{h} \int_x^{x+h} f \, \mathrm{d}\lambda \leq f(x). \tag{6.13}$$

Switching from f to $-f$, we obtain

$$\liminf_{h \downarrow 0} \frac{1}{h} \int_x^{x+h} f \, \mathrm{d}\lambda \geq f(x). \tag{6.14}$$

A combination of Equations (6.13) and (6.14) shows that F is right-differentiable at x, with right derivative $f(x)$. By the same token, starting from

$$\int_{x-h}^{x} f \, \mathrm{d}\lambda \leq rh + \int_{x-h}^{x} f_r \, \mathrm{d}\lambda,$$

for sufficiently small h, one obtains that F is left-differentiable at x with left-derivative also equal to $f(x)$. This finishes the proof of showing that F is almost everywhere differentiable. □

We can go one step further. Also functions F that cannot be written as integrals of f w.r.t. Lebesgue measure may be differentiable outside a set of Lebesgue measure zero. Take for instance for F the Cantor function of Exercise 6.5, then *outside the Cantor set*, F is differentiable with $F' = 0$.

Theorem 6.19 *Let $F : [0, \infty) \to \mathbb{R}$ be an increasing function. Then F is differentiable in Lebesgue almost all points of $(0, \infty)$. Denote by F' the derivative in those points where it exists, and zero else. Then F' is integrable w.r.t. λ on any interval $[0, b]$.*

Proof Let $G : [0, \infty) \to \mathbb{R}$ be defined by $G(x) = F(x+)$, then G is right-continuous and also increasing. Let $H = F - G$. Then H is zero with the exception of at most countably many points. It follows from Lemma 6.17 that H is differentiable with derivative zero (Lebesgue) almost everywhere on $(0, \infty)$.

We continue to show the assertion on differentiability of G on an arbitrary interval $(0, b)$. To avoid trivialities, we assume that $G(b) > 0$. Up to the normalization by $G(b)$, G is a distribution function on $[0, b]$. By Proposition 6.13, G can be decomposed as $G = G_d + G_s + G_{ac}$, where G_d and G_s have a derivative almost everywhere equal to zero, again by Lemma 6.17, whereas one has $G_{ac}(x) = \int_0^x g(y)\, \mathrm{d}y$ for some nonnegative integrable function g on $[0, b]$. Taking $F' = g$ completes the proof of existence of a derivative on an interval $(0, b)$. Since g is unique a.e. on an interval $[0, b]$, we also have a.e. uniqueness of the derivative on the whole positive real line $(0, \infty)$ (take $b = n$ with $n \in \mathbb{N}$ and $n \to \infty$). □

6.8 Dual spaces

Recall that the dual space of a normed space is the linear space of all bounded linear functionals. In this section we show that for $p \in [1, \infty)$ the dual space of $L^p(\Omega, \mathcal{F}, \mathbb{P})$ is isomorphic to $L^q(\Omega, \mathcal{F}, \mathbb{P})$ with $q = \frac{p}{p-1}$ for all $p \geq 1$. The results below can be extended to general L^p spaces.

Theorem 6.20 *Let $p \in [1, \infty)$. The dual space of $L^p(\Omega, \mathcal{F}, \mathbb{P})$ is $L^q(\Omega, \mathcal{F}, \mathbb{P})$, where $\frac{1}{p} + \frac{1}{q} = 1$.*

Proof Let $T : L^p(\Omega, \mathcal{F}, \mathbb{P}) \to \mathbb{R}$ be a bounded linear operator. Define on $\mathcal{F}$ the map ν by

$$\nu(F) = T(\mathbf{1}_F). \tag{6.15}$$

Obviously, ν is by linearity of T an additive map and by continuity of T even σ-additive. Indeed, if $F_n \downarrow \emptyset$, then $|\nu(F_n)| \leq ||T||\, \mathbb{P}(F_n)^{1/p} \downarrow 0$. Hence ν is a finite signed measure on $\mathcal{F}$, that is absolute continuous w.r.t. $\mathbb{P}$. It follows from the Radon-Nikodym theorem that there is $Y \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ such that

$$\nu(F) = \mathbb{E}[\mathbf{1}_F Y], \quad \forall F \in \mathcal{F}. \tag{6.16}$$

Case 1: $p = 1$. We show that Y is a.s. bounded. Let $F = \{Y > c\}$, for some $c > 0$. By continuity of T, we have

$$c\, P(Y > c) \leq \mathbb{E}[\mathbf{1}_F Y] = |T(\mathbf{1}_F)| \leq ||T||\, ||\mathbf{1}_F||_1 = ||T||\, \mathbb{P}(Y > c).$$

Hence, if $\mathbb{P}(Y > c) > 0$ it follows that $||T|| \geq c$. Stated otherwise $||T|| \geq \sup\{c > 0 : \mathbb{P}(Y > c) > 0\}$. A similar argument yields $||T|| \geq \sup\{c > 0 : \mathbb{P}(Y < -c) > 0\}$. It follows that $||Y||_\infty \leq ||T|| < \infty$.

We finally show that for every $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$, it holds that $T(X) = \mathbb{E}[XY]$. By the above construction this is true for X of the form $X = \mathbf{1}_F$. Hence also for (nonnegative) simple functions. Let $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ be arbitrary. Choose a sequence (X_n) of simple functions such that $X_n \to X$ a.s. and $|X_n| \leq |X|$. Then, by dominated convergence, $X_n \to X$ in $L^1(\Omega, \mathcal{F}, \mathbb{P})$ as well and, since Y is a.s. bounded, we also have $X_n Y \to XY$ in $L^1(\Omega, \mathcal{F}, \mathbb{P})$. But then $T(X) = \mathbb{E}[XY]$.

Case 2: $p \in (1, \infty)$. We start from Equation (6.16). It follows that for every $X \in L^\infty((\Omega, \mathcal{F}, \mathbb{P}))$ one has $T(X) = \mathbb{E}XY$. For every $n \in \mathbb{N}$, put $X_n = \operatorname{sgn}(Y)|Y|^{q-1}\mathbf{1}_{E_n}$, where $E_n = \{|Y| \leq n\}$. Note that every X_n is bounded, $X_n Y = |Y|^q \mathbf{1}_{E_n}$ and $|X_n|^p = |Y|^q \mathbf{1}_{E_n}$. We obtain

$$\mathbb{E}|Y|^q \mathbf{1}_{E_n} = T(X_n) \leq ||T|| \cdot ||X_n||_p = ||T|| \cdot (\mathbb{E}|Y|^q \mathbf{1}_{E_n})^{1/p},$$

from which it follows that $(\mathbb{E}|Y|^q \mathbf{1}_{E_n})^{1/q} \leq ||T||$ (here it is used that $p > 1$). By letting $n \to \infty$, we obtain $||Y||_q \leq ||T|| < \infty$, so $Y \in L^q(\Omega, \mathcal{F}, \mathbb{P})$.

Finally, for $X \in L^p(\Omega, \mathcal{F}, \mathbb{P})$ we put $X_n = X\mathbf{1}_{\{|X| \leq n\}}$, so that $X_n \in L^\infty(\Omega, \mathcal{F}, \mathbb{P})$ and $||X - X_n||_p \to 0$. It follows by Hölder's inequality (used in the fourth step) that

$$\begin{aligned} T(X) &= T(X - X_n) + T(X_n) \\ &= T(X - X_n) + \mathbb{E}X_n Y \\ &= T(X - X_n) + \mathbb{E}(X_n - X)Y + \mathbb{E}XY \\ &\to \mathbb{E}XY. \end{aligned}$$

In both cases the random variable Y is a.s. unique. Indeed, if Y and Y' satisfy $\mathbb{E}[XY] = \mathbb{E}[XY'] = T(X)$ for all $X \in L^p(\Omega, \mathcal{F}, \mathbb{P})$, we can choose $X = \operatorname{sgn}(Y - Y')$ to obtain this result. □

Remark 6.21 In the proof of Lemma 6.20 one actually has that $||Y||_q = ||T||$ (Exercise 6.15). The Riesz-Fréchet Theorem 6.1 can be applied to show that the dual space of $L^2(\Omega, \mathcal{F}, \mathbb{P})$ is isomorphic to itself, see also Remark 6.2 for the $\mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$ case.

The dual space of $L^\infty(\Omega, \mathcal{F}, \mathbb{P})$ is less nice. Still, ν as in (6.15) defines a *finitely* additive signed measure, even absolutely continuous w.r.t. $\mathbb{P}$, but the story basically stops at this point, ν will in general not be countably additive and the Radon-Nikodym theorem cannot be applied. For some special $L^\infty(S, \Sigma, \mu)$ alternative characterizations exits. Although any $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ induces a bounded linear functional on $L^\infty(\Omega, \mathcal{F}, \mathbb{P})$, the converse is not true and so $L^1(\Omega, \mathcal{F}, \mathbb{P})$ is not the dual of $L^\infty(\Omega, \mathcal{F}, \mathbb{P})$. Here is a quick sketch of a counterexample.

Take $(\Omega, \mathcal{F}, \mathbb{P}) = ((-\frac{1}{2}, \frac{1}{2}), \mathcal{B}, \lambda)$ and define T by $Tf = f(0)$. For every $x_0 \neq 0$ and $f = \mathbf{1}_{[x_0 - \varepsilon, x_0 + \varepsilon]}$ we have $Tf = 0$ for all sufficiently small positive ε.

Suppose there is $g \in L^1((-\frac{1}{2},\frac{1}{2}),\mathcal{B},\lambda)$ such that $Tf \equiv \int_{(-\frac{1}{2},\frac{1}{2})} fg \, \mathrm{d}\lambda$. Then for $x_0 \neq 0$ and all small ε

$$0 = \frac{1}{2\varepsilon}\int_{[x_0-\varepsilon,x_0+\varepsilon]} g \, \mathrm{d}\lambda \to g(x_0),$$

in view of Theorem 6.18 for all x_0 outside a set of measure zero. Hence $g = 0$ a.s. and then $\int_{(-\frac{1}{2},\frac{1}{2})} fg \, \mathrm{d}\lambda = 0$ for all $f \in L^\infty((-\frac{1}{2},\frac{1}{2}),\mathcal{B},\lambda)$. But T is not identically zero.

6.9 Additional results

We present some more properties of complex measures and their total variation measure.

Proposition 6.22 *Let μ be a complex measure. Then $\mu \ll |\mu|$ and the Radon-Nikodym derivative $h = \frac{\mathrm{d}\mu}{\mathrm{d}|\mu|}$ may be taken such that $|h| = 1$.*

Proof That $\mu \ll |\mu|$ has already been observed in Section 6.3. Let h be any function as in the Radon-Nikodym theorem applied to $\mu \ll |\mu|$. Since $||\mu|(h\mathbf{1}_E)| = |\mu(E)| \leq |\mu|(E)$, it follows from Lemma 6.9 that $|\mu|(\{|h| > 1\}) = 0$. On the other hand, for $A = \{|h| \leq r\}$ $(r > 0)$ and a measurable partition with elements A_j of A, we have

$$\sum_j |\mu(A_j)| = \sum_j |\mu|(\mathbf{1}_{A_j} h) \leq \sum_j |\mu|(\mathbf{1}_{A_j}|h|) \leq r|\mu|(A).$$

Then we find, by taking suprema over such partitions, that $|\mu|(A) \leq r|\mu|(A)$. Hence for $r < 1$ we find $|\mu|(A) = 0$ and we conclude that $|\mu|(\{|h| < 1\}) = 0$. Combining this with the previous result we get $|\mu|(\{|h| \neq 1\}) = 0$. The function that we look for, is $h\mathbf{1}_{\{|h|=1\}} + \mathbf{1}_{\{|h|\neq 1\}}$. □

Corollary 6.23 *Let μ be a real measure, $h = \frac{\mathrm{d}\mu}{\mathrm{d}|\mu|}$. Then for any $E \in \Sigma$ we have $\mu^+(E) = |\mu|(\mathbf{1}_{E\cap\{h=1\}})$ and $\mu^-(E) = |\mu|(\mathbf{1}_{E\cap\{h=-1\}})$ and $\mu^+ \perp \mu^-$. Moreover, if $\mu = \mu_1 - \mu_2$ with positive measures μ_1, μ_2, then $\mu_1 \geq \mu^+$ and $\mu_2 \geq \mu^-$. In this sense the Jordan decomposition is minimal.*

Proof The representation of μ^+ and μ^- follows from the previous proposition. Minimality is proved as follows. Since $\mu \leq \mu_1$, we have $\mu^+(E) = \mu(E \cap \{h = 1\}) \leq \mu_1(E \cap \{h = 1\}) \leq \mu_1(E)$. □

Proposition 6.24 *If μ is a positive measure and ν a complex measure such that $\nu \ll \mu$, then $|\nu| \ll \mu$ and*

$$\frac{\mathrm{d}|\nu|}{\mathrm{d}\mu} = \left|\frac{\mathrm{d}\nu}{\mathrm{d}\mu}\right|.$$

Proof Exercise 6.8. □

6.10 Exercises

6.1 Let μ be a real measure on a space (S, Σ). Define $\nu : \Sigma \to [0, \infty)$ by $\nu(E) = \sup\{\mu(F) : F \in \Sigma, F \subset E, \mu(F) \geq 0\}$. Show that ν is a finite positive measure. Give a characterization of ν.

6.2 Prove Proposition 6.6.

6.3 Prove a version of Proposition 6.8 adapted to the case where $h \in \mathcal{L}^1(S, \Sigma, \mu)$ is complex valued.

6.4 Let X be a symmetric Bernoulli distributed random variable ($\mathbb{P}(X = 0) = \mathbb{P}(X = 1) = \frac{1}{2}$) and Y uniformly distributed on $[0, \theta]$ (for some arbitrary $\theta > 0$). Assume that X and Y are independent.

(a) Show that the laws $\mathcal{L}_\theta$ ($\theta > 0$) of XY are not absolutely continuous w.r.t. Lebesgue measure on $\mathbb{R}$.

(b) Find a fixed dominating σ-finite measure μ such that $\mathcal{L}_\theta \ll \mu$ for all θ and determine the corresponding Radon-Nikodym derivatives.

6.5 Let $X_1, X_2, \ldots$ be an *iid* sequence of Bernoulli random variables, defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with $\mathbb{P}(X_1 = 1) = \frac{1}{2}$. Let

$$X = \sum_{k=1}^{\infty} 2^{-k} X_k.$$

(a) Find the distribution of X.

(b) A completely different situation occurs when we ignore the odd numbered random variables. Let

$$Y = 3 \sum_{k=1}^{\infty} 4^{-k} X_{2k},$$

where the factor 3 only appears for esthetic reasons. Show that the distribution function $F : [0, 1] \to \mathbb{R}$ of Y is constant on $(\frac{1}{4}, \frac{3}{4})$, that $F(1 - x) = 1 - F(x)$ and that it satisfies $F(x) = 2F(x/4)$ for $x < \frac{1}{4}$.

(c) Make a sketch of F and show that F is continuous, but not absolutely continuous w.r.t. Lebesgue measure. (Hence there is no Borel measurable function f such that $F(x) = \int_{[0,x]} f(u)\,\mathrm{d}u$, $x \in [0, 1]$).

6.6 Let $f \in \mathcal{L}^1(S, \Sigma, \mu)$ be such that $\mu(\mathbf{1}_E f) = 0$ for all $E \in \Sigma$. Show that $\mu(\{f \neq 0\}) = 0$. Conclude that the function h in the Radon-Nikodym theorem has the stated uniqueness property.

6.7 Let μ and ν be positive σ-finite measures and ϕ an arbitrary measure on a measurable space (S, Σ). Assume that $\phi \ll \nu$ and $\nu \ll \mu$. Show that $\phi \ll \mu$ and that

$$\frac{\mathrm{d}\phi}{\mathrm{d}\mu} = \frac{\mathrm{d}\phi}{\mathrm{d}\nu} \frac{\mathrm{d}\nu}{\mathrm{d}\mu}.$$

6.8 Prove Proposition 6.24.

6.9 Let ν and μ be positive σ-finite measures on (S, Σ) with $\nu \ll \mu$ and let $h = \frac{\mathrm{d}\nu}{\mathrm{d}\mu}$, the standing assumptions in this exercise. Show that $\nu(\{h = 0\}) = 0$. Show that $\mu(\{h = 0\}) = 0$ iff $\mu \ll \nu$. What is $\frac{\mathrm{d}\mu}{\mathrm{d}\nu}$ if this happens?

6.10 Let μ and ν be positive σ-finite measures and ϕ a complex measure on (S, Σ). Assume that $\phi \ll \mu$ and $\nu \ll \mu$ with Radon-Nikodym derivatives h and k respectively. Let $\phi = \phi_a + \phi_s$ be the Lebesgue decomposition of ϕ w.r.t. μ. Show that (ν-a.e.)

$$\frac{\mathrm{d}\phi_a}{\mathrm{d}\nu} = \frac{h}{k}\mathbf{1}_{\{k>0\}}.$$

6.11 Consider the measurable space $(\Omega, \mathcal{F})$ and a measurable map $X : \Omega \to \mathbb{R}^n$ ($\mathbb{R}^n$ is endowed with the usual Borel σ-algebra $\mathcal{B}^n$). Consider two probability measure $\mathbb{P}$ and $\mathbb{Q}$ on $(\Omega, \mathcal{F})$ and let $\mathbb{P}^X$ and $\mathbb{Q}^X$ be the corresponding distributions (laws) on $(\mathbb{R}^n, \mathcal{B}^n)$. Assume that $\mathbb{P}^X$ and $\mathbb{Q}^X$ are both absolutely continuous w.r.t. some σ-finite measure (e.g. Lebesgue measure), with corresponding Radon-Nikodym derivatives (in this context often called densities) f and g respectively, so $f, g : \mathbb{R}^n \to [0, \infty)$. Assume that $g > 0$. Show that for $\mathcal{F} = \sigma(X)$ it holds that $\mathbb{P} \ll \mathbb{Q}$ and that (look at Exercise 6.10) the Radon-Nikodym derivative here can be taken as the *likelihood ratio*

$$\omega \mapsto \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}(\omega) = \frac{f(X(\omega))}{g(X(\omega))}.$$

6.12 Prove Proposition 6.4.

6.13 Show that the collection of sets Σ in the proof of Lemma 6.15 is a σ-algebra.

6.14 Show by a direct argument that $[0, 1] \setminus \mathbb{Q}$ is regular, see Definition 6.14, for the Lebesgue measure on $([0, 1], \mathcal{B}([0, 1]))$.

6.15 Show the assertion on $||T||$ in Remark 6.21.

6.16 Show that the two displayed formulas in Exercise 5.10 are valid for functions F and G that are of bounded variation over some interval $(a, b]$. The integrals should be taken in the Lebesgue-Stieltjes sense.

6.17 Let a random variable X have distribution function F with the decomposition as in Proposition 6.13.

(a) Suppose that $F_s = 0$. Assume that $\mathbb{E}X$ is well defined. How would one compute this expectation practically? See also the introductory paragraph of Chapter 4 for an example where this occurs, and compute for that case $\mathbb{E}X$ explicitly. Verify the answer by exploiting the independence of Y and Z.

(b) As an example of the other extreme case, suppose that F is the distribution of Y as in Exercise 6.5, so $F = F_s$. What is $\mathbb{E}Y$ here?

7 Convergence and Uniform Integrability

In this chapter we first review a number of convergence concepts for random variables and study how they are interrelated. The important concept of uniform integrability shall enable us to perform a more refined analysis.

7.1 Modes of convergence

Let $X, X_1, X_2, \ldots$ be random variables. We have the following definitions of different modes of convergence. We will always assume that the parameter n tends to infinity, unless stated otherwise.

Definition 7.1 Here a three fundamental convergence concepts.

(i) If $\mathbb{P}(\omega : X_n(\omega) \to X(\omega)) = 1$, then we say that X_n converges to X almost surely (a.s.).

(ii) If $\mathbb{P}(|X_n - X| > \varepsilon) \to 0$ for all $\varepsilon > 0$, then we say that X_n converges to X in probability.

(iii) If $\mathbb{E}|X_n - X|^p \to 0$ (equivalently, $||X_n - X||_p \to 0$) for some $p \geq 1$, then we say that X_n converges to X in p-th mean, or in $\mathcal{L}^p$.

For these types of convergence we use the following notations: $X_n \overset{\text{a.s.}}{\to} X$, $X_n \overset{\mathbb{P}}{\to} X$ and $X_n \overset{\mathcal{L}^p}{\to} X$ respectively.

First we study a bit more in detail almost sure convergence of X_n to X. If this type of convergence takes place we have

$$\mathbb{P}(\omega : \forall \varepsilon > 0 : \exists N : \forall n \geq N : |X_n(\omega) - X(\omega)| < \varepsilon) = 1.$$

But then also (dropping the ω in the notation)

$$\text{for all } \varepsilon > 0\text{: } \mathbb{P}(\exists N : \forall n \geq N : |X_n - X| < \varepsilon) = 1. \tag{7.1}$$

Conversely, if (7.1) holds, we have almost sure convergence. Notice that we can rewrite the probability in (7.1) as $\mathbb{P}(\liminf E_n^\varepsilon) = 1$, with $E_n^\varepsilon = \{|X_n - X| < \varepsilon\}$.

Limits are often required to be unique in an appropriate sense. The natural concept of uniqueness here is that of almost sure uniqueness.

Proposition 7.2 *For each of the convergence concepts in Definition 7.1 the limit, when it exists, is almost surely unique. This means that if there are two candidate limits X and X', one must have $\mathbb{P}(X = X') = 1$.*

Proof Suppose that $X_n \overset{\text{a.s.}}{\to} X$ and $X_n \overset{\text{a.s.}}{\to} X'$. Let Ω_0 be the set of probability one on which $X_n(\omega) \to X(\omega)$ and Ω_0' be the set of probability one on which $X_n(\omega) \to X'(\omega)$. Then also $\mathbb{P}(\Omega_0 \cap \Omega_0') = 1$ and by uniqueness of limits of real numbers we must have that $X(\omega) = X'(\omega)$ for all $\omega \in \Omega_0 \cap \Omega_0'$. Hence $\mathbb{P}(X = X') \geq \mathbb{P}(\Omega_0 \cap \Omega_0') = 1$.

If $X_n \xrightarrow{\mathbb{P}} X$ and $X_n \xrightarrow{\mathbb{P}} X'$, then we have by the triangle inequality for any $\varepsilon > 0$

$$\mathbb{P}(|X - X'| > \varepsilon) \leq \mathbb{P}(|X_n - X| > \varepsilon/2) + \mathbb{P}(|X_n - X'| > \varepsilon/2),$$

and the right hand side converges to zero by assumption.

Finally we consider the third convergence concept. We need the basic inequality $|a + b|^p \leq 2^{p-1}(|a|^p + |b|^p)$. This allows us to write $\mathbb{E}|X - X'|^p \leq 2^{p-1}(\mathbb{E}|X_n - X|^p + \mathbb{E}|X_n - X'|^p)$. It follows that $\mathbb{E}|X - X'|^p = 0$ and hence that $\mathbb{P}(X = X') = 1$. □

The following relations hold between the types of convergence introduced in Definition 7.1.

Proposition 7.3 *The following implications hold.*

(i) If $X_n \xrightarrow{\text{a.s.}} X$, then $X_n \xrightarrow{\mathbb{P}} X$.

(ii) If for all $\varepsilon > 0$ the series $\sum_n \mathbb{P}(|X_n - X| > \varepsilon)$ is convergent, then $X_n \xrightarrow{\text{a.s.}} X$.

(iii) If $X_n \xrightarrow{\mathcal{L}^p} X$, then $X_n \xrightarrow{\mathbb{P}} X$.

(iv) If $p > q > 0$ and $X_n \xrightarrow{\mathcal{L}^p} X$, then $X_n \xrightarrow{\mathcal{L}^q} X$.

Proof (i) Assume $X_n \xrightarrow{\text{a.s.}} X$, fix $\varepsilon > 0$ and let $A_n = \{|X_n - X| \geq \varepsilon\}$. From (7.1) we know that $\mathbb{P}(\liminf A_n^c) = 1$, so $\mathbb{P}(\limsup A_n) = 0$. But $A_n \subset U_n := \bigcup_{m \geq n} A_m$ and the U_n form a decreasing sequence with $\limsup A_n$ as its limit. Hence we have $\limsup \mathbb{P}(A_n) \leq \lim \mathbb{P}(U_n) = 0$ and so $X_n \xrightarrow{\mathbb{P}} X$. An easier argument is the following. Let $Y_n = \mathbf{1}_{A_n}$. Then $Y_n \xrightarrow{\text{a.s.}} 0$ for every $\varepsilon > 0$. Dominated convergence gives $\mathbb{P}(A_n) = \mathbb{E}Y_n \to 0$.

(ii) Fix $\varepsilon > 0$ and let $E_n = \{|X_n - X| > \varepsilon\}$. The first part of the Borel-Cantelli lemma (Lemma 3.14) gives that $\mathbb{P}(\limsup E_n) = 0$, equivalently $\mathbb{P}(\liminf E_n^c) = 1$, but this is just (7.1).

(iii) By Markov's inequality we have

$$\mathbb{P}(|X_n - X| > \varepsilon) = \mathbb{P}(|X_n - X|^p > \varepsilon^p) \leq \frac{1}{\varepsilon^p}\mathbb{E}|X_n - X|^p,$$

and the result follows.

(iv) Proposition 4.44 yields $||X_n - X||_q \leq ||X_n - X||_p$. □

The next proposition gives a partial converse to Proposition 7.3 (iii). Notice that the assertion would trivially follow from the Dominated Convergence Theorem for an a.s. converging sequence. The weaker assumption on convergence in probability makes it slightly less trivial. In Theorem 7.15 we will see a kind of converse to Proposition 7.3(iii) for $p = 1$.

Proposition 7.4 *Let (X_n) be a sequence of random variables that is almost surely bounded, there is $K > 0$ such that $\mathbb{P}(|X_n| > K) = 0$, for all n. Let X be a random variable. If $X_n \xrightarrow{\mathbb{P}} X$, then $|X| \leq K$ a.s. and $X_n \xrightarrow{\mathcal{L}^1} X$.*

Proof The first assertion follows from $\mathbb{P}(|X| > K+\varepsilon) \leq \mathbb{P}(|X_n - X| > \varepsilon)$, valid for every $\varepsilon > 0$. Let $n \to \infty$ to conclude $\mathbb{P}(|X| > K + \varepsilon) = 0$, $\forall \varepsilon > 0$. Then $|X_n - X| \leq 2K$ a.s., which we use to prove the second assertion. Consider for any $\varepsilon > 0$

$$\begin{aligned}\mathbb{E}|X_n - X| &\leq \mathbb{E}|X_n - X|\mathbf{1}_{\{|X_n - X| > \varepsilon\}} + \mathbb{E}|X_n - X|\mathbf{1}_{\{|X_n - X| \leq \varepsilon\}} \\ &\leq 2K\mathbb{P}(|X_n - X| > \varepsilon) + \varepsilon.\end{aligned}$$

By the assumed convergence in probability we obtain $\limsup \mathbb{E}|X_n - X| \leq \varepsilon$, true for every $\varepsilon > 0$, from which the assertion follows. □

The following result tells how to use almost sure convergence when convergence in probability has to be established.

Proposition 7.5 *There is equivalence between*

(i) $X_n \xrightarrow{\mathbb{P}} X$ *and*

(ii) every subsequence of (X_n) *contains a further subsequence that is almost surely convergent to* X*.*

Proof Assume that (i) holds, then for any $\varepsilon > 0$ and any subsequence we also have $\mathbb{P}(|X_{n_k} - X| > \varepsilon) \to 0$. Hence for every $p \in \mathbb{N}$, there is $k_p \in \mathbb{N}$ such that $\mathbb{P}(|X_{n_{k_p}} - X| > 2^{-p}) \leq 2^{-p}$. Now we apply part (ii) of Proposition 7.3, which gives us (ii), once we have verified that $\sum_p \mathbb{P}(|X_{n_{k_p}} - X| > \varepsilon) < \infty$ for all $\varepsilon > 0$. This holds since

$$\sum_p \mathbb{P}(|X_{n_{k_p}} - X| > \varepsilon) = \sum_{p:2^{-p} > \varepsilon} \mathbb{P}(|X_{n_{k_p}} - X| > \varepsilon) + \sum_{p:2^{-p} \leq \varepsilon} \mathbb{P}(|X_{n_{k_p}} - X| > \varepsilon),$$

where the first sum on the right hand side has finitely many terms, whereas the second one is less than $\sum_{p:2^{-p} \leq \varepsilon} \mathbb{P}(|X_{n_{k_p}} - X| > 2^{-p})$, which is finite by construction.

Conversely, assume that (ii) holds. We reason by contradiction. Suppose that (i) doesn't hold. Then there exist an $\varepsilon > 0$ and a level $\delta > 0$ such that along some subsequence (n_k) one has

$$\mathbb{P}(|X_{n_k} - X| > \varepsilon) > \delta, \text{ for all } k. \tag{7.2}$$

But the sequence X_{n_k} by assumption has an almost surely convergent subsequence $(X_{n_{k_p}})$, which, by Proposition 7.3 (i), also converges in probability. This contradicts (7.2). □

The following result cannot be a surprise.

Proposition 7.6 *Let* $X, X_1, X_2, \ldots$ *be random variables and* $g : \mathbb{R} \to \mathbb{R}$ *be continuous. If* $X_n \xrightarrow{\text{a.s.}} X$*, we also have* $g(X_n) \xrightarrow{\text{a.s.}} g(X)$ *and if* $X_n \xrightarrow{\mathbb{P}} X$*, then also* $g(X_n) \xrightarrow{\mathbb{P}} g(X)$*.*

Proof Exercise 7.2. □

Convergence, almost surely or in probability, of random vectors is defined similarly. For instance, if $X, X_1, X_2, \ldots$ are n-dimensional random vectors and $||\cdot||$ is a norm on $\mathbb{R}^n$ (you may also take a metric instead of a norm), then we say that $X_n \xrightarrow{\mathbb{P}} X$ if $||X_n - X|| \xrightarrow{\mathbb{P}} 0$. Here we apply the definition of convergence for real random variables to $||X_n - X||$ (which is truly a random variable!). A nice feature of the convergence concepts introduced above is that appropriate convergence results for real valued random variables carry over to results for random vectors.

Proposition 7.7 *Let $X, X_1, X_2, \ldots$ and $Y, Y_1, Y_2, \ldots$ be two sequence of random variables, defined on a common probability space. Put $Z = (X, Y)$ and $Z_n = (X_n, Y_n)$, $n \in \mathbb{N}$. Let $\xrightarrow{*}$ denote any of the three types of convergence $\xrightarrow{\text{a.s.}}$, $\xrightarrow{\mathbb{P}}$ and $\xrightarrow{\mathcal{L}^p}$. If $X_n \xrightarrow{*} X$ and $Y_n \xrightarrow{*} Y$, then $Z_n \xrightarrow{*} Z$.*

Proof Exercise 7.3. □

7.2 Uniform integrability

This section deals with collections of random variables that in some sense *uniformly* belong to $\mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$. First a preparatory result.

Lemma 7.8 *Let $X \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ and put $\nu(F) := \mathbb{E}|X|\mathbf{1}_F$, $F \in \mathcal{F}$. Then for all $\varepsilon > 0$ there exists $\delta > 0$, such that $\nu(F) < \varepsilon$, if $\mathbb{P}(F) < \delta$.*

Proof We reason by contradiction. If the assertion doesn't hold, there exists $\varepsilon > 0$ such that for all $\delta > 0$, there exists a set $F \in \mathcal{F}$ with $\mathbb{P}(F) < \delta$ and $\nu(F) \geq \varepsilon$. And thus, for all $n \in \mathbb{N}$ there are sets $F_n \in \mathcal{F}$ such that $\mathbb{P}(F_n) < 2^{-n}$ and $\nu(F_n) \geq \varepsilon$. Let $F = \limsup F_n$. The Borel-Cantelli Lemma 3.14 states that $\mathbb{P}(F) = 0$. At the same time, we deduce from Fatou's lemma for sets (Exercise 1.4) the contradiction that $\nu(F) \geq \limsup \nu(F_n) \geq \varepsilon$. □

Remark 7.9 The assertion of Lemma 7.8 justifies what we previously called *absolute continuity* (of ν w.r.t. $\mathbb{P}$).

The following corollary characterizes integrability of a random variable. It explains the concept of *uniform* integrability to be introduced in Definition 7.11.

Corollary 7.10 *Let X be a random variable. Then $X \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ iff*

$$\lim_{K \to \infty} \mathbb{E}|X|\mathbf{1}_{\{|X| \geq K\}} = 0.$$

Proof Exercise 7.4. □

Definition 7.11 Let $\mathcal{C}$ be a collection of random variable defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$. This collection is called *uniformly integrable* (UI) if

$$\lim_{K \to \infty} \sup\{\mathbb{E}|X|\mathbf{1}_{\{|X| > K\}} : X \in \mathcal{C}\} = 0.$$

We give some rather general examples of a uniformly integrable collection $\mathcal{C}$.

Example 7.12 Let $\mathcal{C}$ be a collection of random variables that is bounded in $\mathcal{L}^p(\Omega, \mathcal{F}, \mathbb{P})$ for some $p > 1$. By definition, there is $M > 0$, such that $\mathbb{E}|X|^p \leq M, \forall X \in \mathcal{C}$. Then

$$\begin{aligned}\mathbb{E}|X|\mathbf{1}_{\{|X|>K\}} &= \mathbb{E}\frac{|X|^p}{|X|^{p-1}}\mathbf{1}_{\{|X|>K\}} \\ &\leq \mathbb{E}\frac{|X|^p}{K^{p-1}}\mathbf{1}_{\{|X|>K\}} \\ &\leq \frac{M}{K^{p-1}}.\end{aligned}$$

Let $K \to \infty$ to obtain that $\mathcal{C}$ is UI.

Example 7.13 The content of the present example should be intuitively obvious. Let Y be a nonnegative random variable with $EY < \infty$. Let $\mathcal{C}$ be a collection of random variables X with the property $|X| \leq Y$ a.s. Then $\mathcal{C}$ is UI. Indeed, we have from $|X| \leq Y$ a.s. that

$$\mathbb{E}|X|\mathbf{1}_{\{|X|>K\}} \leq \mathbb{E}Y\mathbf{1}_{\{Y>K\}},$$

hence

$$\sup\{\mathbb{E}|X|\mathbf{1}_{\{|X|>K\}} : X \in \mathcal{C}\} \leq \mathbb{E}Y\mathbf{1}_{\{Y>K\}},$$

and the result follows from Corollary 7.10.

Proposition 7.14 *If $\mathcal{C}$ is a finite collection of random variables in $\mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$, then $\mathcal{C}$ is UI. Also the union of two UI collections is UI again.*

Proof Exercise 7.6 □

The importance of the concept of uniform integrability is a consequence of the next theorem, which contains a converse to Proposition 7.3 (iii) for $p = 1$. Recall also Lebesgue's Theorem 4.19 and think of the f_n as random variables X_n. The stipulated conditions there imply that the sequence (X_n) is UI (Example 7.13), and it is indeed this more general property that is needed for $\mathcal{L}^1$-convergence.

Theorem 7.15 *Let (X_n) be a sequence in $\mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ and $X \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$. Then $X_n \xrightarrow{\mathcal{L}^1} X$ (so $\mathbb{E}|X_n - X| \to 0$) iff*

(i) $X_n \xrightarrow{\mathbb{P}} X$ and

(ii) (X_n) is uniformly integrable.

Proof Assume that $\mathbb{E}|X_n - X| \to 0$. Then (i) was already known from Proposition 7.3. To prove (ii) we consider

$$\mathbb{E}|X_n|\mathbf{1}_{\{|X_n|>K\}} \leq \mathbb{E}|X_n - X| + \mathbb{E}|X|\mathbf{1}_{\{|X_n|>K\}},$$

hence for every $N \in \mathbb{N}$

$$\sup_{n\geq N} \mathbb{E}|X_n|\mathbf{1}_{\{|X_n|>K\}} \leq \sup_{n\geq N} \mathbb{E}|X_n - X| + \sup_{n\geq N} \mathbb{E}|X|\mathbf{1}_{\{|X_n|>K\}}. \tag{7.3}$$

Let $\varepsilon > 0$ and choose N such that the first term is smaller than ε, which can be done by $\mathcal{L}^1$-convergence. To control the second term, we observe that $\mathbb{P}(|X_n| > K) \leq \sup_n \mathbb{E}|X_n|/K$. By $\mathcal{L}^1$-convergence one has $\sup_n \mathbb{E}|X_n| < \infty$ and hence, by selecting $K = K(\delta)$ large enough ($K > \sup_n \mathbb{E}|X_n|/\delta$), one has $\mathbb{P}(|X_n| > K) < \delta$, for any $\delta > 0$. Now choose δ as in Lemma 7.8. Then we get $\sup_{n\geq N} \mathbb{E}|X|\mathbf{1}_{\{|X_n|>K\}} < \varepsilon$. It follows from (7.3) that

$$\sup_{n\geq N} \mathbb{E}|X_n|\mathbf{1}_{\{|X_n|>K\}} < 2\varepsilon.$$

Next we consider the X_n for $n < N$. Since by Proposition 7.14 these form a UI collection, we have for all sufficiently large K

$$\sup_{n<N} \mathbb{E}|X|\mathbf{1}_{\{|X_n|>K\}} < 2\varepsilon.$$

Combining the inequalities in the last two displays, we conclude that for all large enough K

$$\sup_{n\geq 1} \mathbb{E}|X_n|\mathbf{1}_{\{|X_n|>K\}} \leq 2\varepsilon,$$

in other words, the family $(X_n)_{n\geq 1}$ is UI.

For the converse implication we proceed as follows. Assume (i). Instead of (ii) we assume for the time being the stronger assumption that (X_n) is a.s. (uniformly) bounded. Then the results follows from Proposition 7.4. If (X_n) is merely UI, we use *truncation functions* ϕ_K defined by

$$\phi_K(x) = x\mathbf{1}_{\{|x|\leq K\}} + \operatorname{sgn}(x)K\mathbf{1}_{\{|x|>K\}}.$$

These functions have the property that $|\phi_K(x) - x| \leq |x|\mathbf{1}_{\{|x|>K\}}$. By the triangle inequality we get

$$\begin{aligned}\mathbb{E}|X_n - X| &\leq \mathbb{E}|X_n - \phi_K(X_n)| + \mathbb{E}|\phi_K(X_n) - \phi_K(X)| + \mathbb{E}|\phi_K(X) - X| \\ &\leq \mathbb{E}|X_n|\mathbf{1}_{\{|X_n|>K\}} + \mathbb{E}|\phi_K(X_n) - \phi_K(X)| + \mathbb{E}|X|\mathbf{1}_{\{|X|>K\}}.\end{aligned}$$

The first term can be made arbitrary small, by selecting K large enough in view of the assumed uniform integrability and likewise we can control the last term in view of Corollary 7.10. The middle term converges to zero in view of the first part of the proof, no matter what K is (verify directly that $\phi_K(X_n) \xrightarrow{\mathbb{P}} \phi_K(X)$, or use Proposition 7.6). This concludes the proof. □

7.3 Exercises

7.1 Let $X_1, Y_1, X_2, Y_2, \ldots$ be an i.i.d. sequence whose members have a uniform distribution on $[0, 1]$ and let $f : [0, 1] \to [0, 1]$ be continuous. Define $Z_i = \mathbf{1}_{\{f(X_i)>Y_i\}}$.

(a) Show that $\frac{1}{n}\sum_{i=1}^{n} Z_i \to \int_0^1 f(x)\,dx$ a.s.
(b) Show that $\mathbb{E}(\frac{1}{n}\sum_{i=1}^{n} Z_i - \int_0^1 f(x)\,dx)^2 \leq \frac{1}{4n}$.
(c) Explain why these two results are useful.

7.2 Prove Proposition 7.6.

7.3 Prove Proposition 7.7.

7.4 Prove Corollary 7.10.

7.5 Let $\mathcal{C}$ be a collection of uniformly integrable random variables. Show that $\mathcal{C}$ is *bounded in* $\mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$, i.e. $\sup\{\mathbb{E}|X| : X \in \mathcal{C}\} < \infty$. Let $(\Omega, \mathcal{F}, \mathbb{P}) = ([0,1], \mathcal{B}[0,1], \nu)$ and $X_n(\omega) = n\mathbf{1}_{(0,1/n)}(\omega)$. Show that $\{X_n : n \in \mathbb{N}\}$ is bounded in $\mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$, but not uniformly integrable.

7.6 Prove Proposition 7.14.

7.7 Let $\mathcal{C}_1, \ldots, \mathcal{C}_n$ be uniformly integrable collections of random variables on a common probability space. Show that $\bigcup_{k=1}^{n} \mathcal{C}_k$ is uniformly integrable. (In particular is a finite collection in $\mathcal{L}^1$ uniformly integrable).

7.8 If $\mathcal{C}_1$ and $\mathcal{C}_2$ are uniformly integrable collections in $\mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$, so is $\mathcal{C}_1 + \mathcal{C}_2 := \{X_1 + X_2 : X_1 \in \mathcal{C}_1, X_2 \in \mathcal{C}_2\}$. Show this.

7.9 Here is *uniform variation* on Lemma 7.8. Let $\mathcal{C}$ be a class of random variables defined on some $(\Omega, \mathcal{F}, \mathbb{P})$. Put $\nu_X(F) := \mathbb{E}|X|\mathbf{1}_F$, $X \in \mathcal{C}$, $F \in \mathcal{F}$. Show that $\mathcal{C}$ is uniformly integrable iff the following two conditions hold.
(i) $\mathcal{C}$ is bounded in $\mathcal{L}^1$.
(ii) For all $\varepsilon > 0$ there exists $\delta > 0$, such that $\nu_X(F) < \varepsilon$ $\forall X \in \mathcal{C}$, if $\mathbb{P}(F) < \delta$.

7.10 Let $\mathcal{C}$ be a uniformly integrable collection of random variables.
(a) Consider $\bar{\mathcal{C}}$, the closure of $\mathcal{C}$ in $\mathcal{L}^1$. Use Exercise 7.9 to show that also $\bar{\mathcal{C}}$ is uniformly integrable.
(b) Let $\mathcal{D}$ be the convex hull of $\mathcal{C}$, the smallest convex set that contains $\mathcal{C}$. Then both $\mathcal{D}$ and its closure in $\mathcal{L}^1$ are uniformly integrable

7.11 In this exercise you prove (fill in the details) the following characterization: a collection $\mathcal{C}$ is uniformly integrable iff there exists a function $G : \mathbb{R}^+ \to \mathbb{R}^+$ such that $\lim_{t\to\infty} \frac{G(t)}{t} = \infty$ and $M := \sup\{\mathbb{E}\,G(|X|) : X \in \mathcal{C}\} < \infty$.

The necessity you prove as follows. Let $\varepsilon > 0$ choose $a = M/\varepsilon$ and c such that $\frac{G(t)}{t} \geq a$ for all $t > c$. To prove uniform integrability of $\mathcal{C}$ you use that $|X| \leq \frac{G(|X|)}{a}$ on the set $\{|X| \geq c\}$.

It is less easy to prove sufficiency. Proceed as follows. Suppose that we have a sequence (g_n) with $g_0 = 0$ and $\lim_{n\to\infty} g_n = \infty$. Define $g(t) = \sum_{n=0}^{\infty} \mathbf{1}_{[n,n+1)}(t) g_n$ and $G(t) = \int_0^t g(s)\,\mathrm{d}s$. Check that $\lim_{t\to\infty} \frac{G(t)}{t} = \infty$.

With $a_n(X) = \mathbb{P}(|X| > n)$, it holds that $\mathbb{E}G(|X|) \leq \sum_{n=1}^{\infty} g_n a_n(|X|)$. Furthermore, for every $k \in \mathbb{N}$ we have $\int_{|X|\geq k} |X|\,\mathrm{d}\mathbb{P} \geq \sum_{m=k}^{\infty} a_m(X)$. Pick for every n

a constant $c_n \in \mathbb{N}$ such that $\int_{|X|\geq c_n} |X| \, \mathrm{d}\mathbb{P} \leq 2^{-n}$. Then $\sum_{m=c_n}^{\infty} a_m(X) \leq 2^{-n}$ and hence $\sum_{n=1}^{\infty} \sum_{m=c_n}^{\infty} a_m(X) \leq 1$. Choose then the sequence (g_n) as the 'inverse' of (c_n): $g_n = \#\{k : c_k \leq n\}$.

7.12 Prove that a collection $\mathcal{C}$ is uniformly integrable iff there exists an *increasing and convex* function $G : \mathbb{R}^+ \to \mathbb{R}^+$ such that $\lim_{t\to\infty} \frac{G(t)}{t} = \infty$ and $M := \sup\{\mathbb{E}G(|X|) : X \in \mathcal{C}\} < \infty$. (You may use the result of Exercise 7.11.) Let $\mathcal{D}$ be the closure of the convex hull of a uniformly integrable collection $\mathcal{C}$ in $\mathcal{L}^1$. With the function G as above we have $\sup\{\mathbb{E}G(|X|) : X \in \mathcal{D}\} = M$, whence also $\mathcal{D}$ is uniformly integrable.

7.13 Let $p \geq 1$ and let $X, X_1, X_2, \ldots$ be random variables. Then X_n converges to X in $\mathcal{L}^p$ iff the following two conditions are satisfied.

(i) $X_n \to X$ in probability,

(ii) The collection $\{|X_n|^p : n \in \mathbb{N}\}$ is uniformly integrable.

7.14 Here is an extension of Proposition 10.9, but you can do the exercise now. Let $\mathcal{C}$ be a uniformly integrable collection of random variables on some $(\Omega, \mathcal{F}, \mathbb{P})$. Let $\mathbb{G}$ be a family of sub-σ-algebras of $\mathcal{F}$. Let $\mathcal{D} = \{\mathbb{E}[X|\mathcal{G}] : X \in \mathcal{C}, \mathcal{G} \in \mathbb{G}\}$ ('in the sense of versions'). Show that also $\mathcal{D}$ is uniformly integrable. (*Hint:* use Exercise 7.9.)

7.15 Let $X, X_1, X_2, \ldots$ be random variables that are defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Suppose that $X_n \xrightarrow{\mathbb{P}} X$ and that there exists a random variable Y with $\mathbb{E}Y < \infty$ such that $|X_n| \leq Y$ a.s. Show that $\mathbb{P}(|X| \leq Y) = 1$ and that $X_n \xrightarrow{\mathcal{L}^1} X$.

7.16 Assume that $X_1, X_2, \ldots$ is an *iid* sequence and that $\overline{X}_n \xrightarrow{\text{a.s.}} \mu$ for some $\mu \in \mathbb{R}$. Show that $\mathbb{E}|X_1| < \infty$ and that $\mathbb{E}X_1 = \mu$. *Hint:* Show first that $\mathbb{E}|X_1| < \infty$ iff $\sum_{n=0}^{\infty} \mathbb{P}(|X_n| > n) < \infty$ and use Borel-Cantelli. (This exercise is a converse to the strong law of large numbers, Theorem 10.23.)

7.17 Here is a generalisation of Fatou's lemma. Let (X_n) be a sequence of random variables such that the negative parts (X_n^-) are uniformly integrable. Show that $\liminf_{n\to\infty} \mathbb{E}X_n \geq \mathbb{E} \liminf_{n\to\infty} X_n$. *Hint: show first that for given $\varepsilon > 0$ there exists $a < 0$ such that $\mathbb{E}(X_n \vee -a) \leq \mathbb{E}X_n + \varepsilon$ for all n.*

8 Conditional expectation

8.1 A simple, finite case

Let X be a random variable with values in $\{x_1, \ldots, x_n\}$ and Y a random variable with values in $\{y_1, \ldots, y_m\}$. The conditional probability

$$\mathbb{P}(X = x_i | Y = y_j) := \frac{\mathbb{P}(X = x_i, Y = y_j)}{\mathbb{P}(Y = y_j)}$$

is well defined if $\mathbb{P}(Y = y_j) > 0$. Otherwise we define it to be zero. We write E_j for $\{Y = y_j\}$. The conditional expectation $\hat{x}_j := \mathbb{E}[X|E_j]$ is then

$$\hat{x}_j = \sum_i x_i \mathbb{P}(X = x_i | E_j).$$

We define now a new *random variable* $\hat{X}$ by

$$\hat{X} = \sum_j \hat{x}_j \mathbf{1}_{E_j}.$$

Since $\hat{X} = \hat{x}_j$ on each event $\{Y = y_j\}$, we call $\hat{X}$ the conditional expectation of X given Y. It has two remarkable properties. First we see that $\hat{X}$ is $\sigma(Y)$-measurable. The second property, which we prove below, is

$$\mathbb{E}\hat{X}\mathbf{1}_{E_j} = \mathbb{E}X\mathbf{1}_{E_j},$$

the expectation of $\hat{X}$ over the set E_j is the same as the expectation of X over that set. We show this by simple computation. Note first that the values of $X\mathbf{1}_{E_j}$ are zero and x_i, the latter reached on the event $\{X = x_i\} \cap E_j$ that has probability $\mathbb{P}(\{X = x_i\} \cap E_j)$. Note too that $\hat{X}\mathbf{1}_{E_j} = \hat{x}_j \mathbf{1}_{E_j}$. We then get

$$\begin{aligned}
\mathbb{E}\hat{X}\mathbf{1}_{E_j} &= \hat{x}_j \mathbb{P}(E_j) \\
&= \sum_i x_i \mathbb{P}(\{X = x_i\}|E_j)\mathbb{P}(E_j) \\
&= \sum_i x_i \mathbb{P}(\{X = x_i\} \cap E_j) \\
&= \mathbb{E}X\mathbf{1}_{E_j}.
\end{aligned}$$

Every event $E \in \sigma(Y)$ is a finite union of events E_j. It then follows that

$$\mathbb{E}\hat{X}\mathbf{1}_E = \mathbb{E}X\mathbf{1}_E, \forall E \in \sigma(Y). \tag{8.1}$$

The just described two properties of the conditional expectation will lie at the heart of a more general concept, *conditional expectation* of a random variable given a σ-algebra, see Section 8.2.

The random variable $\hat{X}$ is a.s. the only $\sigma(Y)$-measurable random variable that

satisfies (8.1). Indeed, suppose that Z is $\sigma(Y)$-measurable and that $\mathbb{E}Z\mathbf{1}_E = \mathbb{E}X\mathbf{1}_E, \forall E \in \sigma(Y)$. Let $E = \{Z > \hat{X}\}$. Then $(Z - \hat{X})\mathbf{1}_E \geq 0$ and has expectation zero since $E \in \sigma(Y)$, so we have $(Z - \hat{X})\mathbf{1}_{\{Z>\hat{X}\}} = 0$ a.s. Likewise we get $(Z - \hat{X})\mathbf{1}_{\{Z<\hat{X}\}} = 0$ a.s. and it then follows that $Z - \hat{X} = 0$ a.s.

8.2 Conditional expectation for $X \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $\mathcal{G}$ a sub-σ-algebra of $\mathcal{F}$. Assume that $X \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$. Inspired by the results of the previous section we adopt the following definition.

Definition 8.1 A random variable $\hat{X}$ is called a version of the conditional expectation $\mathbb{E}[X|\mathcal{G}]$, if it is $\mathcal{G}$-measurable and if

$$\mathbb{E}\hat{X}\mathbf{1}_G = \mathbb{E}X\mathbf{1}_G, \forall G \in \mathcal{G}. \tag{8.2}$$

If $\mathcal{G} = \sigma(Y)$, where Y is a random variable, then we usually write $\mathbb{E}[X|Y]$ instead of $\mathbb{E}[X|\sigma(Y)]$.

Theorem 8.2 *If $X \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$, then a version of the conditional expectation $\mathbb{E}[X|\mathcal{G}]$ exists and moreover, any two versions are a.s. equal.*

Proof For any $G \in \mathcal{G}$ we define $\nu^+(G) := \mathbb{E}X^+\mathbf{1}_G$ and $\nu^-(G) := \mathbb{E}X^-\mathbf{1}_G$ We have seen that ν^+ and ν^- are finite measures on the measurable space $(\Omega, \mathcal{G})$. Moreover, $\nu^+ \ll \mathbb{P}$ and $\nu^- \ll \mathbb{P}$ on this space. According to the Radon-Nikodym Theorem 6.10 there exist nonnegative $\mathcal{G}$-measurable functions ξ^+ and ξ^- such that $\nu^+(G) = \mathbb{E}\xi^+\mathbf{1}_G$ and $\nu^-(G) = \mathbb{E}\xi^-\mathbf{1}_G$. These functions are a.s. unique. Then $\hat{X} = \xi^+ - \xi^-$ is a version of $\mathbb{E}[X|\mathcal{G}]$. □

Remark 8.3 The ξ^+ and ξ^- in the above proof are in general not equal to the positive and negative parts $\hat{X}^+$ and $\hat{X}^-$ of $\hat{X}$. Think of a simple example.

Remark 8.4 It is common to call a given version of $\mathbb{E}[X|\mathcal{G}]$ *the* conditional expectation of X given $\mathcal{G}$, but one should take care with this custom. In fact one should consider $\mathbb{E}[X|\mathcal{G}]$ as an equivalence class of random variables, where equivalence $Y_1 \sim Y_2$ for $\mathcal{G}$-measurable functions means that $\mathbb{P}(Y_1 = Y_2) = 1$. As such one can consider $\mathbb{E}[X|\mathcal{G}]$ as an element of $L^1(\Omega, \mathcal{G}, \mathbb{P})$. Later on we will often identify a version $\hat{X}$ of $\mathbb{E}[X|\mathcal{G}]$ with $\mathbb{E}[X|\mathcal{G}]$.

Remark 8.5 One can also define versions of conditional expectations for random variables X with $\mathbb{P}(X \in [0, \infty]) = 1$ without requiring that $\mathbb{E}X < \infty$. Again this follows from the Radon-Nikodym theorem. The definition of conditional expectation can also be extended to e.g. the case where $X = X^+ - X^-$, where $\mathbb{E}X^- < \infty$, but $\mathbb{E}X^+ = \infty$.

Let us present the most relevant properties of conditional expectation. As before, we let $X \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$, $\mathcal{G}$ a sub-σ-algebra of $\mathcal{F}$ and $\hat{X}$ is a version of $\mathbb{E}[X|\mathcal{G}]$. Other random variables below that are versions of a conditional expectation given $\mathcal{G}$ are similarly denoted with a 'hat'.

Proposition 8.6 *The following elementary properties hold.*

(i) If $X \geq 0$ a.s., then $\hat{X} \geq 0$ a.s. If $X \geq Y$ a.s., then $\hat{X} \geq \hat{Y}$ a.s.

(ii) $\mathbb{E}\hat{X} = \mathbb{E}X$.

(iii) If $a, b \in \mathbb{R}$ and if $\hat{X}$ and $\hat{Y}$ are versions of $\mathbb{E}[X|\mathcal{G}]$ and $\mathbb{E}[Y|\mathcal{G}]$, then $a\hat{X}+b\hat{Y}$ is a version of $\mathbb{E}[aX + bY|\mathcal{G}]$.

(iv) If X is $\mathcal{G}$-measurable, then X is a version of $\mathbb{E}[X|\mathcal{G}]$.

Proof (i) Let $G = \{\hat{X} < 0\}$. Then we have from (8.2) that $0 \geq \mathbb{E}\mathbf{1}_G\hat{X} = \mathbb{E}\mathbf{1}_G X \geq 0$. Hence $\mathbf{1}_G\hat{X} = 0$ a.s.

(ii) Take $G = \Omega$ in (8.2).

(iii) Just verify that $\mathbb{E}\mathbf{1}_G(a\hat{X} + b\hat{Y}) = \mathbb{E}\mathbf{1}_G(aX + bY)$, for all $G \in \mathcal{G}$.

(iv) Obvious. □

We have taken some care in formulating the assertions of the previous theorem concerning versions. Bearing this in mind and being a bit less precise at the same time, one often phrases e.g. (iii) as $\mathbb{E}[aX + bY|\mathcal{G}] = a\mathbb{E}[X|\mathcal{G}] + b\mathbb{E}[Y|\mathcal{G}]$. Some convergence properties are listed in the following theorem.

Theorem 8.7 *The following convergence properties for conditional expectation given a fixed sub-σ-algebra hold.*

(i) If (X_n) is an a.s. increasing sequence of nonnegative random variables, then the same holds for versions $(\hat{X}_n)$. If moreover $X_n \uparrow X$ a.s., then $\hat{X}_n \uparrow \hat{X}$ a.s. (monotone convergence for conditional expectations)

(ii) If (X_n) is a sequence of a.s. nonnegative random variables, and $(\hat{X}_n)$ are corresponding versions of the conditional expectations, then $\liminf_{n\to\infty} \hat{X}_n \geq \hat{\ell}$ a.s., where $\hat{\ell}$ is a version of the conditional expectation of $\ell := \liminf_{n\to\infty} X_n$. (Fatou's lemma for conditional expectations)

(iii) If (X_n) is a sequence of random variables such that for some X one has $X_n \to X$ a.s. and if there is a random variable Y such that $\mathbb{E}Y < \infty$ and $|X_n| \leq Y$ a.s. for all n. Then $\hat{X}_n \to \hat{X}$ a.s. (dominated convergence for conditional expectations)

Proof (i) From the previous theorem we know that the $\hat{X}_n$ form a.s. an increasing sequence. Let $\hat{X} := \limsup \hat{X}_n$, then $\hat{X}$ is $\mathcal{G}$-measurable and $\hat{X}_n \uparrow \hat{X}$ a.s. We verify that this $\hat{X}$ is a version of $\mathbb{E}[X|\mathcal{G}]$. But this follows by application of the Monotone Convergence Theorem to both sides of $\mathbb{E}\mathbf{1}_G X_n = \mathbb{E}\mathbf{1}_G\hat{X}_n$ for all $G \in \mathcal{G}$.

(ii) and (iii) These properties follow by mimicking the proofs of the ordinary versions of Fatou's Lemma and the Dominated Convergence Theorem, Exercises 8.4 and 8.5. □

Theorem 8.8 *Additional properties of conditional expectations are as follows.*

(i) If $\mathcal{H}$ is a sub-σ-algebra of $\mathcal{G}$, then any version of $\mathbb{E}[\hat{X}|\mathcal{H}]$ is also a version of $\mathbb{E}[X|\mathcal{H}]$ and vice versa (tower property).

(ii) *If Z is $\mathcal{G}$-measurable such that $ZX \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$, then $Z\hat{X}$ is a version of $\mathbb{E}[ZX|\mathcal{G}]$. We write $Z\mathbb{E}[X|\mathcal{G}] = \mathbb{E}[ZX|\mathcal{G}]$.*

(iii) *Let $\hat{X}$ be a version of $\mathbb{E}[X|\mathcal{G}]$. If $\mathcal{H}$ is independent of $\sigma(X) \vee \mathcal{G}$, then $\hat{X}$ is a version of $\mathbb{E}[X|\mathcal{G} \vee \mathcal{H}]$. In particular, $\mathbb{E}X$ is a version of $\mathbb{E}[X|\mathcal{H}]$ if $\sigma(X)$ and $\mathcal{H}$ are independent.*

(iv) *Let X be a $\mathcal{G}$-measurable random variable and let the random variable Y be independent of $\mathcal{G}$. Assume that $h \in \mathcal{B}(\mathbb{R}^2)$ is such that $h(X, Y) \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$. Put $\hat{h}(x) = \mathbb{E}[h(x, Y)]$. Then $\hat{h}$ is a Borel function and $\hat{h}(X)$ is a version of $\mathbb{E}[h(X, Y)|\mathcal{G}]$.*

(v) *If $c : \mathbb{R} \to \mathbb{R}$ is a convex function and $\mathbb{E}|c(X)| < \infty$, then $c(\hat{X}) \leq C$, a.s., where C is any version of $\mathbb{E}[c(X)|\mathcal{G}]$. We often write $c(\mathbb{E}[X|\mathcal{G}]) \leq \mathbb{E}[c(X)|\mathcal{G}]$ (Jensen's inequality for conditional expectations).*

(vi) *$||\hat{X}||_p \leq ||X||_p$, for every $p \geq 1$.*

Proof (i) Let $\tilde{X}$ be a version of $\mathbb{E}[\hat{X}|\mathcal{H}]$. By definition, we have $\mathbb{E}\mathbf{1}_H\tilde{X} = \mathbb{E}\mathbf{1}_H\hat{X}$, for all $H \in \mathcal{H}$. But since $\mathcal{H} \subset \mathcal{G}$, it also holds that $\mathbb{E}\mathbf{1}_H\hat{X} = \mathbb{E}\mathbf{1}_H X$, by (8.2). Hence $\tilde{X}$ is a version of $\mathbb{E}[X|\mathcal{H}]$.

(ii) We give the proof for bounded Z. Certainly ZX is integrable in this case and its conditional expectation exists. Without loss of generality we may then even assume that $Z \geq 0$ a.s. (Add a constant c to Z to have $Z + c \geq 0$, if this is not the case and the result will follow from the case of nonnegative Z). Assume first that also X is nonnegative. If $Z = \mathbf{1}_G$ for some $G \in \mathcal{G}$, then the result directly follows from the definition. By linearity the assertion holds for nonnegative simple Z. For arbitrary $Z \geq 0$, we choose simple Z_n such that $Z_n \uparrow Z$. Since we know (in the sense of versions) $Z_n\hat{X} = \mathbb{E}[Z_n X|\mathcal{G}]$, we apply Theorem 8.7 (i)) to settle the case for $X \geq 0$. If X is arbitrary, linearity yields the assertion by applying the previous results for X^+ and X^-.

(iii) It is sufficient to show this for nonnegative X. Let $G \in \mathcal{G}$ and $H \in \mathcal{H}$. By the independence assumption, we have $\mathbb{E}\mathbf{1}_G\mathbf{1}_H X = \mathbb{E}\mathbf{1}_G X\,\mathbb{P}(H)$ and $\mathbb{E}\mathbf{1}_G\mathbf{1}_H\hat{X} = \mathbb{E}\mathbf{1}_G\hat{X}\,\mathbb{P}(H)$. It follows that $\mathbb{E}\mathbf{1}_G\mathbf{1}_H X = \mathbb{E}\mathbf{1}_G\mathbf{1}_H\hat{X}$, since $\hat{X}$ is version of $\mathbb{E}[X|\mathcal{G}]$. Recall from Exercise 1.6 that the collection $\mathcal{C} := \{G \cap H : G \in \mathcal{G}, H \in \mathcal{H}\}$ is a π-system that generates $\mathcal{G} \vee \mathcal{H}$. Observe that $E \mapsto \mathbb{E}\mathbf{1}_E X$ and $E \mapsto \mathbb{E}\mathbf{1}_E\hat{X}$ both define measures on $\mathcal{G} \vee \mathcal{H}$ and that these measures have been seen to coincide on $\mathcal{C}$. It follows from Theorem 1.15 that these measures are the same. The second statement follows by taking $\mathcal{G} = \{\emptyset, \Omega\}$.

(iv) We use the Monotone Class Theorem, Theorem 3.6 and for simplicity of notation we take X and Y real valued. Let V be the collection of all bounded measurable functions for which the statement holds true. Using (iii), one easily checks that $h = \mathbf{1}_{B\times C} \in V$, where B, C are Borel sets in $\mathbb{R}$. The sets $B \times C$ form a π-system that generates $\mathcal{B}(\mathbb{R}^2)$. The collection V is obviously a vector space and the constant functions belong to it. Let (h_n) be an increasing sequence of nonnegative functions in V that converge to some bounded function h. If $\hat{h}_n(x) = \mathbb{E}h_n(x, Y)$ and $\hat{h}(x) = \mathbb{E}h(x, Y)$, then we also have $\hat{h}_n(x) \uparrow \hat{h}(x)$ for all x by the Monotone Convergence Theorem. We will see that $\hat{h}(X)$ is a version of $\mathbb{E}[h(X, Y)|\mathcal{G}]$. Let $G \in \mathcal{G}$. For all n it holds that $\mathbb{E}\mathbf{1}_G\hat{h}_n(X) =$

$\mathbb{E}\mathbf{1}_G h_n(X,Y)$. Invoking the Monotone Convergence Theorem again results in $\mathbb{E}\mathbf{1}_G \hat{h}(X) = \mathbb{E}\mathbf{1}_G h(X,Y)$. Since all $\hat{h}_n(X)$ are $\mathcal{G}$-measurable, the same holds for $\hat{h}(X)$ and we conclude that $h \in V$. The remainder of the proof is Exercise 8.12.

(v) Since c is convex, there are sequences (a_n) and (b_n) in $\mathbb{R}$ such that $c(x) = \sup\{a_n x + b_n : n \in \mathbb{N}\}$, $\forall x \in \mathbb{R}$. Hence for all n we have $c(X) \geq a_n X + b_n$ and by the monotonicity property of conditional expectation, we also have $C \geq a_n \hat{X} + b_n$ a.s. If N_n is the set of probability zero, where this inequality is violated, then also $\mathbb{P}(N) = 0$, where $N = \cup_{n=1}^{\infty} N_n$. Outside N we have $C \geq \sup_n(a_n \hat{X} + b_n) = c(\hat{X})$.

(vi) The statement concerning the p-norms follows upon choosing $c(x) = |x|^p$ in (v) and taking expectations. □

Here is an illustrative example.

Example 8.9 Let $X_1, X_2, \ldots$ be an *iid* sequence of integrable random variables. Put $S_n = \sum_{i=1}^n X_i$, $n \in \mathbb{N}$. We claim that for every n the following reasonable identity (to be interpreted in the sense of versions)

$$\mathbb{E}[X_1|S_n] = \frac{S_n}{n} \text{ a.s.}$$

holds true. Argue as follows. By symmetry we have for all sets $G = \{S_n \in B\}$ the equality $\mathbb{E}\mathbf{1}_G X_1 = \mathbb{E}\mathbf{1}_G X_j$, for every $j \in \{1, \ldots, n\}$. Hence we obtain $\mathbb{E}[X_1|S_n] = \mathbb{E}[X_j|S_n]$ and then $S_n = \mathbb{E}[S_n|S_n] = \sum_{j=1}^n \mathbb{E}[X_j|S_n] = n\mathbb{E}[X_1|S_n]$. Even more is true, $\mathbb{E}[X_1|S_n]$ also equals $\mathbb{E}[X_1|\mathcal{G}_n]$, where $\mathcal{G}_n = \sigma(S_n, S_{n+1}, \ldots)$. To see this, observe first that $\mathcal{G}_n = \sigma(S_n) \vee \mathcal{H}_n$, where $\mathcal{H}_n = \sigma(X_{n+1}, X_{n+2}, \ldots)$ and that $\mathcal{H}_n$ is independent of $\sigma(X_1, S_n)$). Now apply Theorem 8.8(iii).

Let $P : L^1(\Omega, \mathcal{F}, \mathbb{P}) \to L^1(\Omega, \mathcal{G}, \mathbb{P})$ be the linear map that transforms X into $\mathbb{E}[X|\mathcal{G}]$. If $\hat{X}$ is a version of $\mathbb{E}[X|\mathcal{G}]$, then it is also a version of $\mathbb{E}[\hat{X}|\mathcal{G}]$. So, we get $P^2 = P$, meaning that P is a *projection*. In the next proposition we give this a geometric interpretation in a slightly narrower context.

Proposition 8.10 *Let $X \in \mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$ and $\mathcal{G}$ a sub-σ-algebra of $\mathcal{F}$. If $\hat{X}$ is a version of $\mathbb{E}[X|\mathcal{G}]$, then $\hat{X} \in \mathcal{L}^2(\Omega, \mathcal{G}, \mathbb{P})$ and*

$$\mathbb{E}(X - Y)^2 = \mathbb{E}(X - \hat{X})^2 + \mathbb{E}(\hat{X} - Y)^2, \, \forall Y \in \mathcal{L}^2(\Omega, \mathcal{G}, \mathbb{P}).$$

Hence, $\mathbb{E}(X - Y)^2 \geq \mathbb{E}(X - \hat{X})^2$, $\forall Y \in \mathcal{L}^2(\Omega, \mathcal{G}, \mathbb{P})$. Conditional expectations of square integrable random variables are thus orthogonal projections onto $\mathcal{L}^2(\Omega, \mathcal{G}, \mathbb{P})$, in view of Corollary 4.52.

Proof Exercise 8.3. □

We conclude this section with the following loose statement, whose message should be clear from the above results. A conditional expectation is a *random variable* that has properties similar to those of ordinary expectation.

8.3 Conditional probabilities

Let $F \in \mathcal{F}$ and $\mathcal{G}$ a sub-σ-algebra of $\mathcal{F}$. We define $\mathbb{P}(F|\mathcal{G}) := \mathbb{E}[\mathbf{1}_F|\mathcal{G}]$, the conditional probability of F given $\mathcal{G}$. So a version of $\mathbb{P}(F|\mathcal{G})$ is a $\mathcal{G}$-measurable random variable $\hat{\mathbb{P}}(F)$ that satisfies

$$\mathbb{P}(F \cap G) = \mathbb{E}[\hat{\mathbb{P}}(F)\mathbf{1}_G], \, \forall G \in \mathcal{G}.$$

Likewise, one can define conditional distributions of a random variable X. For a Borel set B one defines $\mathbb{P}^X(B|\mathcal{G}) := \mathbb{P}(X^{-1}[B]|\mathcal{G})$.

Of course all versions of $\mathbb{P}(F|\mathcal{G})$ are almost surely equal. Moreover, if $F_1, F_2, \ldots$ is a sequence of disjoint events, and $\hat{\mathbb{P}}(F_n)$ are versions of the conditional probabilities, then one easily shows that $\sum_{n=1}^{\infty} \hat{\mathbb{P}}(F_n)$ is a version of the conditional probability $\mathbb{P}(\cup_{n=1}^{\infty} F_n|\mathcal{G})$. So, if $\hat{\mathbb{P}}(\cup_{n=1}^{\infty} F_n)$ is any version of $\mathbb{P}(\cup_{n=1}^{\infty} F_n|\mathcal{G})$, then outside a set N of probability zero, we have

$$\hat{\mathbb{P}}(\cup_{n=1}^{\infty} F_n) = \sum_{n=1}^{\infty} \hat{\mathbb{P}}(F_n). \tag{8.3}$$

A problem is that the set N in general depends on the sequence of events $F_1, F_2, \ldots$ Since there are usually uncountably many of such sequences, it is not clear (and in fact not always true!, see Exercise 8.9) that there is one (fixed) set of probability zero such that outside this set for *all* disjoint sequences (F_n) the equality (8.3) holds true. But if it does, this means that for every $F \in \mathcal{F}$, there exists a random variable $\hat{\mathbb{P}}(F)$ that is a version of $\mathbb{P}(F|\mathcal{G})$ and such that for all ω outside a set N with $\mathbb{P}(N) = 0$ the map $F \mapsto \hat{\mathbb{P}}(F)(\omega)$ is a probability measure on $\mathcal{F}$. In this case, the map

$$\mathcal{F} \times \Omega \ni (F, \omega) \mapsto \hat{\mathbb{P}}(F)(\omega)$$

is called a *regular conditional probability* given $\mathcal{G}$.

In the above setup for regular conditional probabilities, relation (8.3) is assumed to hold outside a set N of probability zero. Of course, if $N = \emptyset$, this relation holds everywhere. But also of $N \neq \emptyset$, this relation can be turned into one that is everywhere true. Suppose that $N \neq \emptyset$. Redefine $\hat{\mathbb{P}}$ by taking $\hat{\mathbb{P}}(F)(\omega)$ as given on N^c, but for all $\omega \in N$ we take instead $\hat{\mathbb{P}}(\cdot)(\omega)$ as any fixed probability measure on $\mathcal{F}$ (for instance a Dirac measure). Since we change the map $\hat{\mathbb{P}}(F)$ on the null set N only, we keep on having a conditional probability of F, whereas (8.3) now holds everywhere. One easily checks that the modification $\hat{\mathbb{P}}(\cdot)(\cdot)$ enjoys the following properties. For any fixed ω, $\hat{\mathbb{P}}(\cdot)(\omega)$ is a probability measure, whereas for any fixed $F \in \mathcal{F}$, $\hat{\mathbb{P}}(F)(\cdot)$ is a $\mathcal{G}$-measurable function. These two properties are often cast by saying that $(F, \omega) \mapsto \hat{\mathbb{P}}(F)(\omega)$ is a *probability kernel* defined on $\mathcal{F} \times \Omega$.

As mentioned before, regular conditional probabilities do not always exist. But when it happens to be the case, conditional expectations can be computed through integrals.

Theorem 8.11 *Let X be a (real) random variable with law $\mathbb{P}^X$, a probability measure on $(\mathbb{R}, \mathcal{B})$. There exists a* regular conditional distribution *of X given $\mathcal{G}$. That is, there exists a probability kernel $\hat{\mathbb{P}}^X$ on $\mathcal{B} \times \Omega$ with the property that $\hat{\mathbb{P}}^X(B)$ is a version of $\mathbb{P}(X^{-1}[B]|\mathcal{G})$.*

Proof We split the proof into two parts. First we show the existence of a conditional distribution function, after which we show that it generates a regular conditional distribution of X given $\mathcal{G}$.

We will construct a conditional distribution function on the rational numbers. For each $q \in \mathbb{Q}$ we select a version of $\mathbb{P}(X \leq q|\mathcal{G})$, call it $G(q)$. Let $E_{rq} = \{G(r) < G(q)\}$. Assume that $r > q$. Then $\{X \leq r\} \supset \{X \leq q\}$ and hence $G(r) \geq G(q)$ a.s. and so $\mathbb{P}(E_{rq}) = 0$. Hence we obtain that $\mathbb{P}(E) = 0$, where $E = \cup_{r>q} E_{rq}$. Note that E is the set where the random variables $G(q)$ fail to be increasing in the argument q. Let $F_q = \{\inf_{r>q} G(r) > G(q)\}$. Let $\{q_1, q_2, \ldots\}$ be the set of rationals strictly bigger then q and let $r_n = \inf\{q_1, \ldots, q_n\}$. Then $r_n \downarrow q$, as $n \to \infty$. Since the indicators $\mathbf{1}_{\{X \leq r_n\}}$ are bounded, we have $G(r_n) \downarrow G(q)$ a.s. It follows that $\mathbb{P}(F_q) = 0$, and then $\mathbb{P}(F) = 0$, where $F = \cup_{q \in \mathbb{Q}} F_q$. Note that F is the event on which $G(\cdot)$ is not right-continuous. Let then H be the set on which $\lim_{q \to \infty} G(q) < 1$ or $\lim_{q \to -\infty} G(q) > 0$. By a similar argument, we have $\mathbb{P}(H) = 0$. On the set $\Omega_0 := (E \cup F \cup H)^c$, the random function $G(\cdot)$ has the properties of a distribution function on the rationals. Note that $\Omega_0 \in \mathcal{G}$. Let F^0 be an arbitrary distribution function and define for $x \in \mathbb{R}$

$$\hat{F}(x) = \mathbf{1}_{\Omega_0^c} F^0(x) + \mathbf{1}_{\Omega_0} \inf_{q>x} G(q).$$

It is easy to check that $\hat{F}(\cdot)$ is a distribution function for each hidden argument ω. Moreover, $\hat{F}(x)$ is $\mathcal{G}$-measurable and since $\inf_{q>x} \mathbf{1}_{\{X \leq q\}} = \mathbf{1}_{\{X \leq x\}}$, we obtain that $\hat{F}(x)$ is a version of $\mathbb{P}(X \leq x|\mathcal{G})$. This finishes the proof of the construction of a conditional distribution function of X given $\mathcal{G}$.

For every ω, the distribution function $\hat{F}(\cdot)(\omega)$ generates a probability measure $\mathbb{P}^X(\cdot)(\omega)$ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Let $\mathcal{C}$ be the class of Borel-measurable sets B for which $\mathbb{P}^X(B)$ is a version of $\mathbb{P}(X \in B|\mathcal{G})$. It follows that all intervals $(-\infty, x]$ belong to $\mathcal{C}$. Moreover, $\mathcal{C}$ is a d-system. By virtue of Dynkin's Lemma 1.13, $\mathcal{C} = \mathcal{B}(\mathbb{R})$. □

Proposition 8.12 *Let X be a random variable and $h : \mathbb{R} \to \mathbb{R}$ be a Borel-measurable function. Let $\hat{\mathbb{P}}^X$ be a regular conditional distribution of X given $\mathcal{G}$. If $h(X) \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$, then*

$$\int h(x)\, \hat{\mathbb{P}}^X(\mathrm{d}x) \tag{8.4}$$

is a version of the conditional expectation $\mathbb{E}[h(X)|\mathcal{G}]$.

Proof Consider the collection $\mathcal{H}$ of all Borel functions h for which (8.4) is a version of $\mathbb{E}[h(X)|\mathcal{G}]$. Clearly, in view of Theorem 8.11 the indicator functions

$\mathbf{1}_B$ for $B \in \mathcal{B}(\mathbb{R})$ belong to $\mathcal{H}$ and so do linear combinations of them. If $h \geq 0$, then we can find nonnegative simple functions h_n that convergence to h in a monotone way. Monotone convergence for conditional expectations yields $h \in \mathcal{H}$. If h is arbitrary, we split as usual $h = h^+ - h^-$ and apply the previous step. □

Once more we emphasize that regular conditional probabilities in general don't exist. The general definition of conditional expectation would be pointless if every conditional expectation could be computed by Proposition 8.12. The good news is that in most common situations Proposition 8.12 can be applied.

In Exercise 8.8 you find an explicit expression for the regular conditional distribution of a random variable X given another random variable Y.

8.4 Exercises

8.1 Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $\mathcal{A} = \{A_1, \ldots, A_n\}$ be a partition of Ω, where the A_i belong to $\mathcal{F}$. Let $X \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ and $\mathcal{G} = \sigma(\mathcal{A})$. Show that any version of $\mathbb{E}[X|\mathcal{G}]$ is of the form $\sum_{i=1}^n a_i \mathbf{1}_{A_i}$ and determine the a_i.

8.2 Let Y be a (real) random variable or random vector on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Assume that Z is another random variable that is $\sigma(Y)$-measurable. Use the standard machine to show that there exists a Borel-measurable function h on $\mathbb{R}$ such that $Z = h(Y)$. Conclude that for integrable X it holds that $\mathbb{E}[X|Y] = h(Y)$ for some Borel-measurable function h.

8.3 Prove Proposition 8.10.

8.4 Prove the conditional version of Fatou's lemma, Theorem 8.7(ii).

8.5 Prove the conditional Dominated Convergence theorem, Theorem 8.7(iii).

8.6 Let (X, Y) have a bivariate normal distribution with $\mathbb{E}X = \mu_X$, $\mathbb{E}Y = \mu_Y$, $\operatorname{Var} X = \sigma_X^2$, $\operatorname{Var} Y = \sigma_Y^2$ and $\operatorname{Cov}(X, Y) = c$. Let

$$\hat{X} = \mu_x + \frac{c}{\sigma_Y^2}(Y - \mu_Y).$$

Show that $\mathbb{E}(X - \hat{X})Y = 0$. Show also (use a special property of the bivariate normal distribution) that $\mathbb{E}(X - \hat{X})g(Y) = 0$ if g is a Borel-measurable function such that $\mathbb{E}g(Y)^2 < \infty$. Conclude that $\hat{X}$ is a version of $\mathbb{E}[X|Y]$.

8.7 Let $X, Y \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ and assume that $\mathbb{E}[X|Y] = Y$ and $\mathbb{E}[Y|X] = X$ (or rather, versions of them are a.s. equal). Show that $\mathbb{P}(X = Y) = 1$. *Hint:* Start to work on $\mathbb{E}(X - Y)\mathbf{1}_{\{X > z, Y \leq z\}} + \mathbb{E}(X - Y)\mathbf{1}_{\{X \leq z, Y \leq z\}}$ for arbitrary $z \in \mathbb{R}$.

8.8 Let X and Y be random variables and assume that (X, Y) admits a density f w.r.t. Lebesgue measure on $(\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2))$. Let f_Y be the marginal density of Y. Define $\hat{f}(x|y)$ by

$$\hat{f}(x|y) = \begin{cases} \frac{f(x,y)}{f_Y(y)} & \text{if } f_Y(y) > 0 \\ 0 & \text{else.} \end{cases}$$

Assume that $\mathbb{E}|h(X)| < \infty$. Put $\hat{h}(y) = \int_{\mathbb{R}} h(x)\hat{f}(x|y)\,\mathrm{d}x$. Show that $\hat{h}(Y)$ is a version of $\mathbb{E}[h(X)|Y]$. Show also that

$$\hat{\mathbb{P}}(E) = \int_E \hat{f}(x|Y)\,\mathrm{d}x$$

defines a regular conditional probability on $\mathcal{B}(\mathbb{R})$ given Y. What is the exceptional set N of Section 8.3?

8.9 Consider $(\Omega, \mathcal{F}, \mathbb{P}) = ([0,1], \mathcal{B}, \lambda)$, where λ denotes both the Lebesgue measure on $\mathcal{B}$ and the outer measure as in (2.3). Let E be a subset of $[0,1]$ for which $\nu(E) = \lambda(E^c) = 1$ (clearly E is not λ-measurable, see Exercise 2.6 for existence of such a set). Let $\mathcal{F}$ be the smallest σ-algebra that contains $\mathcal{B}$ and E. Show that $F \in \mathcal{F}$ iff there are $B_1, B_2 \in \mathcal{B}$ such that $F = (B_1 \cap E) \cup (B_2 \cap E^c)$. For such a F we *define* $\hat{\mathbb{P}}(F) = \frac{1}{2}(\lambda(B_1) + \lambda(B_2))$. Check that this definition is independent of the specific B_1 and B_2 and that $\mathbb{P}$ is a probability measure on $\mathcal{F}$. Show that there exists no regular conditional probability $\hat{\mathbb{P}}$ on $\mathcal{F}$ given $\mathcal{B}$.

8.10 Consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Two random variables X and Y are called conditionally independent given a sub-σ-algebra $\mathcal{G}$ of $\mathcal{F}$ if for all bounded Borel functions $f, g : \mathbb{R} \to \mathbb{R}$ it holds that $\mathbb{E}[f(X)g(Y)|\mathcal{G}] = \mathbb{E}[f(X)|\mathcal{G}]\mathbb{E}[g(Y)|\mathcal{G}]$.

(a) Show that X and Y are conditionally independent given $\mathcal{G}$ iff for every bounded measurable function $f : \mathbb{R} \to \mathbb{R}$ it holds that $\mathbb{E}[f(X)|\sigma(Y) \vee \mathcal{G}] = \mathbb{E}[f(X)|\mathcal{G}]$.

(b) Show (by examples) that in general conditional independence is not implied by independence, nor vice versa.

(c) If X and Y are given random variables, give an example of a σ-algebra $\mathcal{G}$ that makes X and Y conditionally independent.

(d) Propose a definition of conditional independence of two σ-algebras $\mathcal{H}_1$ and $\mathcal{H}_2$ given $\mathcal{G}$ that is such that conditional independence of X and Y given $\mathcal{G}$ can be derived from it as a special case.

8.11 (*Hölder's inequality for conditional expectations*) Let $X \in \mathcal{L}^p(\Omega, \mathcal{F}, \mathbb{P})$ and $Y \in \mathcal{L}^q(\Omega, \mathcal{F}, \mathbb{P})$, where $p, q \in [1, \infty]$, $\frac{1}{p} + \frac{1}{q} = 1$. Then

$$\mathbb{E}[|XY|\big|\mathcal{G}] \leq (\mathbb{E}[|X|^p|\mathcal{G}])^{1/p}(\mathbb{E}[|Y|^q|\mathcal{G}])^{1/q}. \tag{8.5}$$

It is sufficient to show this for nonnegative X and Y, so assume $X, Y \geq 0$ a.s.

(a) Let $U = \mathbb{E}[X^p|\mathcal{G}]$ and $V = \mathbb{E}[Y^q|\mathcal{G}]$ and $H = \{U, V > 0\}$. Suppose that

$$\mathbf{1}_H\mathbb{E}[XY|\mathcal{G}] \leq \mathbf{1}_H(\mathbb{E}[X^p|\mathcal{G}])^{1/p}(\mathbb{E}[Y^q|\mathcal{G}])^{1/q}]. \tag{8.6}$$

Show that Hölder's inequality (8.5) follows from (8.6).

(b) Show that

$$\mathbb{E}[\mathbf{1}_G\mathbf{1}_H\frac{\mathbb{E}[XY|\mathcal{G}]}{UV} \leq \mathbb{E}[\mathbf{1}_G\mathbf{1}_H]$$

holds for all $G \in \mathcal{G}$ and deduce (8.6).

8.12 Finish the proof of Theorem 8.8 (iv), i.e. show that the assertion also holds without the boundedness condition on h.

9 Martingales and their relatives

In this chapter we define *martingales, sub- and supermartingales*. In the next chapter we formulate convergence theorems for them and see how these can be applied to give elegant proofs of some central results in probability theory. The power of martingales is the combination of their main defining property, that is shared by a rich class of special cases, and the strong convergence results that can be obtained in spite of a seemingly innocent definition.

9.1 Basic concepts and definition

As we shall see below, a martingale is a *stochastic process* with certain defining properties. A *stochastic process*, or simply a *process*, (in discrete time) is nothing else but a sequence of random variables defined on some underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The time set is often taken as $\{0, 1, 2, \ldots\}$ in which case we have e.g. a sequence of random variables $X_0, X_1, X_2, \ldots$. Such a sequence as a whole is often denoted by X. So we have $X = (X_n)_{n\geq 0}$. Unless otherwise stated, all process have their values in $\mathbb{R}$, while the extension to $\mathbb{R}^d$-valued processes should be clear.

We shall need a sequence of sub-σ-algebras of $\mathcal{F}$, that form a *filtration* $\mathbb{F}$. This means that $\mathbb{F} = (\mathcal{F}_n)_{n\geq 0}$, where each $\mathcal{F}_n$ is a σ-algebra satisfying $\mathcal{F}_n \subset \mathcal{F}$, and moreover $\mathcal{F}_n \subset \mathcal{F}_{n+1}$, for all $n \geq 0$. The sequence $\mathbb{F}$ is thus increasing. Recall that in general a union of σ-algebras is not a σ-algebra itself. We define $\mathcal{F}_\infty := \sigma(\cup_{n=0}^\infty \mathcal{F}_n)$. Obviously $\mathcal{F}_n \subset \mathcal{F}_\infty$ for all n. If X is a stochastic process, then one defines $\mathcal{F}_n^X := \sigma(X_0, \ldots, X_n)$. It is clear that $\mathbb{F}^X := (\mathcal{F}_n^X)_{n\geq 0}$ is a filtration.

Given a filtration $\mathbb{F}$, we shall often consider $\mathbb{F}$-adapted processes. A process Y is $\mathbb{F}$-adapted (or adapted to $\mathbb{F}$, or just adapted), if for all n the random variable Y_n is $\mathcal{F}_n$-measurable ($Y_n \in \mathcal{F}_n$). If $\mathbb{F} = \mathbb{F}^X$ for some process X, then another process Y is $\mathbb{F}^X$-adapted, iff for all n, there exists a Borel function $f_n : \mathbb{R}^{n+1} \to \mathbb{R}$ such that $Y_n = f_n(X_0, \ldots, X_n)$, see Exercise 8.2. Obviously X is adapted to $\mathbb{F}^X$.

A filtration can be interpreted as an information flow, where each $\mathcal{F}_n$ represents the available information up to time n. For $\mathbb{F} = \mathbb{F}^X$, the information comes from the process X and the information at time n is presented by events in terms of $X_0, \ldots, X_n$.

Having introduced all the relevant underlying terminology, we are ready to define martingales.

Definition 9.1 A stochastic process $M = (M_n)_{n\geq 0}$ is called a *martingale* (or $\mathbb{F}$-martingale), if it is adapted to a filtration $\mathbb{F}$, if $M_n \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ for all $n \geq 0$ and if

$$\mathbb{E}[M_{n+1}|\mathcal{F}_n] = M_n \text{ a.s.} \tag{9.1}$$

Equation (9.1), valid for all $n \geq 0$ is called the *martingale property* of M.

Remark 9.2 The equality (9.1) should be read in the sense that M_n is a version of the conditional expectation $\mathbb{E}[M_{n+1}|\mathcal{F}_n]$. Although we have always been careful in formulating properties of conditional expectations in terms of version, we will drop this care and leave it to the reader to properly interpret statements given below concerning conditional expectations.

Remark 9.3 The definition of martingales has been given in terms of 'one-step-ahead' conditional expectations. If we change (9.1) in the sense that we replace on the left hand side $\mathbb{E}[M_{n+1}|\mathcal{F}_n]$ with $\mathbb{E}[M_m|\mathcal{F}_n]$, $m \geq n+1$ arbitrary, we obtain an equivalent definition. (Use the tower property to check this!) A similar remark applies to the definitions of sub- and supermartingales, that we will meet shortly.

Let us give some concrete examples of martingales.

Example 9.4 Let X be a process consisting of independent random variables X_n, with $n \geq 1$ and assume that $X_n \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ for all n. Put $X_0 = 0$ and $S_n = \sum_{k=0}^n X_k = \sum_{k=1}^n X_k$. Take $\mathbb{F} = \mathbb{F}^X$. Obviously, S is adapted to $\mathbb{F}$, each S_n belongs to $\mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ and we have from properties of conditional expectation (see Proposition 8.6 and Theorem 8.8) a.s. the following chain of equalities.

$$\begin{aligned}\mathbb{E}[S_{n+1}|\mathcal{F}_n] &= \mathbb{E}[S_n + X_{n+1}|\mathcal{F}_n] \\ &= \mathbb{E}[S_n|\mathcal{F}_n] + \mathbb{E}[X_{n+1}|\mathcal{F}_n] \\ &= S_n + \mathbb{E}X_{n+1}.\end{aligned}$$

Hence, S is a martingale iff $\mathbb{E}X_n = 0$ for all n. We conclude that a martingale is an extension of the partial sum process generated by a sequence of independent random variables having expectation zero.

The previous example was in terms of sums. The next one involves products.

Example 9.5 Let X be a process consisting of independent random variables X_n, with $n \geq 1$ and assume that $X_n \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ for all n. Put $X_0 = 1$ and $P_n = \prod_{k=0}^n X_k = \prod_{k=1}^n X_k$. Take $\mathbb{F} = \mathbb{F}^X$. Obviously, P is adapted to $\mathbb{F}$. We even have that each P_n belongs to $\mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$, because of the product rule for products of independent random variables, see Proposition 4.43. From properties of conditional expectation we obtain a.s.

$$\begin{aligned}\mathbb{E}[P_{n+1}|\mathcal{F}_n] &= \mathbb{E}[P_n\, X_{n+1}|\mathcal{F}_n] \\ &= P_n\, \mathbb{E}[X_{n+1}|\mathcal{F}_n] \\ &= P_n\, \mathbb{E}X_{n+1}.\end{aligned}$$

Hence, P is a martingale iff $\mathbb{E}X_n = 1$ for all n.

Here is another fundamental example.

Example 9.6 Let X be a random variable with $\mathbb{E}|X| < \infty$ and $\mathbb{F}$ a filtration. Put $M_n = \mathbb{E}[X|\mathcal{F}_n]$, $n \geq 0$. By the tower property (see Theorem 8.8(i)) of conditional expectation we obtain $\mathbb{E}[M_{n+1}|\mathcal{F}_n] = \mathbb{E}[\mathbb{E}[X|\mathcal{F}_{n+1}]|\mathcal{F}_n] = \mathbb{E}[X|\mathcal{F}_n] = M_n$, where all equalities are to be understood in the a.s. sense. The process M is thus a martingale.

Assume further that $X \in \mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$. We can interpret M_n as the *best prediction* of X given the 'information' $\mathcal{F}_n$. Take as a measure of 'prediction error' the mean square loss $\mathbb{E}(X-Y)^2$, $Y \in \mathcal{L}^2(\Omega, \mathcal{F}_n, \mathbb{P})$, see Proposition 8.10. Put $v_n := \mathbb{E}(M_n - X)^2$. One can show that the v_n are decreasing (Exercise 9.2), which supports our intuitive understanding that with more information one should be able to predict better.

Next to martingales also super- and submartingales are of considerable interest.

Definition 9.7 A stochastic process $X = (X_n)_{n\geq 0}$ is called a *submartingale* (or $\mathbb{F}$-submartingale), if it is adapted to a filtration $\mathbb{F}$, if $X_n \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ for all $n \geq 0$ and if

$$\mathbb{E}[X_{n+1}|\mathcal{F}_n] \geq X_n \text{ a.s.} \tag{9.2}$$

Inequality (9.2), valid for all $n \geq 0$ is called the *submartingale property* of X. A stochastic process $X = (X_n)_{n\geq 0}$ is called a *supermartingale* (or $\mathbb{F}$-supermartingale), if $-X$ is a submartingale.

The expressions (9.1) and (9.2) can be interpreted by saying that a martingale follows a constant *trend*, whereas a submartingale displays an increasing trend. Of course, a supermartingale fluctuates around a decreasing trend. Notice that it also follows from (9.1) that a martingale has constant expectation, $\mathbb{E}M_n = \mathbb{E}M_0$, for all n. On the other hand, a submartingale has increasing expectations, as follows from (9.2).

Example 9.8 We revisit the first two examples given above. In Example 9.4 we obtain a submartingale if the X_n have positive expectation, resulting in an increasing trend, whereas a negative expectation for the X_n turns S into a supermartingale. In Example 9.5 we now restrict the X_n to be positive. Then P will become a submartingale if $\mathbb{E}X_n \geq 1$ and a supermartingale for $\mathbb{E}X_n \leq 1$.

If X is any process, we define the process ΔX by

$$\Delta X_n = X_n - X_{n-1},\ n \geq 1.$$

It trivially follows that $X_n = X_0 + \sum_{k=1}^n \Delta X_k$. Sometimes it is convenient to adopt the convention $\Delta X_0 = X_0$, from which we then obtain $X_n = \sum_{k=0}^n \Delta X_k$. The martingale property of an adapted integrable process X can then be formulated as $\mathbb{E}[\Delta X_{n+1}|\mathcal{F}_n] = 0$ a.s. for $n \geq 0$. For submartingales it holds that $\mathbb{E}[\Delta X_{n+1}|\mathcal{F}_n] \geq 0$ a.s.

If you want to interpret random variables ξ_n as the payoffs (profits or losses, depending on the sign) of your bets in the n-th game of a series, then $S_n = \sum_{k=1}^n \xi_k$

would be your accumulated total capital after n games. Here we have $\Delta S_n = \xi_n$. If S is a submartingale, you are playing a favorable game, and if S is martingale you are playing a fair game. It should be clear what an unfavorable game is. In the next subsection we will investigate whether it is possible by playing clever strategies to turn an unfavorable game into a favorable one.

9.2 Stopping times and martingale transforms

Somebody who would know at time $n-1$ what is going to be the outcome of a random experiment to be held at time n, is in that sense clairvoyant. He is able to 'predict' this future outcome. This motivates the notion of a *predictable* (also called *previsible*) process.

Definition 9.9 Given a filtration $\mathbb{F}$, a process $Y = (Y_n)_{n\geq 1}$ is called $\mathbb{F}$-predictable (or just predictable) if $Y_n \in \mathcal{F}_{n-1}$, $n \geq 1$. A convenient additional convention is to set $Y_0 = 0$.

You may consider a predictable process Y to be a *strategy*, it tells you what your action at time n is going to be, given that you use your information available at time $n-1$. In a trivial sense, you 'perfectly' predict Y_n at time $n-1$.

Suppose that a sequence of random variables ξ_n (with $\xi_0 = 0$) represents the payoff of a game at time n when you make a *unit bet*, then $Y_n\xi_n$ would be the payoff when you bet Y_n units at time n. When you are not a clairvoyant and have no insider information, your bet Y_n cannot depend on future outcomes ξ_m, $m \geq n$ of the game, but you are allowed to let them depend on what has been realized before, i.e. to take Y as $\mathbb{F}^\xi$-predictable. The accumulated, or total, earnings up to time n are $S_n = \sum_{k=1}^n Y_k\xi_k$ (with $S_0 = 0$). If we let $X_n = \sum_{k=0}^n \xi_k$, we get $\Delta S_n = Y_n \Delta X_n$. We also have $S_n = \sum_{k=1}^n Y_k \Delta X_k$. This notation is a discrete time analogue of an expression like $S_t = \int_0^t Y_s \, \mathrm{d}X_s$. Such an expression, for suitably defined stochastic processes with a continuous time parameter are called *stochastic integrals*. Because of the analogy, a process like S above, is sometimes called a *discrete time stochastic integral*, in particular when X is a martingale. In that case one also speaks of a *martingale transform*. A common notation is to write in this case $S = Y \cdot X$ for the process S. Note that the 'dot' here is *not* the multiplication operator. If Z is another (predictable) process, we have $Z \cdot S = (ZY) \cdot X$. The term *martingale transform* is justified by the following proposition.

Proposition 9.10 *Let X be an adapted process and Y a predictable process. Assume that the X_n are in $\mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ as well as the $Y_n \Delta X_n$. Let $S = Y \cdot X$. The following results hold.*

(i) If X is martingale, so is S.

(ii) If X is a submartingale (supermartingale) and if Y is nonnegative, also S is a submartingale (supermartingale).

Proof Clearly, S is adapted. All three statements follow from the identity $\mathbb{E}[\Delta S_n|\mathcal{F}_{n-1}] = Y_n\mathbb{E}[\Delta X_n|\mathcal{F}_{n-1}]$ a.s., which holds by virtue of Theorem 8.8(ii), and the definitions of martingale and sub-, supermartingale. □

Back to the interpretation in terms of betting and games. If you play an unfavorable game, the accumulated 'gains per unit bet' process X is a supermartingale. A (predictable) strategy Y has to be nonnegative (negative bets are usually not allowed). Proposition 9.10 tells us that whatever strategy you play, your accumulated gains process $Y \cdot X$ will still be a supermartingale, an unfavorable game.

Although, as we have explained above, you are not able to change the nature of the game by using predictable strategies, it is still a good question to ask for *optimal strategies*. A strategy could be called optimal, if it maximizes S_N at some fixed end time N. Questions like this are answered in the theory of optimal stopping. We don't treat this theory in this course, but we do pay attention to one of the basic ingredients, *stopping times*, sometimes also called optional times. A formal definition follows.

Definition 9.11 Let $\mathbb{F}$ be a filtration. A mapping $T : \Omega \to \{0, 1, 2, \ldots\} \cup \{\infty\}$ is called a stopping time if for all $n \in \{0, 1, 2, \ldots\}$ it holds that $\{T = n\} \in \mathcal{F}_n$.

Let us make a few observations. Recall the inclusions $\mathcal{F}_n \subset \mathcal{F}_\infty \subset \mathcal{F}$ hold for all n. The event $\{T = \infty\}$ can be written as $(\cup_{n=0}^{\infty}\{T = n\})^c$. Since $\{T = n\} \in \mathcal{F}_n \subset \mathcal{F}_\infty$, we have $\{T = \infty\} \in \mathcal{F}_\infty$. Hence the requirement $\{T = n\} \in \mathcal{F}_n$ in Definition 9.11 extends to $n = \infty$.

Another observation is that T is a stopping time iff $\{T \leq n\} \in \mathcal{F}_n$ is true for all $n \in \{0, 1, 2, \ldots\}$. One implication follows from $\{T \leq n\} = \cup_{k=0}^{n}\{T = k\}$, the other from $\{T = n\} = \{T \leq n\} \setminus \{T \leq n - 1\}$.

The name stopping time can be justified as follows. If you bet, you want to reach a goal, which could be trying to gain a profit of at least 1000 euro, and you stop playing, once you have reached your goal (if it ever happens). If your gains process is S, in this case you will stop playing at time T, where $T = \inf\{n \geq 0 : S_n \geq 1000\}$, with the convention $\inf \emptyset = \infty$. Indeed, as we will see in the next proposition, that describes a slightly more general situation, T is a stopping time if S is adapted.

Proposition 9.12 *Let $\mathbb{F}$ be a filtration and X an adapted process. Let $B \in \mathcal{B}$ be a Borel set in $\mathbb{R}$ and let $T = \inf\{n \geq 0 : X_n \in B\}$. Then T is a stopping time.*

Proof The event $\{T \leq n\}$ can alternatively be expressed as $\cup_{k=0}^{n}\{X_k \in B\}$ and here $\{X_k \in B\} \in \mathcal{F}_k \subset \mathcal{F}_n$ by adaptedness of X. □

Let T be a stopping time and $n \in \{0, 1, 2, \ldots\}$. Define $T_n(\omega) := T(\omega) \wedge n$, the minimum of $T(\omega)$ and n. Then T_n is also a stopping time. Indeed, for $k < n$ we have $\{T_n \leq k\} = \{T \leq k\} \in \mathcal{F}_k$, whereas $\{T_n \leq k\} = \Omega$ for $k \geq n$. Usually we write $T \wedge n$ for T_n.

If X is an adapted process and T a stopping time, we define the *stopped process* X^T by $X_n^T(\omega) := X_{T(\omega)\wedge n}(\omega)$, $n \geq 0$. Note that $X_0^T = X_0$ and $X_n^T(\omega) = X_{T(\omega)}(\omega)$ for $n \geq T(\omega)$, abbreviated $X_n^T = X_T$ on $\{T \leq n\}$. This explains the terminology.

Proposition 9.13 *If X is an adapted process and T a stopping time, then X^T is adapted too. Moreover, if X is a supermartingale, so is X^T and then $\mathbb{E}X_n^T \leq \mathbb{E}X_0$. If X is a martingale, then X^T is a martingale too and $\mathbb{E}X_n^T = \mathbb{E}X_0$.*

Proof Let X be adapted. Write $X_n^T = \sum_{k=0}^n \mathbf{1}_{\{T=k\}} X_k + \mathbf{1}_{\{T>n\}} X_n$ and $\mathcal{F}_n$-measurability of X_n^T easily follows. To show the other assertions, we define the process $Y_n = \mathbf{1}_{\{T\geq n\}}$, $n \geq 1$. Note that $\{Y_n = 0\} = \{T \leq n-1\} \in \mathcal{F}_{n-1}$. Hence Y is predictable. A simple computation shows that $\Delta X_k^T = Y_k \Delta X_k$, hence $X^T = X_0 + Y \cdot X$. The assertions now follow from Proposition 9.10. $\square$

For a stopping time T and an adapted process X, the obvious definition of the random variable X_T (the value of the process at the stopping time) would be $X_T(\omega) = X_{T(\omega)}(\omega)$, so $X_T = X_n$ on $\{T = n\}$. But since T may assume the value ∞, there is a problem, because we often only have X_n for $n < \infty$. The problem can be circumvented by defining X_T only on $\{T < \infty\}$ and setting it equal to zero outside that set, which leads to $X_T = X_T \mathbf{1}_{\{T<\infty\}}$. Another way out could be to assume that we also have a random variable X_∞, in which case X_T is properly defined everywhere. Of course the problem disappears if T is finite, or even bounded.

Example 9.14 Here is a seemingly winning strategy for a fair game. Let $(\xi)_{n\geq 1}$ be an *iid* sequence of random variable with $\mathbb{P}(\xi_n = \pm 1) = \frac{1}{2}$. Then the process X, $X_n = \sum_{k=1}^n \xi_k$ with $X_0 = 0$, is a martingale with $\mathbb{E}X_n = 0$. For any strategy Y, the total earnings process starting from zero is again $S = Y \cdot X$. Note that $S_n = S_{n-1} + Y_n \xi_n$. The strategy of interest is defined by the predictable process Y, where $Y_n = 1 - S_{n-1}$ and $Y_1 = 1$. What happens at time n if $\xi_n = 1$? In that case, $S_n = S_{n-1} + (1 - S_{n-1}) = 1$ and then $Y_{n+1} = 0$, so $S_{n+1} = S_n = 1$, and so $S_k = 1$ for $k \geq n$. If $\xi_n = -1$, we obtain $S_n = S_{n-1} - (1 - S_{n-1}) = 2S_{n-1} - 1$. Hence $Y_{n+1} = 2(1 - S_{n-1}) = 2Y_n$. It follows that the strategy doubles, as long as the ξ_n are equal to -1, which results in $Y_n = 2^{n-1}$ on the event $\{\xi_1 = \cdots = \xi_{n-1} = -1\}$ and zero on its complement. As soon as $\xi_n = 1$, you stop playing and go home with your profit $S_n = 1$. Let $T = \inf\{n : S_n = 1\}$. One checks that $\mathbb{P}(T = n) = \mathbb{P}(\xi_1 = \cdots = \xi_{n-1} = -1, \xi_n = 1) = 2^{-n}$. Hence $\mathbb{P}(T < \infty) = 1$. Moreover $S_T = 1$. Here is the pitfall of this strategy. The last (non-zero) bet is equal to $Y_T = 2^{T-1} = \sum_{n=1}^\infty 2^{n-1} \mathbf{1}_{\{T=n\}}$, which assumes arbitrary large values and has expectation $\mathbb{E}Y_T = \sum_{n=1}^\infty 2^{n-1} \mathbb{P}(T = n) = \infty$. Therefore, you need an infinite capital to successfully play this strategy to the very end.

We have seen an example of a martingale, S, that has expectation zero, $\mathbb{E}S_n = 0$ for all $n \geq 1$, whereas $\mathbb{E}S_T = 1$. This was essentially caused by the facts that S is not bounded, $\mathbb{P}(S_n = 1 - 2^n) > 0$, and that T is not bounded.

Theorem 9.15 *Let X be supermartingale and T an a.s. finite stopping time. Then $\mathbb{E}X_T \leq \mathbb{E}X_0$ under either of the assumptions*

(i) X is a.s. bounded from below by random variable $Y \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$, or

(ii) T is bounded, i.e. there exists $N < \infty$ such that $\mathbb{P}(T \leq N) = 1$ or

(iii) The process ΔX is bounded by a constant C and $\mathbb{E}T < \infty$.

If X is a martingale, then $\mathbb{E}X_T = \mathbb{E}X_0$ under (ii) and (iii) and also under the assumption (iv) that X is bounded.

Proof Let X be a supermartingale. We know from Proposition 9.13 that X^T is a supermartingale, hence $\mathbb{E}|X_n^T| < \infty$ for all n, and $\mathbb{E}X_n^T \leq \mathbb{E}X_0$. Note that $X_n^T \overset{\text{a.s.}}{\to} X_T$, since $T < \infty$ a.s.

Assume (i). Then $X_n^T - Y \geq 0$. Hence we can apply Fatou's lemma and Proposition 9.13 to get $\mathbb{E}(X_T - Y) = \mathbb{E}\lim_{n\to\infty}(X_n^T - Y) \leq \liminf_{n\to\infty} \mathbb{E}(X_n^T - Y) = \liminf_{n\to\infty} \mathbb{E}X_n^T - \mathbb{E}Y \leq \mathbb{E}X_0 - \mathbb{E}Y$. Hence $\mathbb{E}X_T \leq \mathbb{E}X_0$.

If we assume (ii), then we only need $X_T = X_N^T$, whose expectation is at most equal to $\mathbb{E}X_0$ by Proposition 9.13.

Finally, we assume (iii) under which we have $|X_n^T - X_0| \leq CT$. Therefore we can apply the Dominated Convergence Theorem to get $\mathbb{E}(X_T - X_0) = \lim_{n\to\infty} \mathbb{E}(X_n^T - X_0) \leq 0$.

If X is a martingale, then the assertion follows as above under (ii) and (iii), whereas under (iv) it follows by dominated convergence. □

9.3 Doob's decomposition

The celebrated decomposition theorem by Doob for submartingales is presented in Theorem 9.16. It says that a submartingale can be decomposed as the sum of a predictable increasing *trend* and a martingale.

Theorem 9.16 *Let X be an adapted process and assume that $\mathbb{E}|X_n| < \infty$ for all n. Then there exists a predictable process A and a martingale M such that $X_n = A_n + M_n$ a.s. for $n \geq 1$. The process A is a.s. unique in the sense that if $A' + M'$ is another additive decomposition of X with the same properties, then $\mathbb{P}(A_n = A'_n, \forall n \geq 1) = 1$. Moreover, A is a.s. an increasing process iff X is a submartingale.*

Proof Note first that $\mathbb{E}|\Delta X_n| < \infty$ for all n. We define the predictable process A is follows. For $n \geq 1$, we put $\Delta A_n := \mathbb{E}[\Delta X_n | \mathcal{F}_{n-1}]$ (or rather, any version of it) and $A_n = \sum_{k=1}^n \Delta A_k$. That A is predictable should be immediately clear. Knowing this, we define M, by setting $M_0 = X_0$ and $\Delta M_n := \Delta X_n - \Delta A_n$ for $n \geq 1$ and finally $M_n = M_0 + \sum_{k=1}^n \Delta M_k$. By its definition, M is a martingale, since $\mathbb{E}[\Delta M_n | \mathcal{F}_{n-1}] = 0$ a.s. Note that $\Delta A_n \geq 0$ a.s. if X is a submartingale, in which case A becomes increasing. The converse statement is as easy to prove.

To prove uniqueness, we argue as follows. Since $X_n = A_n + M_n = A'_n + M'_n$, we have $A'_n - A_n = M'_n - M_n$ and so $A' - A$ becomes a predictable martingale. These properties yield $A'_n - A_n = \mathbb{E}[A'_n - A_n | \mathcal{F}_{n-1}] = A'_{n-1} - A_{n-1}$ a.s. It follows that $\mathbb{P}(\Delta A'_n = \Delta A_n) = 1$ for each individual n. But then also $\mathbb{P}(\Delta A'_n =$

$\Delta A_n, \forall n \geq 1) = 1$, since it is the countable union of events with probability one. Since $A'_0 = A_0 = 0$, we get the assertion about unicity. $\square$

A martingale is called *square integrable* if $\mathbb{E}M_n^2 < \infty$ for all n.

Corollary 9.17 *Let M be a square integrable martingale. Then there exists a unique (in the sense of Theorem 9.16) increasing predictable process $\langle M \rangle$ with $\langle M \rangle_0 = 0$ such that $M^2 - \langle M \rangle$ is a martingale. Moreover, for $n \geq 1$ the random variable $\Delta\langle M \rangle_n$ is (a version of) the conditional variance of M_n given $\mathcal{F}_{n-1}$, i.e.*

$$\Delta\langle M \rangle_n = \mathbb{E}[(M_n - \mathbb{E}[M_n|\mathcal{F}_{n-1}])^2|\mathcal{F}_{n-1}] = \mathbb{E}[(M_n - M_{n-1})^2|\mathcal{F}_{n-1}] \text{ a.s.}$$

It follows that 'Pythagoras's theorem' holds for square integrable martingales,

$$\mathbb{E}M_n^2 = \mathbb{E}(M_0 + \sum_{k=1}^{n} \Delta M_k)^2 = \mathbb{E}M_0^2 + \sum_{k=1}^{n} \mathbb{E}(\Delta M_k)^2,$$

Proof First we note that M^2 is a submartingale, see Exercise 9.3. Hence the previous proposition applies and the only thing left to prove is the statement about the conditional variance. Since M is a martingale, we have a.s.

$$\begin{aligned}\mathbb{E}[(M_n - M_{n-1})^2|\mathcal{F}_{n-1}] &= \mathbb{E}[M_n^2 - 2M_nM_{n-1} + M_{n-1}^2|\mathcal{F}_{n-1}] \\ &= \mathbb{E}[M_n^2|\mathcal{F}_{n-1}] - M_{n-1}^2 \\ &= \mathbb{E}[M_n^2 - M_{n-1}^2|\mathcal{F}_{n-1}],\end{aligned}$$

which is by the proof of Theorem 9.16 just the definition of $\Delta\langle M \rangle_n$. $\square$

The process $\langle M \rangle$ of Corollary 9.17 is also called the *predictable quadratic variation process* of M. If M and N are two square integrable martingales, then there exists a unique predictable process $\langle M, N \rangle$ with $\langle M, N \rangle_0 = 0$, called the predictable covariation process of M and N, such that $MN - \langle M, N \rangle$ is martingale. See Exercise 9.13.

Remark 9.18 Under the conditions of Corollary 9.17 it holds that $\mathbb{E}M_n^2 = \mathbb{E}M_0^2 + \mathbb{E}\langle M \rangle_n$. Since $\langle M \rangle$ is an a.s. increasing process, it has an almost sure limit $\langle M \rangle_\infty \leq \infty$. It follows that M is bounded in $\mathcal{L}^2$ ($\sup_n \mathbb{E}M_n^2 < \infty$) iff $\mathbb{E}\langle M \rangle_\infty < \infty$.

9.4 Optional sampling

Let $\mathbb{F}$ be a filtration. For a stopping time T we define the σ-algebra

$$\mathcal{F}_T := \{F \subset \Omega : F \cap \{T \leq n\} \in \mathcal{F}_n \text{ for every } n\}.$$

If S and T are stopping times with $S \leq T$, then $\mathcal{F}_S \subset \mathcal{F}_T$. If X is a process with index set $\mathbb{N}$, we define $X_T = \sum_{n=0}^{\infty} X_n \mathbf{1}_{\{T=n\}}$ and so $X_T = X_T \mathbf{1}_{\{T<\infty\}}$. If also X_∞ is defined, we include $n = \infty$ in the last summation. In both cases X_T is a well-defined random variable and even $\mathcal{F}_T$-measurable (check!). All results below are versions of the *optional sampling theorem*, the (sub)martingale property of a process is, under appropriate conditions, preserved when deterministic time indices are replaced with stopping times.

Lemma 9.19 *Let X be a submartingale and T a bounded stopping time, $T \leq N$ say for some $N \in \mathbb{N}$. Then $\mathbb{E}|X_T| < \infty$ and*

$$X_T \leq \mathbb{E}[X_N|\mathcal{F}_T] \quad \text{a.s.} \tag{9.3}$$

Proof Integrability of X_T follows from $|X_T| \leq \sum_{n=0}^{N} |X_n|$. Let $F \in \mathcal{F}_T$. Then, because $F \cap \{T = n\} \in \mathcal{F}_n$ and the fact that X is a submartingale, we have

$$\begin{aligned}
\mathbb{E}[X_N \mathbf{1}_F] &= \sum_{n=0}^{N} \mathbb{E}[X_N \mathbf{1}_F \mathbf{1}_{\{T=n\}}] \\
&\geq \sum_{n=0}^{N} \mathbb{E}[X_n \mathbf{1}_F \mathbf{1}_{\{T=n\}}] \\
&= \sum_{n=0}^{N} \mathbb{E}[X_T \mathbf{1}_F \mathbf{1}_{\{T=n\}}] \\
&= \mathbb{E}[X_T \mathbf{1}_F],
\end{aligned}$$

which is what we wanted to prove. □

Theorem 9.20 *Let X be a uniformly integrable martingale with a last element X_∞, so $X_n = \mathbb{E}[X_\infty|\mathcal{F}_n]$ a.s. for every n. Let T and S be stopping times with $S \leq T$. Then X_T and X_S are integrable and*

(i) $X_T = \mathbb{E}[X_\infty|\mathcal{F}_T]$ *a.s.*

(ii) $X_S = \mathbb{E}[X_T|\mathcal{F}_S]$ *a.s.*

Proof First we show that X_T is integrable. Notice that $\mathbb{E}|X_T|\mathbf{1}_{\{T=\infty\}} = \mathbb{E}|X_\infty|\mathbf{1}_{\{T=\infty\}} \leq \mathbb{E}|X_\infty| < \infty$. Next, because $|X|$ is a submartingale with last element $|X_\infty|$,

$$\begin{aligned}
\mathbb{E}|X_T|\mathbf{1}_{\{T<\infty\}} &= \sum_{n=0}^{\infty} \mathbb{E}|X_n|\mathbf{1}_{\{T=n\}} \\
&\leq \sum_{n=0}^{\infty} \mathbb{E}|X_\infty|\mathbf{1}_{\{T=n\}} \\
&= \mathbb{E}|X_\infty|\mathbf{1}_{\{T<\infty\}} < \infty.
\end{aligned}$$

We proceed with the proof of (i). Notice that $T \wedge n$ is a bounded stopping time for every n. But then by Lemma 9.19 it holds a.s. that

$$\begin{aligned}
\mathbb{E}[X_\infty|\mathcal{F}_{T\wedge n}] &= \mathbb{E}[\mathbb{E}[X_\infty|\mathcal{F}_n]|\mathcal{F}_{T\wedge n}] \\
&= \mathbb{E}[X_n|\mathcal{F}_{T\wedge n}] \\
&= X_{T\wedge n}.
\end{aligned}$$

Let now $F \in \mathcal{F}_T$, then $F \cap \{T \leq n\} \in \mathcal{F}_{T\wedge n}$ and by the above display, we have

$$\mathbb{E}[X_\infty \mathbf{1}_{F\cap\{T\leq n\}}] = \mathbb{E}[X_{T\wedge n}\mathbf{1}_{F\cap\{T\leq n\}}] = \mathbb{E}[X_T \mathbf{1}_{F\cap\{T\leq n\}}].$$

Let $n \to \infty$ and apply the Dominated convergence theorem to get

$$\mathbb{E}[X_\infty \mathbf{1}_F \mathbf{1}_{\{T<\infty\}}] = \mathbb{E}[X_T \mathbf{1}_F \mathbf{1}_{\{T<\infty\}}].$$

Together with the trivial identity $\mathbb{E}[X_\infty \mathbf{1}_F \mathbf{1}_{\{T=\infty\}}] = \mathbb{E}[X_T \mathbf{1}_F \mathbf{1}_{\{T=\infty\}}]$ this yields $\mathbb{E}[X_\infty \mathbf{1}_F] = \mathbb{E}[X_T \mathbf{1}_F]$ and (i) is proved.

For the proof of (ii) we use (i) two times and obtain

$$\mathbb{E}[X_T|\mathcal{F}_S] = \mathbb{E}[\mathbb{E}[X_\infty|\mathcal{F}_T]|\mathcal{F}_S] = \mathbb{E}[X_\infty|\mathcal{F}_S] = X_S.$$

□

Theorem 9.21 *Let X be a submartingale such that $X_n \leq 0$ for all $n = 0, 1, \ldots$ and let $X_\infty = 0$. Let T and S be stopping times with $S \leq T$. Then X_T and X_S are integrable and $X_S \leq \mathbb{E}[X_T|\mathcal{F}_S]$ a.s.*

Proof Because of Lemma 9.19 we have $\mathbb{E}[-X_{T\wedge n}] \leq \mathbb{E}[-X_0]$ for every $n \geq 0$, which implies by virtue of Fatou's lemma $0 \leq \mathbb{E}[-X_T \mathbf{1}_{\{T<\infty\}}] < \infty$.

Let $E \in \mathcal{F}_S$, then $E \cap \{S \leq n\} \in \mathcal{F}_{S\wedge n}$. An application of Lemma 9.19 and non-positivity of X yields

$$\mathbb{E}[X_{S\wedge n} \mathbf{1}_E \mathbf{1}_{\{S\leq n\}}] \leq \mathbb{E}[X_{T\wedge n} \mathbf{1}_E \mathbf{1}_{\{S\leq n\}}] \leq \mathbb{E}[X_{T\wedge n} \mathbf{1}_E \mathbf{1}_{\{T\leq n\}}]$$

and hence

$$\mathbb{E}[X_S \mathbf{1}_E \mathbf{1}_{\{S\leq n\}}] \leq \mathbb{E}[X_T \mathbf{1}_E \mathbf{1}_{\{T\leq n\}}].$$

The Monotone convergence theorem yields $\mathbb{E}[X_S \mathbf{1}_E] \leq \mathbb{E}[X_T \mathbf{1}_E]$. □

Theorem 9.22 *Let X be a submartingale with a last element X_∞, so $X_n \leq \mathbb{E}[X_\infty|\mathcal{F}_n]$ a.s. for every n. Let T and S be stopping times with $S \leq T$. Then X_T and X_S are integrable and*

(i) $X_T \leq \mathbb{E}[X_\infty|\mathcal{F}_T]$ *a.s.*

(ii) $X_S \leq \mathbb{E}[X_T|\mathcal{F}_S]$ *a.s.*

Proof Let $M_n = \mathbb{E}[X_\infty|\mathcal{F}_n]$. By Theorem 9.20, we get $M_S = \mathbb{E}[M_T|\mathcal{F}_S]$. Put then $Y_n = X_n - M_n$. Then Y is a submartingale with $Y_n \leq 0$. From Theorem 9.21 we get $Y_S \leq \mathbb{E}[Y_T|\mathcal{F}_S]$. Since $X_T = M_T + Y_T$ and $X_S = M_S + Y_S$, the result follows. □

9.5 Exercises

9.1 Let $\Omega = \{0, 1\}^{\mathbb{N}}$ and denote by $\omega = (\omega_1, \omega_2, \ldots)$ a typical element of Ω. Let $X_n : \Omega \to \mathbb{R}$ be defined by $X_n(\omega) = \omega_n$ $(n \geq 1)$ and put $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$. Write down the elements of $\mathcal{F}_1$ and $\mathcal{F}_2$ explicitly and describe the elements of $\mathcal{F}_n$ for arbitrary n. How many elements does $\mathcal{F}_n$ have?

9.2 Show that the sequence (v_n) of Example 9.6 is decreasing.

9.3 Let X be a submartingale and f an increasing convex function. Assume that $\mathbb{E}|f(X_n)| < \infty$ for all n. Show that $f(X) = (f(X_n))_{n\geq 0}$ is a submartingale too. If X is a martingale, then $f(X)$ is a submartingale even if f is not increasing, but still convex.

9.4 Let X be an adapted process and T a stopping time that is finite. Show that X_T is $\mathcal{F}$-measurable.

9.5 For every n we have a measurable function f_n on $\mathbb{R}^n$. Let $Z_1, Z_2, \ldots$ be independent random variables and $\mathcal{F}_n = \sigma(Z_1, \ldots, Z_n)$. Show that $X_n = f_n(Z_1, \ldots, Z_n)$ defines a martingale under the conditions that $\mathbb{E}|X_n| < \infty$ and $\mathbb{E}f_n(z_1, \ldots, z_{n-1}, Z_n) = f_{n-1}(z_1, \ldots, z_{n-1})$ for every n.

9.6 If S and T are stopping times, then also $S+T$, $S\vee T$ and $S\wedge T$ are stopping times. Show this.

9.7 Show that an adapted process X is a martingale iff $\mathbb{E}[X_{n+m}|\mathcal{F}_n] = X_n$ for all $n, m \geq 0$.

9.8 Let M be a martingale such that ΔM is a bounded process and Y a bounded predictable process. Let $X = Y \cdot M$. Show that $\mathbb{E}X_T = 0$ if T is a stopping time with finite expectation.

9.9 Let $X_1, X_2, \ldots$ be an *iid* sequence of Bernoulli random variables with probability of success equal to p. Put $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$, $n \geq 1$. Let M be a martingale adapted to this filtration. Show that the *Martingale Representation Property* holds: there exists a constant m and a predictable process Y such that $M_n = m + (Y \cdot S)_n$, $n \geq 1$, where $S_n = \sum_{k=1}^n (X_k - p)$.

9.10 Let $X_1, X_2, \ldots$ be a sequence of independent random variables with $\sigma_n^2 = \mathbb{E}X_n^2 < \infty$ and $\mathbb{E}X_n = 0$ for all n. Let the filtration generated by X and define the martingale M by $M_n = \sum_{k=1}^n X_k$. Determine $\langle M\rangle$.

9.11 Let M be a martingale with $\mathbb{E}M_n^2 < \infty$ for every n. Let C be a bounded predictable process and $X = C \cdot M$. Show that $\mathbb{E}X_n^2 < \infty$ for all n and that $\langle X\rangle = C^2 \cdot \langle M\rangle$.

9.12 Let M be a martingale with $\mathbb{E}M_n^2 < \infty$ for all n and T a stopping time. We know that the stopped process M^T is a martingale too. Show that $\mathbb{E}(M_n^T)^2 < \infty$ for all n and that $\langle M^T\rangle_n = \langle M\rangle_{n\wedge T}$ for every n.

9.13 Let M and N be two square integrable martingales. Show that there exists a unique predictable process $\langle M, N\rangle$ with $\langle M, N\rangle_0 = 0$ such that $MN - \langle M, N\rangle$ is martingale. Show also that for $n \geq 1$

$$\Delta\langle M, N\rangle_n = \mathbb{E}[\Delta M_n \Delta N_n|\mathcal{F}_{n-1}] \text{ a.s.}$$

10 Convergence theorems

The key idea behind the convergence theorems of this chapter is explained first. Consider a sequence of real numbers (x_n). Suppose that $x_n \to x \in \mathbb{R}$ and let (a, b) be any open interval containing x. Then there is $N > 0$ such that $x_n \in (a, b)$ for $n \geq N$. Hence there will be only finitely many fluctuations of the sequence between values smaller than a and values larger than b. This is obviously also true, when $x \notin (a, b)$. Let's see what happens if (x_n) doesn't have a limit. In that case, $\underline{x} = \liminf x_n < \bar{x} = \limsup x_n$. Let $(a, b) \subset (\underline{x}, \bar{x})$. Then there exists a subsequence (x_{n_k}) such that $x_{n_k} > b$ for all k, and a subsequence (x_{m_k}) such that $x_{m_k} < a$. Hence there are infinitely many fluctuations between values below a and values above b. We conclude that convergence of the sequence is equivalent to have only finitely many fluctuations from below a to above b, for any pair of real (even rational) numbers a, b with $a < b$. Something similar can be said if the limit is equal to $\pm\infty$.

Below we use *upcrossings*, these are defined next. Recall that $\inf \emptyset = \infty$.

Definition 10.1 Let (x_n) be a sequence of real numbers and $N \in \mathbb{N}$. Let $a < b$ be real numbers. Put $B_1 = \inf\{n \geq 0 : x_n < a\}$, $S_1 = \inf\{n > B_1 : x_n > b\}$ and then recursively for $n \geq 2$: $B_n = \inf\{k > S_{n-1} : x_k < a\}$, and $S_n = \inf\{k > B_n : x_k > b\}$. Then we define $U_N(a, b) = \max\{n : S_n \leq N\}$ and $U(a, b) = \lim_{N\to\infty} U_N(a, b)$. $U_N(a, b)$ is called the number of upcrossings over the interval (a, b) up to time N, whereas $U(a, b)$ is the total number of upcrossings of the sequence (x_n) over (a, b). An upcrossing is an interval $(B_k, S_k]$ in $\mathbb{N}$. A downcrossing is then an interval $(S_k, B_{k+1}]$.

It follows from the discussion above that a sequence converges, possibly with limits $\pm\infty$, iff $U(a, b) < \infty$ for all $a < b$. If X is a stochastic process, then we can apply the definition of upcrossings to any sequence $(X_n(\omega))$. Then all B_n and S_n will depend on ω, which we then view as mappings $B_n, S_n : \Omega \to \{0, 1, 2, \ldots\} \cup \{\infty\}$. The same holds for the $U_N(a, b)$ and $U(a, b)$. In fact they are all random variables.

10.1 Doob's convergence theorem

The first result is that the random times B_n and S_n are stopping times, if the underlying process X is adapted.

Proposition 10.2 *Let X be an adapted process. Fix $a < b$. Then the B_n and S_n are stopping times. Furthermore, $U_N(a, b)$ is $\mathcal{F}_N$-measurable and $U(a, b)$ is $\mathcal{F}_\infty$-measurable.*

Proof Exercise 10.1. □

We introduce the predictable process Y, defined by $Y_n = \sum_{k=1}^{\infty} \mathbf{1}_{\{B_k < n \leq S_k\}}$. Notice that Y_n takes values in $\{0, 1\}$ and that $\{B_k < n \leq S_k\} \in \mathcal{F}_{n-1}$, for all k and n. Hence $\{Y_n = 1\} \in \mathcal{F}_{n-1}$ as well, which entails that Y is predictable. Let

$Z = Y \cdot X$. If $S_k < \infty$, then $Z_{S_k} - Z_{B_k} > (b-a)$. You may think of Y as a 'buy low, sell high' strategy, if X has the interpretation of a stock price. During an upcrossing your profit will be at least $b - a$.

Lemma 10.3 *It holds that* $Z_N \geq (b-a)U_N(a,b) - (X_N - a)^-$.

Proof We discern two cases. If N belongs to a downcrossing, then we have $Z_N \geq (b-a)U_N(a,b)$, since there are exactly $U_N(a,b)$ upcrossings to the left of N. Note that in this case we have $N \in (S_{U_N}, B_{U_N+1}]$. If N falls in an upcrossing, then we can write $Z_N = Z_{B_{U_N+1}} + (Z_N - Z_{B_{U_N+1}}) = Z_{B_{U_N+1}} + (X_N - X_{B_{U_N+1}}) \geq (b-a)U_N(a,b) + X_N - a \geq (b-a)U_N(a,b) - (X_N - a)^-$. Combining the two cases, we arrive at the assertion. □

Proposition 10.4 *Let X be a supermartingale that is bounded in $\mathcal{L}^1$ (i.e. there is an $M > 0$ such that $\sup_n \mathbb{E}|X_n| < M$). Then for all $a < b$ it holds that $\mathbb{E}U(a,b) < \infty$ and thus $U(a,b) < \infty$ a.s.*

Proof If X is a supermartingale, then so is Z by virtue of Proposition 9.10. It follows that $\mathbb{E}Z_N \leq \mathbb{E}Z_0 = 0$. From Lemma 10.3 we obtain $0 \geq \mathbb{E}Z_N \geq (b-a)\mathbb{E}U_N(a,b) - \mathbb{E}(X_N - a)^-$. Hence $(b-a)\mathbb{E}U_N(a,b) \leq \sup_n \mathbb{E}(X_n - a)^- \leq |a| + \sup_n \mathbb{E}|X_n| \leq |a| + M$. Since the $U_N(a,b)$ form an increasing sequence, the Monotone Convergence Theorem yields $(b-a)\mathbb{E}U(a,b) \leq |a| + M$. □

Here is the first convergence result for supermartingales, Doob's convergence theorem.

Theorem 10.5 *Let X be a supermartingale which is bounded in $\mathcal{L}^1$. Then there exists a random variable $X_\infty \in \mathcal{L}^1(\Omega, \mathcal{F}_\infty, \mathbb{P})$ such that $X_n \overset{\text{a.s.}}{\to} X_\infty$.*

Proof Define $X_\infty := \liminf X_n^+ - \liminf X_n^-$. Then $|X_\infty| \leq \liminf X_n^+ + \liminf X_n^-$ and from Fatou's lemma we deduce that $\mathbb{E}|X_\infty| \leq \liminf \mathbb{E}X_n^+ + \liminf \mathbb{E}X_n^- \leq 2 \liminf \mathbb{E}|X_n|$, which is finite by the assumption that X is bounded in $\mathcal{L}^1$. Note that if (X_n) has an a.s. limit, it must be a.s. equal to X_∞. Hence, the limit -if it exists- has finite expectation and is thus a.s. finite. Let N be the set of ω's such $X_n(\omega)$ doesn't have a limit in $[-\infty, \infty]$. Then $N = \{\liminf X_n < \limsup X_n\}$. We can write $N = \cup_{a<b,\, a,b\in\mathbb{Q}} N_{a,b}$, where $N_{a,b} = \{\liminf X_n < a < b < \limsup X_n\}$. On $N_{a,b}$ it holds that $U(a,b) = \infty$. But the latter event has probability zero, in view of Proposition 10.4. Hence, N, being a countable union of events with probability zero also has probability zero, which concludes the proof. □

Remark 10.6 Notice that the theorem only states that a.s. convergence holds, no other type of convergence, except convergence in probability, necessarily holds true. If X is a martingale, it is attractive to add X_∞ to the sequence X to obtain a process with time index set that includes infinity. It would be very nice that the martingale property as in Remark 9.3 would extend to $m = \infty$, i.e. $\mathbb{E}[X_\infty|\mathcal{F}_n] = X_n$ a.s. But this is not true in general, see Exercise 10.2. In the next section we will give necessary and sufficient conditions under which the extended martingale property holds.

The following corollary can be seen as a stochastic version of the elementary result that every decreasing sequence of real numbers that is bounded from below has a (finite) limit.

Corollary 10.7 *If X is a supermartingale that is bounded from below, then there exists a random variable $X_\infty \in \mathcal{L}^1(\Omega, \mathcal{F}_\infty, \mathbb{P})$ such that $X_n \overset{\text{a.s.}}{\to} X_\infty$.*

Proof Take a constant c such that $Y_n = X_n + c \geq 0$ for all n. Then Y is also a supermartingale and $\mathbb{E}|Y_n| = \mathbb{E}Y_n \leq \mathbb{E}Y_0$. Hence, by virtue of Theorem 10.5, Y admits an a.s. limit and so does X. □

10.2 Uniformly integrable martingales and convergence

Recall the definition of a uniformly integrable collection of random variables $\mathcal{C}$, Definition 7.11, and that a uniformly integrable collection $\mathcal{C}$ is also bounded in $\mathcal{L}^1$, Exercise 7.5. These facts account for half of the proof of

Theorem 10.8 *Let X be a uniformly integrable supermartingale. Then there exists a random variable $X_\infty \in \mathcal{L}^1(\Omega, \mathcal{F}_\infty, \mathbb{P})$ such that $X_n \to X_\infty$ both almost surely and in $\mathcal{L}^1$. Moreover, the extended supermartingale property holds, i.e.*

$$\mathbb{E}[X_\infty|\mathcal{F}_n] \leq X_n \text{ a.s.} \tag{10.1}$$

If X is a martingale then we even have $\mathbb{E}[X_\infty|\mathcal{F}_n] = X_n$.

Proof Let X be a supermartingale. Existence of X_∞ and a.s. convergence follows from Theorem 10.5. Then $\mathcal{L}^1$ convergence holds by virtue of Theorem 7.15. We now show (10.1). Since $\mathbb{E}[X_m|\mathcal{F}_n] \leq X_n$ a.s. when $m > n$ (Remark 9.3), it holds by definition of conditional expectation that $\mathbb{E}\mathbf{1}_G X_m \leq \mathbb{E}\mathbf{1}_G X_n, \forall G \in \mathcal{F}_n$. From this it follows that

$$\begin{aligned}\mathbb{E}\mathbf{1}_G X_\infty - \mathbb{E}\mathbf{1}_G X_n &\leq \mathbb{E}\mathbf{1}_G X_\infty - \mathbb{E}\mathbf{1}_G X_m \\ &\leq \mathbb{E}|\mathbf{1}_G(X_\infty - X_m)| \\ &\leq \mathbb{E}|X_\infty - X_m|,\end{aligned}$$

which tends to zero for $m \to \infty$ by $\mathcal{L}^1$ convergence. We conclude that $\mathbb{E}\mathbf{1}_G X_\infty - \mathbb{E}\mathbf{1}_G X_n \leq 0$. By definition of conditional expectation, also $\mathbb{E}\mathbf{1}_G \mathbb{E}[X_\infty|\mathcal{F}_n] - \mathbb{E}\mathbf{1}_G X_n \leq 0$, or $\mathbb{E}\mathbf{1}_G(\mathbb{E}[X_\infty|\mathcal{F}_n] - X_n) \leq 0$, for all $G \in \mathcal{F}_n$. Take $G = \{\mathbb{E}[X_\infty|\mathcal{F}_n] > X_n\}$ to get $\mathbb{E}\mathbf{1}_G(\mathbb{E}[X_\infty|\mathcal{F}_n] - X_n) = 0$, since the integrand has fixed sign, we obtain $\mathbf{1}_G(\mathbb{E}[X_\infty|\mathcal{F}_n] - X_n) = 0$ a.s. Hence $\mathbb{E}[X_\infty|\mathcal{F}_n] - X_n \overset{\text{a.s.}}{=} \mathbf{1}_{G^c}(\mathbb{E}[X_\infty|\mathcal{F}_n] - X_n) \leq 0$ and the result for supermartingales follows. When X is martingale, one applies the same reasoning to $|\mathbb{E}\mathbf{1}_G X_\infty - \mathbb{E}\mathbf{1}_G X_n|$. □

We conclude that every uniformly integrable martingale is of the form $X_n = \mathbb{E}[X_\infty|\mathcal{F}_n]$, where X_∞ is the a.s. and $\mathcal{L}^1$ limit of the X_n. In the next proposition we present a converse statement.

Proposition 10.9 *Let $\xi \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ and let $\mathbb{G}$ be a family of sub-σ-algebras of $\mathcal{F}$. Write $X_\mathcal{G}$ for any version of $\mathbb{E}[\xi|\mathcal{G}]$. Then the collection $\mathcal{C} = \{X_\mathcal{G} : \mathcal{G} \in \mathbb{G}\}$ is uniformly integrable. In particular, if $\mathbb{G}$ is a filtration $(\mathcal{F}_n)$ and $X_n := \mathbb{E}[\xi|\mathcal{F}_n]$, then the process X is a uniformly integrable martingale.*

Proof Let $\varepsilon > 0$. We have to show the existence of $k > 0$ such that

$$\sup\{\mathbb{E}|X_\mathcal{G}|\mathbf{1}_{\{|X_\mathcal{G}|>k\}} : \mathcal{G} \in \mathbb{G}\} < \varepsilon. \tag{10.2}$$

We exploit the integrability of ξ. Choose δ as in Lemma 7.8 and $k > \mathbb{E}|\xi|/\delta$. By Jensen's inequality for conditional expectations, see Theorem 8.8, we have

$$|X_\mathcal{G}| \leq \mathbb{E}[|\xi|\,|\mathcal{G}] \tag{10.3}$$

and then $\mathbb{E}|X_\mathcal{G}| \leq \mathbb{E}|\xi|$. By Markov's inequality, we obtain

$$\mathbb{P}(|X_\mathcal{G}| > k) \leq \frac{\mathbb{E}|X_\mathcal{G}|}{k} \leq \frac{\mathbb{E}|\xi|}{k} < \delta.$$

Write $G = \{|X_\mathcal{G}| > k\}$. Note that $G \in \mathcal{G}$ and by (10.3) we get

$$\mathbb{E}|X_\mathcal{G}|\mathbf{1}_G \leq \mathbb{E}|\xi|\mathbf{1}_G < \varepsilon,$$

in view of Lemma 7.8. This proves (10.2). For the case, where $\mathbb{G}$ is a filtration, it remains to show that X is a martingale, but we have already done that in Example 9.6. □

In the next theorem we connect the results of Theorem 10.8 and Proposition 10.9. It is known as Lévy's upward convergence theorem.

Theorem 10.10 *Let $\xi \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$, $\mathbb{F}$ a filtration and $X_n = \mathbb{E}[\xi|\mathcal{F}_n]$. Then the a.s. and $\mathcal{L}^1$ limit X_∞ of the X_n is (a version of) $\mathbb{E}[\xi|\mathcal{F}_\infty]$.*

Proof In view of preceding theorems we know of the existence of a limit X_∞, both in the a.s. and $\mathcal{L}^1$ sense. To show that it is a.s. equal to $\mathbb{E}[\xi|\mathcal{F}_\infty]$, we invoke Theorem 1.15. Without loss of generality we suppose that $\xi \geq 0$ a.s., then also $X_\infty \geq 0$ a.s. Since both ξ and X_∞ are integrable, $\nu_\xi(G) := \mathbb{E}\mathbf{1}_G\xi$ and $\nu_\infty(G) := \mathbb{E}\mathbf{1}_G X_\infty$ define finite measures on $\mathcal{F}_\infty$ (Exercise 4.9) with $\nu_\xi(\Omega) = \nu_\infty(\Omega) = \mathbb{E}\xi$. Moreover, they coincide on the algebra (and thus a π-system!) $\cup_{n=1}^\infty \mathcal{F}_n$, since for $G \in \mathcal{F}_n$ one has $\nu_\xi(G) = \mathbb{E}\mathbf{1}_G X_n$ by the definition of X and $\nu_\infty(G) = \mathbb{E}\mathbf{1}_G X_n$ by Theorem 10.8. We conclude that ν_ξ and ν_∞ are the same on $\mathcal{F}_\infty$. By definition of conditional expectation, we then have $\nu_\xi(G) = \mathbb{E}\mathbf{1}_G\mathbb{E}[\xi|\mathcal{F}_\infty]$, $\forall G \in \mathcal{F}_\infty$. Since both $\mathbb{E}[\xi|\mathcal{F}_\infty]$ and X_∞ are $\mathcal{F}_\infty$-measurable, they must be equal. □

Next we need an extension of the concept of filtration. Recall that a filtration is an increasing sequence of σ-algebras, $\mathcal{F}_n \subset \mathcal{F}_{n+1}$. A filtration is thus growing to the right. We extend this in the sense that we now also allow the time index n to be negative. Hence we have $\mathcal{F}_n \subset \mathcal{F}_{n+1}$ for all $n \in \mathbb{Z}$. Hence a filtration is shrinking to the left. Notice that $\mathcal{F}_{-\infty} := \cap_{n<0}\mathcal{F}_n$ is a σ-algebra as well and can

be considered as an infimum. Similarly we extend the notion of a martingale to have negative indexes too. So, a martingale $X = (X_n)_{n\in\mathbb{Z}}$ is a sequence of integrable random variables adapted to a filtration for which the martingale property (9.1) is valid for all $n \in \mathbb{Z}$. Below we need these concepts only for $n < 0$. The next result is known as Lévy's downward theorem.

Theorem 10.11 *Let $\mathbb{F}$ be a filtration on the negative integers and X an $\mathbb{F}$-adapted martingale. Then there is a $\mathcal{F}_{-\infty}$-measurable random variable $X_{-\infty}$ such that $X_n \to X_{-\infty}$ both a.s. and in $\mathcal{L}^1$ as $n \to -\infty$. Moreover $X_{-\infty} = \mathbb{E}[X_{-1}|\mathcal{F}_{-\infty}]$.*

Proof Since $X_n = \mathbb{E}[X_{-1}|\mathcal{F}_n]$ for $n < -1$ we have that X is uniformly integrable in view of Proposition 10.9. Hence $\mathcal{L}^1$ convergence follows from a.s. convergence (Proposition 7.15), the latter to be established now. In the proof of Theorem 10.5, we used the upcrossings inequality of Lemma 10.3. This applies to the present case as well, if we shift the time over a distance of N to the left. Denote the number of upcrossings over (a, b) in a time interval from $-N$ to 0 by $U_{-N}(a, b)$. Taking expectations as in the proof of Theorem 10.5, we obtain $\mathbb{E}U_{-N}(a, b) \le \mathbb{E}(X_{-1}-a)^-/(b-a)$. Hence also $U_{-\infty}(a, b) := \lim_{N\to\infty} U_{-N}(a, b)$ has finite expectation. The rest of the proof of existence of the a.s. limit is as before. The characterization of the limit as a conditional expectation is as in the proof of Theorem 10.8. □

In Section 10.4 we will see a nice application of this theorem.

10.3 $\mathcal{L}^p$ convergence results

In this section our aim is to specialize preceding results to the case where a martingale X is bounded in $\mathcal{L}^p$, i.e. $\sup_n \mathbb{E}|X_n|^p < \infty$, where $p > 1$. We have already seen in Example 7.12 that X is uniformly integrable. By Theorem 10.8 the X_n converge to a limit X_∞ almost surely and in $\mathcal{L}^1$. Our goal is to establish $\mathcal{L}^p$ convergence, meaning $||X_n - X_\infty||_p \to 0$, equivalently $\mathbb{E}|X_n - X_\infty|^p \to 0$. It turns out useful to formulate most results in terms of submartingales. Here is the first one, containing Doob's supremal inequality (10.5).

Proposition 10.12 *Let X be a nonnegative submartingale and write $X_n^* = \max\{X_0, \dots, X_n\}$. For $\lambda > 0$ it holds that*

$$\lambda\mathbb{P}(X_n^* \ge \lambda) \le \mathbb{E}\mathbf{1}_{\{X_n^*\ge\lambda\}}X_n. \tag{10.4}$$

It then follows that

$$\lambda\mathbb{P}(X_n^* \ge \lambda) \le \mathbb{E}X_n. \tag{10.5}$$

Proof Let $T = \inf\{n \ge 0 : X_n \ge \lambda\}$. Then T is a stopping time and $\{X_n^* \ge \lambda\} = \{T \le n\}$. It holds that $E_k := \{T = k\} \subset \{X_k \ge \lambda\}$ and by the submartingale property of X we have $\mathbb{E}\mathbf{1}_{E_k}X_k \le \mathbb{E}\mathbf{1}_{E_k}X_n$, since $E_k \in \mathcal{F}_k$. Then we get $\lambda\mathbb{P}(E_k) = \lambda\mathbb{E}\mathbf{1}_{E_k} \le \mathbb{E}\mathbf{1}_{E_k}X_k \le \mathbb{E}\mathbf{1}_{E_k}X_n$. Equation (10.4) follows by summing over k. □

Remark 10.13 Let $\xi_1, \xi_2, \ldots$ be a sequence of independent random variables with expectation zero and finite second moment. Let $M_n = \sum_{k=1}^n \xi_k$ and $M_n^* = \max\{|M_1|, \ldots, |M_n|\}$. Then $\lambda^2 \mathbb{P}(M_n^* \geq \lambda) \leq \mathbb{E}M_n^2$, as follows by taking $X_n = M_n^2$ in Proposition 10.12. This can be viewed as a 'supremum version' of Chebychev's inequality.

On our path to establishing $\mathcal{L}^p$ convergence we need the following lemma. Observe that (10.4) is of the type $\lambda \mathbb{P}(Y \geq \lambda) \leq \mathbb{E}\mathbf{1}_{\{Y \geq \lambda\}} X$. Such an inequality has a surprising consequence.

Lemma 10.14 *Let X and Y be nonnegative random variables for which the inequality*

$$\lambda \mathbb{P}(Y \geq \lambda) \leq \mathbb{E}\mathbf{1}_{\{Y \geq \lambda\}} X \tag{10.6}$$

holds for all $\lambda > 0$. If $X \in \mathcal{L}^p(\Omega, \mathcal{F}, \mathbb{P})$ for $1 < p \leq \infty$, then also $Y \in \mathcal{L}^p(\Omega, \mathcal{F}, \mathbb{P})$ and moreover, $||Y||_p \leq q||X||_p$, where $q = \frac{p}{p-1}$, as in Hölder's inequality.

Proof We give the proof for $p < \infty$, the case $p = \infty$ is left to the reader. Since (10.6) holds for all $\lambda > 0$ we can integrate both sides after multiplication with $p\lambda^{p-2}$. On the left we get the integral $\int_0^\infty p\lambda^{p-1} \mathbb{P}(Y \geq \lambda)\, \mathrm{d}\lambda = \mathbb{E}Y^p$, in view of Exercise 5.11. On the right we compute using Fubini's theorem (you check that $(\omega, \lambda) \mapsto \lambda^{p-2} \mathbf{1}_{\{Y(\omega) \geq \lambda\}} X(\omega)$ is jointly measurable in ω and λ) the integral

$$\begin{aligned}
\int_0^\infty p\lambda^{p-2} \mathbb{E}\mathbf{1}_{\{Y \geq \lambda\}} X\, \mathrm{d}\lambda &= \mathbb{E} \int_0^\infty p\lambda^{p-2} \mathbf{1}_{\{Y \geq \lambda\}} X\, \mathrm{d}\lambda \\
&= \mathbb{E}(X \int_0^Y p\lambda^{p-2}\, \mathrm{d}\lambda) \\
&= \frac{p}{p-1} \mathbb{E}XY^{p-1}.
\end{aligned}$$

We have established $||Y||_p^p = \mathbb{E}Y^p \leq q\, \mathbb{E}XY^{p-1}$. Hölder's inequality (Theorem 4.46) yields $\mathbb{E}XY^{p-1} \leq (\mathbb{E}X^p)^{1/p} (\mathbb{E}Y^p)^{1/q} = ||X||_p (||Y||_p)^{p/q}$ and we obtain $||Y||_p^p \leq q||X||_p (||Y||_p)^{p/q}$. Would we already know that $||Y||_p < \infty$, then the result follows upon dividing by $(||Y||_p)^{p/q}$. Here is the trick to establish that. First we truncate and consider $Y_n = Y \wedge n$ instead of Y. Certainly $||Y_n||_p < \infty$. Notice that the event $\{Y_n \geq \lambda\}$ is empty for $n < \lambda$ and equal to $\{Y \geq \lambda\}$ otherwise. It follows that (10.6) is valid for Y_n instead of Y. The above reasoning thus yields that $||Y_n||_p \leq q||X||_p$. Since $Y_n \uparrow Y$ we apply the Monotone Convergence Theorem to get the result. □

Here is the result we were aiming at. Inequality (10.7) is known as Doob's $\mathcal{L}^p$-inequality.

Theorem 10.15 *Let X be a nonnegative submartingale that is bounded in $\mathcal{L}^p$ for $1 < p < \infty$. Let $X^* = \sup_n X_n$.*

(i) It holds that $X^* \in \mathcal{L}^p(\Omega, \mathcal{F}_\infty, \mathbb{P})$ *and*

$$||X^*||_p \leq q \sup_n ||X_n||_p, \tag{10.7}$$

where $q = \frac{p}{p-1}$.

(ii) The a.s. limit $X_\infty = \lim_{n\to\infty} X_n$ *exists and moreover,* $X_n \overset{\mathcal{L}^p}{\to} X_\infty$ *and* $||X_\infty||_p = \sup_n ||X_n||_p = \lim_{n\to\infty} ||X_n||_p$.

Proof (i) Put $X_n^* = \max\{X_1, \ldots, X_n\}$. Combining Proposition 10.12 and Lemma 10.14, we get the inequality $||X_n^*||_p \leq q\,||X_n||_p$, with the right hand side less than or equal to $q \sup_n ||X_n||_p$. By monotone convergence ($X_n^* \uparrow X^*$) one obtains the result.

(ii) Existence of the a.s. limit X_∞ is guaranteed by for instance Theorem 10.8. From $|X_n - X_\infty| \leq 2X^*$, we get $|X_n - X_\infty|^p \leq 2^p(|X^*|)_p < \infty$, which has finite expectation. By the Dominated Convergence Theorem we get $X_n \overset{\mathcal{L}^p}{\to} X_\infty$. Finally, the inequalities $||X_\infty||_p \leq ||X_\infty - X_n||_p + \sup_n ||X_n||_p$ and $||X_n||_p \leq ||X_n - X_\infty||_p + ||X_\infty||_p$ together with $||X_n||_p$ increasing in n yield the last assertion. □

Corollary 10.16 *Let* M *be a martingale that is bounded in* $\mathcal{L}^p$ *for* $p > 1$. *The a.s. limit* $M_\infty = \lim_{n\to\infty} M_n$ *exists and moreover,* $M_n \overset{\mathcal{L}^p}{\to} M_\infty$.

Proof Again existence of M_∞ as an a.s. limit can be deduced from e.g. Theorem 10.8. Moreover, $|M_n - M_\infty| \leq 2X^*$, where $X^* = \sup_n |M_n|$. An application of the previous theorem to the nonnegative submartingale $X = |M|$ shows that $||X^*||_p < \infty$. The rest of the proof is as in the proof of Theorem 10.15. □

10.4 The strong law of large numbers

In this section we present versions of a strong law of large numbers for martingales and independent sequences of random variables. We start with a simple analytic lemma.

Lemma 10.17

(i) Let $(w_{nk})_{n,k\geq 1}$ *be a double array of nonnegative real numbers satisfying* $\lim_{n\to\infty} w_{nk} = 0$ *for every* $k \geq 1$ *and* $\lim_{n\to\infty} \sum_{k=1}^n w_{nk} = 1$. *Let* (x_n) *be sequence of real numbers with* $\lim_{n\to\infty} x_n = x \in \mathbb{R}$. *Put* $\bar{x}_n = \sum_{k=1}^n w_{kn} x_k$. *Then* $\lim_{n\to\infty} \bar{x}_n = x$.

(ii) Let $(w_n)_{n\geq 0}$ *be an increasing sequence of nonnegative real numbers with* $w_0 = 0$ *and* $w_n \to \infty$. *Let* (x_n) *be a sequence of real numbers for which the series* $\sum_{n=1}^\infty \frac{x_n}{w_n}$ *is convergent. Then* $\frac{1}{w_n} \sum_{k=1}^n x_k \to 0$.

Proof (i) Let $\varepsilon > 0$ and choose m such that $|x_n - x| < \varepsilon$ for $n > m$. Then we

have for $n > m$

$$\begin{aligned}|\bar{x}_n - x| &\le \sum_{k=1}^{n} w_{kn}|x_k - x| + |\sum_{k=1}^{n} w_{kn} - 1||x| \\ &\le \sum_{k=1}^{m} w_{kn}|x_k - x| + \sum_{k=m+1}^{n} w_{kn}|x_k - x| + |\sum_{k=1}^{n} w_{kn} - 1||x| \\ &\le \sum_{k=1}^{m} w_{kn}|x_k - x| + \varepsilon \sum_{k=m+1}^{n} w_{kn} + |\sum_{k=1}^{n} w_{kn} - 1||x|.\end{aligned}$$

It follows from the assumptions that $\limsup_{n\to\infty} |\bar{x}_n - x| \le \varepsilon$. Since $\varepsilon > 0$ is arbitrary the result follows.

(ii) Given $n \ge 1$, define $w_{kn} = \frac{w_k - w_{k-1}}{w_n}$, for $k = 1, \dots, n$ and $y_n = \sum_{k=n+1}^{\infty} \frac{x_k}{w_k}$. Note that $y_n \to 0$. We compute

$$\begin{aligned}\frac{1}{w_n}\sum_{k=1}^{n} x_k &= -\frac{1}{w_n}\sum_{k=1}^{n} w_k(y_k - y_{k-1}) \\ &= \frac{1}{w_n}\big(- w_n y_n + \sum_{k=1}^{n} y_k(w_k - w_{k-1})\big) \\ &= -y_n + \sum_{k=1}^{n} w_{kn} y_k.\end{aligned}$$

Application of part (i) yields the result. □

Remark 10.18 The first part of the previous proposition is known as Toeplitz' lemma, and the second part is known as Kronecker's lemma. There are many variations on the assertions of this proposition known. The special case $w_{kn} = \frac{1}{n}$ for $1 \le k \le n$ yields for the first part Cesaro's lemma.

Proposition 10.19 *Let M be a square integrable martingale, $\mathbb{E}M_n^2 < \infty$ for all n. Let $\langle M\rangle$ be its predictable variation process. Then $\frac{1}{\langle M\rangle_n} M_n \to 0$ a.s. on the set $\{\langle M\rangle_n \to \infty\}$.*

Proof Let X be the martingale transform $X = \frac{1}{1+\langle M\rangle} \cdot M$. Then

$$\begin{aligned}\mathbb{E}[\Delta X_n^2|\mathcal{F}_{n-1}] &= \frac{1}{(1+\langle M\rangle)^2}\mathbb{E}[\Delta M_n^2|\mathcal{F}_{n-1}] \\ &= \frac{1}{(1+\langle M\rangle)^2}\Delta\langle M\rangle_n \\ &\le \frac{1}{1+\langle M\rangle_{n-1}} - \frac{1}{1+\langle M\rangle_n}.\end{aligned}$$

It follows that $\Delta\langle X\rangle_n \le \frac{1}{1+\langle M\rangle_{n-1}} - \frac{1}{1+\langle M\rangle_n}$ and hence that $\langle X\rangle_n \le 1$ for all n. Therefore $\sup_n \mathbb{E}X_n^2 < \infty$ (see Remark 9.18) and X converges a.s. in view of Theorem 10.5 (and also in $\mathcal{L}^2$ by virtue of Theorem 10.15). An application of Kronecker's lemma yields the assertion. □

Corollary 10.20 *Let $X_1, X_2, \ldots$ be an independent sequence with $\mathbb{E}X_k = 0$ and $\sigma_k^2 := \operatorname{Var} X_k < \infty$ for all k. Let $\alpha_n = \sum_{k=1}^n \sigma_k^2$. If $\alpha_n \to \infty$, then $\frac{1}{\alpha_n}\sum_{k=1}^n X_k \to 0$ a.s.*

Proof Let M be the martingale (the filtration should be obvious) defined by $M_n = \sum_{k=1}^n X_k$. It follows that $\langle M\rangle_n = \alpha_n$ and the assertion immediately follows by an application of Proposition 10.19. □

The assertion of Corollary 10.20 is the strong law for a sequence of independent random variables with a finite second moment. If the sequence is moreover *iid*, then we get $\frac{1}{n}\sum_{k=1}^n X_k \to 0$, the usual strong law of large numbers. The assumption that an *iid* sequence has finite second moments can be dropped, whereas the strong law still holds. This is the content of Theorem 10.23 below, whose proof is based on completely different arguments.

We introduce some terminology. Let $X_1, X_2, \ldots$ be a sequence of random variables. Define $\mathcal{T}_n = \sigma(X_{n+1}, X_{n+2}, \ldots)$ and $\mathcal{T} = \cap_{n=0}^\infty \mathcal{T}_n$, $\mathcal{T}$ is called the *tail σ-algebra* of the sequence. The following proposition is known as Kolmogorov's 0-1 law. It tells us that the tail σ-algebra of an independent sequence is a.s. trivial. Its proof is an easy consequence of Lévy's theorem.

Proposition 10.21 *Let $F \in \mathcal{T}$, the tail σ-algebra of an independent sequence. Then $\mathbb{P}(F) \in \{0, 1\}$.*

Proof Put $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$. Let $F \in \mathcal{T}$ and observe the triviality $F \in \mathcal{F}_\infty$. Put $\xi = \mathbf{1}_F$, then $X_n = \mathbb{E}[\xi|\mathcal{F}_n]$ defines a martingale whose a.s. limit (see Theorem 10.10) is equal to $\mathbb{E}[\xi|\mathcal{F}_\infty] = \xi$ a.s. It is easy to show (Exercise 3.10) that $\mathcal{F}_n$ and $\mathcal{T}_n$ are independent and therefore $\mathcal{F}_n$ and $\sigma(\xi)$ are independent. But then $X_n = \mathbb{E}\xi = \mathbb{P}(F)$ a.s. We conclude that $\mathbf{1}_F$ is a.s. equal to the constant $\mathbb{P}(F)$, and then it has expectation $\mathbb{P}(F)$ and variance zero. So $0 = \mathbb{E}\mathbf{1}_F^2 - (\mathbb{E}\mathbf{1}_F)^2 = \mathbb{P}(F) - \mathbb{P}(F)^2$. The result follows. □

Example 10.22 Here is a first application. Let $X_1, X_2, \ldots$ be an *iid* sequence. Put $\bar{X}_n = \frac{1}{n}\sum_{k=1}^n X_k$ and $\bar{X} = \limsup \bar{X}_n$. Then for every $m \geq 1$, it holds that $\bar{X} = \limsup \frac{1}{n}\sum_{k=m+1}^n X_k$, which belongs to $\mathcal{T}_m$. It follows that $\bar{X} \in \mathcal{T}$ and hence that $\bar{X}$ is a.s. equal to a constant in view of Exercise 3.5. The same holds for the $\liminf$ and then also for the limit of the averages, if it exists.

All the heavy machinery that we have developed so far now pays off by having a relatively simple proof of the Strong Law of Large Numbers for *iid* sequences.

Theorem 10.23 *Let $X_1, X_2, \ldots$ be an* iid *sequence in $\mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$. Let $\mu = \mathbb{E}X_1$ and $\bar{X}_n = \frac{1}{n}\sum_{k=1}^n X_k$. Then $\bar{X}_n \to \mu$ a.s. and in $\mathcal{L}^1$ as $n \to \infty$.*

Proof We'd like to apply Theorem 10.11. The first thing we need is a filtration that is defined on the negative integers. The following choice turns out be a clever one. Let $S_n = \sum_{k=1}^n X_k$ and put $\mathcal{F}_{-n} = \sigma(S_n, S_{n+1}, \ldots)$, $n \geq 1$. Put $M_{-n} = \bar{X}_n$. In Example 8.9 we have seen that $\mathbb{E}[X_1|\mathcal{F}_{-n}] = M_{-n}$. It follows

from Theorem 10.11 that there exists $M_{-\infty}$ such that $M_{-n} \to M_{-\infty}$ both a.s. and in $\mathcal{L}^1$. We proceed by identifying the limit. From Example 10.22, we know that $M_{-\infty}$ has to be equal to a constant a.s. But Theorem 10.11 also tells us that $\mathbb{E}M_{-\infty} = \mathbb{E}M_{-1}$, which is equal to μ. Hence $M_{-\infty} = \mu$ a.s. □

10.5 Exercises

10.1 Prove Proposition 10.2. Show also that the process Y below that proposition is predictable.

10.2 Consider the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with $\Omega = [0,1)$, $\mathcal{F}$ the Borel sets of $[0,1)$ and $\mathbb{P}$ the Lebesgue measure. Let $I_k^n = [k2^{-n}, (k+1)2^{-n})$ for $k = 0, \ldots, 2^n - 1$ and $\mathcal{F}_n$ be the σ-algebra by the I_k^n for $k = 0, \ldots, 2^n - 1$. Define $X_n = \mathbf{1}_{I_0^n} 2^n$. Show that X_n is a martingale and that the conditions of Theorem 10.5 are satisfied. What is X_∞ in this case? Do we have $X_n \to X_\infty$ in $\mathcal{L}^1$?

10.3 Let X be a submartingale with $\sup_{n\geq 0} \mathbb{E}|X_n| < \infty$. Show that there exists a random variable X_∞ such that $X_n \to X_\infty$ a.s.

10.4 Show that for a supermartingale X the condition $\sup\{\mathbb{E}|X_n| : n \in \mathbb{N}\} < \infty$ is equivalent to the condition $\sup\{\mathbb{E}X_n^- : n \in \mathbb{N}\} < \infty$.

10.5 Let $Y \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ and a filtration $(\mathcal{F}_n)$ be given. Define for all $n \in \mathbb{N}$ the random variable $X_n = \mathbb{E}[Y|\mathcal{F}_n]$. We know that there is X_∞ such that $X_n \to X_\infty$ a.s. Show that for $Y \in \mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$, we have $X_n \overset{\mathcal{L}^2}{\to} X_\infty$. Find a condition such that $X_\infty = Y$. Give also an example in which $\mathbb{P}(X_\infty = Y) = 0$.

10.6 Let $X = (X_n)_{n\leq 0}$ a (backward) supermartingale.

(a) Show equivalence of the next two properties:
(i) $\sup_n \mathbb{E}|X_n| < \infty$ and (ii) $\lim_{n\to-\infty} \mathbb{E}X_n < \infty$.
(Use that $x \mapsto x^+$ is convex and increasing.)

(b) Under the condition $\sup_n \mathbb{E}|X_n| =: A < \infty$ the supermartingale X is uniformly integrable. To show this, you may proceed as follows (*but other solutions are equally welcome).* Let $\varepsilon > 0$ and choose $K \in \mathbb{Z}$ such that for all $n < K$ one has $0 \leq \mathbb{E}X_n - \mathbb{E}X_K < \varepsilon$. It is then sufficient to show that $(X_n)_{n\leq K}$ is uniformly integrable. Let $c > 0$ be arbitrary and $F_n = \{|X_n| > c\}$. Using the supermartingale inequality you show that

$$\int_{F_n} |X_n| \, d\mathbb{P} \leq \int_{F_n} |X_K| \, d\mathbb{P} + \varepsilon.$$

Because $\mathbb{P}(F_n) \leq \frac{A}{c}$ you conclude the proof.

10.7 Suppose that $\mathbb{Q}$ is a probability measure on $(\Omega, \mathcal{F})$ such that $\mathbb{Q} \ll \mathbb{P}$ with $\mathrm{d}\mathbb{Q}/\,\mathrm{d}\mathbb{P} = Z$. Denote by $\mathbb{P}_n$ and $\mathbb{Q}_n$ the restrictions of $\mathbb{P}$ and $\mathbb{Q}$ to $\mathcal{F}_n$ $(n \geq 1)$. Show that $\mathbb{Q}_n \ll \mathbb{P}_n$ and that

$$\frac{\mathrm{d}\mathbb{Q}_n}{\mathrm{d}\mathbb{P}_n} = M_n,$$

where $M_n = \mathbb{E}_{\mathbb{P}}[Z|\mathcal{F}_n]$.

10.8 Let M be a nonnegative martingale with $\mathbb{E}M_n = 1$ for all n. Define $\mathbb{Q}_n(F) = \mathbb{E}\mathbf{1}_F M_n$ for $F \in \mathcal{F}_n$ $(n \geq 1)$. Show that for all $n \geq 1$ and $k \geq 1$ one has $\mathbb{Q}_{n+k}(F) = \mathbb{Q}_n(F)$ for $F \in \mathcal{F}_n$. Assume that M is uniformly integrable. Show that there exists a probability measure $\mathbb{Q}$ on $\mathcal{F}_\infty = \sigma(\bigcup_n \mathcal{F}_n)$ that is absolutely continuous w.r.t. $\mathbb{P}$ and that is such that for all n the restriction of $\mathbb{Q}$ to $\mathcal{F}_n$ coincides with $\mathbb{Q}_n$. Characterize $\mathrm{d}\mathbb{Q}/\,\mathrm{d}\mathbb{P}$.

10.9 Consider Theorem 10.15. Show that $||X_n||_p$ is increasing in n.

10.10 Let (X_n) be a sequence of random variables with finite a.s. limit X. Assume the existence of a random variable $Y \geq 0$ with $\mathbb{E}Y < \infty$ such that for all n it holds that $|X_n| \leq Y$. Let $(\mathcal{F}_n)$ be an *arbitrary* filtration. *Hunt's lemma* states

$$\mathbb{E}[X_n|\mathcal{F}_n] \overset{\text{a.s.}}{\to} \mathbb{E}[X|\mathcal{F}_\infty].$$

(a) Put $Z_m = \sup_{k \geq m} |X_k - X|$. Show that Z_m converges to zero, both in $\mathcal{L}^1$ and a.s.

(b) Show also that for $n \geq m$:

$$|\mathbb{E}[X_n|\mathcal{F}_n] - \mathbb{E}[X|\mathcal{F}_\infty]| \leq |\mathbb{E}[X|\mathcal{F}_n] - \mathbb{E}[X|\mathcal{F}_\infty]| + |\mathbb{E}[Z_m|\mathcal{F}_n]|$$

(c) Finish the proof of Hunt's lemma.

10.11 Let M be a martingale with $M_0 = 0$ and assume the existence of constants c_k such that for all k it holds that $|M_k - M_{k-1}| \leq c_k$. Let $x > 0$. The *Azuma-Hoeffding inequality* is

$$\mathbb{P}(\sup_{k \leq n} M_k \geq x) \leq \exp(-\frac{x^2}{2\sum_{i=1}^n c_k^2}).$$

Prove this inequality by the following two steps.

(a) Let $c > 0$ and Y a random variable with $\mathbb{P}(|Y| \leq c) = 1$ and $\mathbb{E}\,Y = 0$. Let $f(y) = e^{\theta y}$. Use convexity of f to obtain for $y \in [-c, c]$

$$f(y) \leq \frac{c-y}{2c} f(-c) + \frac{c+y}{2c} f(c).$$

Show that $\mathbb{E}f(Y) \leq \cosh \theta c \leq \exp(\frac{1}{2}\theta^2 c^2)$.

(b) Show that $\mathbb{E}Z_n \leq \exp(\frac{1}{2}\theta^2 \sum_{k=1}^n c_k^2)$, where $Z_n = \exp(\theta M_n)$.

(c) Give a bound on the probability in Hoeffding's inequality in terms of $\mathbb{E}Z_n$ and minimize over $\theta > 0$ to finish the proof.

10.12 Let $Z_1, Z_2, \ldots$ be independent nonnegative random variables defined on some $(\Omega, \mathcal{F}, \mathbb{P})$ with $\mathbb{E}Z_n = 1$ for all n. The process M defined by $M_n = \prod_{i=1}^n Z_i$ is a nonnegative martingale. We know that M_∞ exist as an almost sure limit of the M_n.

(a) Let $R_n = Z_n^{\frac{1}{2}}$. Show that $r_n := \mathbb{E}R_n \leq 1$.
(b) Let N be the martingale defined by $N_n = \prod_{i=1}^n \frac{R_i}{r_i}$. Assume that

$$\prod_{k=1}^{\infty} r_k > 0. \tag{10.8}$$

Show that N is bounded in $\mathcal{L}^2$ and that consequently M is uniformly integrable.
(c) Show that (10.8) implies that $\mathbb{E}M_\infty = 1$.
(d) Show that $\mathbb{E}M_\infty = 1$ implies that (10.8) holds. *Hint:* Reason by contradiction.

10.13 Let $X_1, X_2, \ldots$ be real valued functions defined on some Ω and take $\mathcal{F} = \sigma(X_1, X_2, \ldots)$. Assume there exist probability measures $\mathbb{P}$ and $\mathbb{Q}$ on $(\Omega, \mathcal{F})$, such that the X_n are *iid* under both $\mathbb{P}$ and $\mathbb{Q}$. Assume that X_1 admits strictly positive densities f and g under $\mathbb{P}$ and $\mathbb{Q}$ respectively. Let $Z_n = \frac{g(X_n)}{f(X_n)}$ and $M_n = \prod_{k=1}^n Z_k$. Use Exercise 10.12 to show that either $\mathbb{P} = \mathbb{Q}$ or $\mathbb{P} \perp \mathbb{Q}$ (in this case r_k is the same for all k). This phenomenon is known as *Kakutani's dichotomy*, or *Kakutani's alternatives*.

10.14 Consider the setup of Exercise 10.13. Assume that

$$\prod_{k=1}^{n} \mathbb{E}_{\mathbb{P}} \sqrt{\frac{g(X_k)}{f(X_k)}} \to 0.$$

Suppose one observes $X_1, \ldots, X_n$. Consider the testing problem H_0: the densities of the X_k are the f_k against H_1: the density of X_k is f and the test that rejects H_0 if $M_n > c_n$, where $\mathbb{P}(M_n > c_n) = \alpha \in (0, 1)$ (likelihood ratio test). Show that this test is *consistent*: $\mathbb{Q}(M_n \leq c_n) \to 0$. (Side remark: the content of the Neyman-Pearson lemma is that this test is most powerful among all test with significance level less than or equal to α.)

10.15 Let (H_n) be a predictable sequence of random variables with $\mathbb{E}H_n^2 < \infty$ for all n. Let (ε_n) be a sequence with $\mathbb{E}\varepsilon_n^2 = 1$, $\mathbb{E}\varepsilon_n = 0$ and ε_n independent of $\mathcal{F}_{n-1}$ for all n. Let $M_n = \sum_{k \leq n} H_k \varepsilon_k$, $n \geq 0$. Compute the conditional variance process A of M. Take $p > 1/2$ and consider $N_n = \sum_{k \leq n} \frac{1}{(1+A_k)^p} H_k \varepsilon_k$. Show that there exists a random variable N_∞ such that $N_n \to N_\infty$ a.s. Show (use Kronecker's lemma) that $\frac{M_n}{(1+A_n)^p}$ has an a.s. finite limit.

11 Local martingales and Girsanov's theorem

The purpose of this section is to formulate and prove Girsanov's theorem on absolutely continuous changes of measure for discrete time processes.

11.1 Local martingales

As always we have a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ on which all random variables are defined. Before defining local martingales, we relax the notion of conditional expectation of a random variable X by dropping the requirement that $\mathbb{E}|X|$ is finite. Suppose $X \geq 0$ and that $\mathcal{G}$ is a sub-σ-algebra of $\mathcal{F}$. Then there exists a $\mathcal{G}$-measurable random variable $\hat{X}$ such that $\mathbb{E}X\mathbf{1}_G = \mathbb{E}\hat{X}\mathbf{1}_G \leq \infty$ for all $G \in \mathcal{G}$ (Exercise 11.1). We adopt again the notation $\hat{X} = \mathbb{E}[X|\mathcal{G}]$ for any version $\hat{X}$. If X is such that $\mathbb{E}[X^+|\mathcal{G}] < \infty$ a.s. or $\mathbb{E}[X^-|\mathcal{G}] < \infty$ a.s., we define $\hat{X} = \mathbb{E}[X^+|\mathcal{G}] - \mathbb{E}[X^-|\mathcal{G}]$, the *generalized conditional expectation* of X given $\mathcal{G}$, denoted (again) $\mathbb{E}[X|\mathcal{G}]$. Note that also here we have different versions, and that these are a.s. equal.

In what follows we work with a fixed filtration $\mathbb{F}$. Without explicit reference to it further, one should understand that an adapted (predictable) process is adapted (predictable) w.r.t. $\mathbb{F}$, etc.

Definition 11.1 Let X be an adapted process. If there exists a sequence of stopping times T^n such that $T^n \to \infty$ and such that the stopped processes X^{T^n} are all martingales, we call the process X a *local* martingale. The sequence (T^n) is called a fundamental or localizing sequence.

It is obvious that any martingale is a local martingale, but the converse is not true, see Exercise 11.2. Note also that the definition implies that $X_0 = X_0^{T^n}$ belongs to $\mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$. There exist alternative definitions for local martingales that also allow for non-integrable X_0, but we don't consider these.

Proposition 11.2 *A real valued adapted process X is a local martingale iff the following two assertions are hold true. (i) $X_0 \in \mathcal{L}^1(\Omega, \mathcal{F}_0, \mathbb{P})$ and (ii) $\mathbb{E}[|X_{n+1}|\,|\mathcal{F}_n] < \infty$ a.s. and $\mathbb{E}[X_{n+1}|\mathcal{F}_n] = X_n$ for all n in the sense of generalized conditional expectations.*

Proof Suppose X is a local martingale and (T^n) a localizing sequence. Then (i) immediately follows and we also have $\mathbb{E}[X_k^{T^n}|\mathcal{F}_{k-1}] = X_{k-1}^{T^n}$, which yields $\mathbf{1}_{\{T^n > k-1\}}\mathbb{E}[X_k|\mathcal{F}_{k-1}] = X_{k-1}\mathbf{1}_{\{T^n > k-1\}}$. Letting $n \to \infty$ gives $\mathbb{E}[X_k|\mathcal{F}_{k-1}] = X_{k-1}$. From this generalized martingale property we see that the conditional expectation $\mathbb{E}[X_k|\mathcal{F}_{k-1}]$ is finite a.s. As a consequence, one automatically has $\mathbb{E}[|X_k|\,|\mathcal{F}_{k-1}] < \infty$ a.s.

Conversely we assume (i) and (ii). Let $T^n = \inf\{m : \sum_{k=1}^{m+1} \mathbb{E}[|X_k|\,|\mathcal{F}_{k-1}] \geq n\}$. Then T^n is a stopping time and $\mathbb{E}|X_k^{T_n}| \leq \mathbb{E}|X_0| + n < \infty$, so $X_k^{T_n} \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$. Since $\{T^n \geq k\} \in \mathcal{F}_{k-1}$ we also have from (ii) with k instead of $n+1$ that $\mathbb{E}[X_k^{T^n}|\mathcal{F}_{k-1}] = X_{k-1}^{T^n}$, which yields the martingale property of X^{T^n}. □

Proposition 11.3 *The following results on martingale transforms hold true.*

(i) If Y is a predictable process and M a martingale, then $X = Y \cdot M$ is a local martingale.

(ii) If X is a local martingale, then there exists a predictable process Y and a martingale M such that $X = X_0 + Y \cdot M$.

(iii) If Y is a predictable process and M a local martingale, then $Y \cdot M$ is a local martingale.

Proof (i) Let $T^n = \inf\{k \geq 0 : |Y_{k+1}| > n\}$. Then (T^n) is a fundamental sequence of stopping times and Y^{T^n} is a bounded process for every n. Moreover, the stopped processes M^{T^n} are martingales as well, and one easily verifies $X^{T^n} = Y^{T^n} \cdot M^{T^n}$. It follows from Proposition 9.13 that X^{T^n} is a martingale for every n and so X is a local martingale.

(ii) Let $A_k^n = \{\mathbb{E}[|X_n| \,|\mathcal{F}_{n-1}] \in [k, k+1)\}$, $k = 0, 1, \ldots$. For every n, the A_k^n are disjoint and by Proposition 11.2 we have $\mathbb{P}(\bigcup_k A_k^n = \Omega) = 1$. Put $u_n := \sum_{k\geq 0}(k+1)^{-3}\Delta X_n \mathbf{1}_{A_k^n}$. Then $u_n \in \mathcal{F}_n$, $\mathbb{E}|u_n| < \infty$ and $\mathbb{E}[u_n|\mathcal{F}_{n-1}] = 0$. Thus (M_n) with $M_n = \sum_{i=1}^n u_k$ is a martingale with $M_0 = 0$. Put then $Y_n = \sum_{k\geq 0}(k+1)^3 \mathbf{1}_{A_k^n}$ and we see that $\Delta X_n = Y_n \Delta M_n$.

(iii) From (ii) it follows that, M being a local martingale, must be of the form $M = M_0 + Y' \cdot M'$ for a predictable process Y' and a martingale M'. This implies $X = Y \cdot (Y' \cdot M') = (YY') \cdot M'$ and the result follows from (i). □

Proposition 11.4 *If X is a nonnegative local martingale with $\mathbb{E}X_0 < \infty$, then it is a true martingale.*

Proof We first show that $\mathbb{E}X_k < \infty$ for all k. Let (T^n) be a localizing sequence. By Fatou's lemma and by the martingale property of X^{T^n} we have

$$\mathbb{E}X_k \leq \liminf_{n\to\infty} \mathbb{E}X_k^{T^n} = \mathbb{E}X_0.$$

Furthermore, $X_k^{T^n} \leq \sum_{i=0}^k X_i$, of which we now know that the sum in the upper bound has finite expectation. Hence we can apply the Dominated Convergence Theorem for conditional expectations to $\mathbb{E}[X_k^{T^n}|\mathcal{F}_{k-1}] = X_{k-1}^{T^n}$ to get $\mathbb{E}[X_k|\mathcal{F}_{k-1}] = X_{k-1}$. □

An adapted sequence (ΔM_n), with $n \geq 1$, is called a local martingale difference sequence if $\mathbb{E}[\Delta M_n|\mathcal{F}_{n-1}] = 0$ (and, although redundant, $\mathbb{E}[|\Delta M_n| \,|\mathcal{F}_{n-1}] < \infty$) for all $n \geq 1$.

Proposition 11.5 *Let M be an adapted process and put $\Delta M_n = M_n - M_{n-1}$, $n \geq 1$. Equivalent are*

(i) M is a local martingale,

(ii) $\mathbb{E}|M_0| < \infty$ and (ΔM_n) is a local martingale difference sequence.

Proof Exercise 11.5 □

11.2 Quadratic variation

If X and Y are two stochastic processes, their *optional* quadratic covariation process $[X, Y]$ is defined as

$$[X, Y]_n = \sum_{k=1}^{n} \Delta X_k \Delta Y_k.$$

For $X = Y$ we write $[X]$ instead of $[X, Y]$. Note the integration by parts formula

$$X_n Y_n = X_0 Y_0 + (X_- \cdot Y)_n + (Y_- \cdot X)_n + [X, Y]_n,$$

where the (predictable) process X_- at time n equals X_{n-1}, defined for $n \geq 1$.

If X and Y are square integrable martingales, we have encountered in Section 9.3 their *predictable* covariation process $\langle X, Y \rangle$, which was the unique predictable process such that $XY - \langle X, Y \rangle$ is a martingale too. It follows from the above integration by parts formula, recall that $(X_- \cdot Y)$ and $(Y_- \cdot X)$ are martingales, that also $[X, Y] - \langle X, Y \rangle$ is a martingale. In fact, $\langle X, Y \rangle$ is also the unique predictable process that, subtracted from $[X, Y]$, yields a martingale, already known from Corollary 9.17,

$$\mathbb{E}[\Delta X_n \Delta Y_n | \mathcal{F}_{n-1}] = \Delta \langle X, Y \rangle_n. \tag{11.9}$$

This property carries over to the case where X and Y are local martingales under the extra conditions $\mathbb{E}[\Delta X_n^2 | \mathcal{F}_{n-1}] < \infty$ and $\mathbb{E}[\Delta Y_n^2 | \mathcal{F}_{n-1}] < \infty$, Exercise 11.3, and under this condition, $XY - \langle X, Y \rangle$ is a local martingale.

11.3 Measure transformation

In this section we suppose that $\mathbb{P}$ and $\mathbb{Q}$ are probability measures on $(\Omega, \mathcal{F})$. Their restrictions to $\mathcal{F}_n$ are denoted $\mathbb{P}_n$ and $\mathbb{Q}_n$ respectively. We will always assume that $\mathbb{P}_0 = \mathbb{Q}_0$, which is automatically the case if $\mathcal{F}_0$ is trivial. We start with a definition whose contents are supposed to be in force throughout this section. See Exercise 10.7 for the more restrictive setting with $\mathbb{Q} \ll \mathbb{P}$.

Definition 11.6 The probability measure $\mathbb{Q}$ is called locally absolutely continuous w.r.t. the probability measure $\mathbb{P}$ if $\mathbb{Q}_n \ll \mathbb{P}_n$ for all $n \geq 0$. In this case we can define the $\mathcal{F}_n$-measurable random variables (Radon-Nikodym derivatives, densities) $Z_n = \frac{\mathrm{d}\mathbb{Q}_n}{\mathrm{d}\mathbb{P}_n}$. The process $Z = (Z_n)_{n \geq 0}$ is called the *density process*. If $\mathbb{P}_n \sim \mathbb{Q}_n$ for all n, $\mathbb{P}$ and $\mathbb{Q}$ are called locally equivalent.

One easily verifies that Z is a martingale under the probability measure $\mathbb{P}$. Note that $\mathbb{Q}(Z_n > 0) = 1$, but $\mathbb{P}(Z_n > 0)$ may be less than 1, unless $\mathbb{P}_n \sim \mathbb{Q}_n$, in which case $\mathbb{P}(F) = \mathbb{E}_{\mathbb{Q}} \frac{1}{Z_n} \mathbf{1}_F$ for $F \in \mathcal{F}_n$. Furthermore, on the set $\{Z_n = 0\}$ also $Z_{n-1} = 0$, as follows from the martingale property of the nonnegative process Z. Hence we can define Z_n / Z_{n-1} with the understanding that it is put equal to zero on $\{Z_{n-1} = 0\}$.

If M is martingale under $\mathbb{P}$, it is usually not a martingale under $\mathbb{Q}$. We shall see below how to transform M parallel to the measure change from $\mathbb{P}$ to $\mathbb{Q}$ such that the changed process is a martingale again, or a local martingale in a more general setting. The following lemma generalizes Equation (6.11) to Equation (11.10) for conditional expectations.

Lemma 11.7 *Let $\mathbb{Q} \ll \mathbb{P}$ with Radon-Nikodym derivative $Z := \frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}$, $\mathcal{G}$ a sub-σ-algebra of $\mathcal{F}$. Let $\hat{\mathbb{P}}$ and $\hat{\mathbb{Q}}$ be the restrictions of $\mathbb{P}$ and $\mathbb{Q}$ to $\mathcal{G}$ and let $\hat{Z} = \mathbb{E}[Z|\mathcal{G}]$.*

(i) It holds that $\hat{\mathbb{Q}} \ll \hat{\mathbb{P}}$, $\frac{\mathrm{d}\hat{\mathbb{Q}}}{\mathrm{d}\hat{\mathbb{P}}} = \hat{Z}$, and for $Y \in \mathcal{L}^1(\Omega, \mathcal{G}, \mathbb{Q})$ one has $\mathbb{E}_{\mathbb{Q}} Y = \mathbb{E}[Y\hat{Z}]$.

(ii) Suppose $X \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{Q})$. Then $\mathbb{Q}(\hat{Z} > 0) = 1$ and a.s. under $\mathbb{Q}$ one has

$$\mathbb{E}_{\mathbb{Q}}[X|\mathcal{G}] = \frac{\mathbb{E}[XZ|\mathcal{G}]}{\hat{Z}} = \mathbb{E}[X\tilde{Z}|\mathcal{G}], \tag{11.10}$$

where $\tilde{Z} := \frac{Z}{\hat{Z}}$.

Proof (i) Absolute continuity is trivial and the other statements follow from $\mathbb{E}_{\mathbb{Q}} Y = \mathbb{E}[YZ] = \mathbb{E}\mathbb{E}[YZ|\mathcal{G}] = \mathbb{E}[Y\mathbb{E}[Z|\mathcal{G}]] = \mathbb{E}[Y\hat{Z}]$.

(ii) Observe first, use (i) with $Y = \mathbf{1}_{\{\hat{Z}=0\}}$, that

$$\mathbb{Q}(\hat{Z} = 0) = \mathbb{E}_{\mathbb{Q}}\mathbf{1}_{\{\hat{Z}=0\}} = \mathbb{E}[\hat{Z}\mathbf{1}_{\{\hat{Z}=0\}}] = 0.$$

Take $G \in \mathcal{G}$ and use (i) with $Y = \mathbb{E}_{\mathbb{Q}}[X\mathbf{1}_G|\mathcal{G}]$ to get

$$\begin{aligned}\mathbb{E}[XZ\mathbf{1}_G] &= \mathbb{E}_{\mathbb{Q}}[X\mathbf{1}_G] \\ &= \mathbb{E}_{\mathbb{Q}}\mathbb{E}_{\mathbb{Q}}[X\mathbf{1}_G|\mathcal{G}] \\ &= \mathbb{E}[\hat{Z}\,\mathbb{E}_{\mathbb{Q}}[X\mathbf{1}_G|\mathcal{G}]] \\ &= \mathbb{E}[\hat{Z}\mathbf{1}_G\mathbb{E}_{\mathbb{Q}}[X|\mathcal{G}]].\end{aligned}$$

Since $\hat{Z}\,\mathbb{E}_{\mathbb{Q}}[X|\mathcal{G}] \in \mathcal{G}$ and G is arbitrary, we conclude that $\mathbb{E}[XZ|\mathcal{G}] = \hat{Z}\,\mathbb{E}_{\mathbb{Q}}[X|\mathcal{G}]$ a.s. under $\mathbb{P}$ and therefore also under $\mathbb{Q}$. The just proved property $\mathbb{Q}(\hat{Z} > 0) = 1$ allows division by $\hat{Z}$ under $\mathbb{Q}$, which then yields Equation (11.10), as also $\frac{\mathbb{E}[XZ|\mathcal{G}]}{\hat{Z}} = \mathbb{E}[X\frac{Z}{\hat{Z}}|\mathcal{G}]$. □

Remark 11.8 One easily shows that $Z = 0$ $\mathbb{P}$-a.s. on the set $N := \{\hat{Z} = 0\}$. Hence $\mathbb{E}[XZ|\mathcal{G}]\mathbf{1}_N = \mathbb{E}[XZ\mathbf{1}_N|\mathcal{G}] = 0$ $\mathbb{P}$-a.s. It also follows that $\mathbb{E}_{\mathbb{Q}}[X|\mathcal{G}]\mathbf{1}_N = 0$ $\mathbb{Q}$-a.s. For the conditionally normalized version of Z, $\tilde{Z} := \frac{Z}{\hat{Z}}$, it holds that the conditional expectation $\mathbb{E}[\tilde{Z}|\mathcal{G}] = 1$ a.s.

In the proof of Proposition 11.10 we shall use the following lemma.

Lemma 11.9 *Let $\mathbb{Q}$ be locally absolutely continuous w.r.t. $\mathbb{P}$ with density process Z and let T be a stopping time. Then for $F \in \mathcal{F}_T$ it holds that*

$$\mathbb{Q}(F \cap \{T < \infty\}) = \mathbb{E}Z_T\mathbf{1}_{F\cap\{T<\infty\}}.$$

Proof Exercise 11.4. □

Proposition 11.10 *Let X be an adapted process and assume $\mathbb{Q}$ locally absolutely continuous w.r.t. $\mathbb{P}$ with density process Z.*

(i) The process X is a martingale under $\mathbb{Q}$ iff the process XZ is a martingale under $\mathbb{P}$.

(ii) The process X is a local martingale under $\mathbb{Q}$ if the process XZ is a local martingale under $\mathbb{P}$.

(iii) If moreover $\mathbb{P}$ is also locally absolutely continuous w.r.t. $\mathbb{Q}$, then the process XZ is a local martingale under $\mathbb{P}$ if the process X is a local martingale under $\mathbb{Q}$.

Proof To prove (i) we use Lemma 11.7 and recall that $\mathbb{E}[Z_n|\mathcal{F}_{n-1}] = Z_{n-1}$. Note that $\mathbb{E}_{\mathbb{Q}}|X_n| < \infty$ iff $\mathbb{E}|X_n|Z_n < \infty$. We have

$$\mathbb{E}_{\mathbb{Q}}[X_n|\mathcal{F}_{n-1}] = \frac{\mathbb{E}[X_n Z_n|\mathcal{F}_{n-1}]}{Z_{n-1}},$$

from which the assertion immediately follows.

(ii) Let (T^n) be a localizing sequence for XZ. Then $\mathbb{P}(\lim_{n\to\infty} T^n < \infty) = 0$ and then also $\mathbb{Q}(\lim_{n\to\infty} T^n < \infty) = 0$ by virtue of Lemma 11.9. Because

$$X_k^{T^n} Z_k = X_k^{T^n} Z_k^{T^n} + X_{T^n}(Z_k - Z_{T^n})\mathbf{1}_{\{k\geq T^n\}} \tag{11.11}$$

and both(!) terms on the right hand side are martingales (Exercise 11.6) under $\mathbb{P}$, we can apply part (i) to deduce that X^{T^n} is a martingale under $\mathbb{Q}$. Hence X is a local martingale under $\mathbb{Q}$.

(iii) Let X be a local martingale under $\mathbb{Q}$ and (T^n) be a localizing sequence for it. Then $\mathbb{Q}(\lim_{n\to\infty} T^n < \infty) = 0$ and by the argument in the proof of (ii) also $\mathbb{P}(\lim_{n\to\infty} T^n < \infty) = 0$, since $\mathbb{P}$ is given to be locally absolutely continuous w.r.t. $\mathbb{Q}$. Similar to Equation (11.11) we have

$$X_k^{T^n} Z_k^{T^n} = X_k^{T^n} Z_k - X_{T^n}(Z_k - Z_{T^n})\mathbf{1}_{\{k\geq T^n\}},$$

again with two martingales under $\mathbb{P}$ on the right hand side, and so XZ is a local martingale under $\mathbb{P}$. □

The next theorem may be considered as Girsanov's theorem in discrete time.

Theorem 11.11 *Let M be a local martingale under $\mathbb{P}$ and let $\mathbb{Q}$ be locally absolutely continuous w.r.t. $\mathbb{P}$. The processes*

$$M^{\mathbb{Q}} = M - \frac{1}{Z} \cdot [M, Z]$$

and

$$\widetilde{M}^{\mathbb{Q}} := M - \frac{1}{Z_-} \cdot \langle M, Z\rangle$$

are well defined under $\mathbb{Q}$ and

(i) $M^{\mathbb{Q}}$ *is a local martingale under* $\mathbb{Q}$*, and*

(ii) if moreover $\mathbb{E}[|\Delta M_n| Z_n \,|\mathcal{F}_{n-1}] < \infty$ *a.s. for all* n*, then also* $\widetilde{M}^{\mathbb{Q}}$ *is a local martingale under* $\mathbb{Q}$*.*

Proof The property $\mathbb{Q}(Z_n > 0) = 1$ allows for division by Z_n a.s. under $\mathbb{Q}$.
(i) Let us compute

$$\begin{aligned}\Delta M_n^{\mathbb{Q}} &= \Delta M_n - \frac{\Delta M_n \Delta Z_n}{Z_n} \\ &= \frac{\Delta M_n}{Z_n}(Z_n - \Delta Z_n) \\ &= \frac{\Delta M_n}{Z_n} Z_{n-1}.\end{aligned}$$

Hence, by Lemma 11.7 with $Z = Z_n$, $\mathcal{G} = \mathcal{F}_{n-1}$ and $\frac{Z}{\mathbb{E}[Z|\mathcal{G}]} = \frac{Z_n}{Z_{n-1}}$ we get

$$\mathbb{E}_{\mathbb{Q}}[\Delta M_n^{\mathbb{Q}} | \mathcal{F}_{n-1}] = \mathbb{E}[\Delta M_n | \mathcal{F}_{n-1}] = 0.$$

(ii) Recall that $\Delta\langle M, Z\rangle_n = \mathbb{E}[\Delta M_n \Delta Z_n | \mathcal{F}_{n-1}]$ and

$$\Delta(\widetilde{M}^{\mathbb{Q}} Z)_n = Z_{n-1}\Delta\widetilde{M}_n^{\mathbb{Q}} + \widetilde{M}_{n-1}^{\mathbb{Q}}\Delta Z_n + \Delta\widetilde{M}_n^{\mathbb{Q}}\Delta Z_n.$$

Hence

$$\begin{aligned}\Delta(\widetilde{M}^{\mathbb{Q}} Z)_n &= Z_{n-1}(\Delta M_n - \frac{\Delta\langle M, Z\rangle_n}{Z_{n-1}}) + \widetilde{M}_{n-1}^{\mathbb{Q}}\Delta Z_n + \Delta\widetilde{M}_n^{\mathbb{Q}}\Delta Z_n \\ &= Z_{n-1}\Delta M_n + (\widetilde{M}_{n-1}^{\mathbb{Q}} - \frac{\Delta\langle M, Z\rangle_n}{Z_{n-1}})\Delta Z_n \\ &\quad + \Delta M_n \Delta Z_n - \Delta\langle M, Z\rangle_n.\end{aligned}$$

The first two terms are obviously local martingale differences under $\mathbb{P}$ by virtue of Proposition 11.3, whereas the same holds true for the difference $\Delta\widetilde{M}_n^{\mathbb{Q}}\Delta Z_n - \Delta\langle M, Z\rangle_n$ under the stated assumption, see Section 11.2. □

The second assertion of Theorem 11.11 is the most appealing one, since it allows to write

$$M = \widetilde{M}^{\mathbb{Q}} + \frac{1}{Z_-} \cdot \langle M, Z\rangle,$$

where the first term is a local martingale under $\mathbb{Q}$ and the second term is predictable. This gives the (generalized) Doob decomposition of M under $\mathbb{Q}$.

The terms Z and Z_- in the denominator in $M^{\mathbb{Q}}$ and $\widetilde{M}^{\mathbb{Q}}$ are annoying in the sense that they may be zero with positive $\mathbb{P}$-probability, although with $\mathbb{Q}$-probability one they are strictly positive. This implies that in general $M^{\mathbb{Q}}$ and $\widetilde{M}^{\mathbb{Q}}$ are not well defined under $\mathbb{P}$. Clearly, this annoyance disappears under the additional condition that $\mathbb{P}$ and $\mathbb{Q}$ are locally equivalent. Alternatively, there

exists another way to transform the given local martingale under $\mathbb{P}$ into a local martingale under $\mathbb{Q}$ in which all terms involved are also well defined under $\mathbb{P}$. We use the following notation.

$$\alpha_k = \frac{Z_k}{Z_{k-1}} \mathbf{1}_{\{Z_{k-1}>0\}}.$$

As we observed before, on the set $\{Z_{k-1} = 0\}$ also $Z_k = 0$. Hence, adopting the convention $\frac{0}{0} = 0$, we may also write $\alpha_k = \frac{Z_k}{Z_{k-1}}$.

Theorem 11.12 *Let M be a local martingale under $\mathbb{P}$ and let $\mathbb{Q}$ be locally absolutely continuous w.r.t. $\mathbb{P}$. Assume $\mathbb{E}[|\Delta M_n|\alpha_n|\mathcal{F}_{n-1}] < \infty$ $\mathbb{P}$-a.s. for all $n \geq 1$. Then the process $\hat{M}^{\mathbb{Q}}$ defined by*

$$\hat{M}^{\mathbb{Q}}_n := M_n - \sum_{k=1}^{n} \mathbb{E}[\alpha_k \Delta M_k|\mathcal{F}_{k-1}]$$

is a local martingale under $\mathbb{Q}$.

Proof The proof is analogous to the proof of Theorem 11.11. With Lemma 11.7 we compute

$$\mathbb{E}_{\mathbb{Q}}[\Delta M_n|\mathcal{F}_{n-1}] = \mathbb{E}[\Delta M_n \alpha_n|\mathcal{F}_{n-1}].$$

Hence

$$\mathbb{E}_{\mathbb{Q}}[\Delta \hat{M}^{\mathbb{Q}}_n|\mathcal{F}_{n-1}] = 0.$$

Note that we have

$$\mathbb{E}_{\mathbb{Q}}[|M_n|\,|\mathcal{F}_{n-1}] \leq |M_{n-1}| + \mathbb{E}[|\Delta M_n|\alpha_n|\mathcal{F}_{n-1}] < \infty,$$

Hence also $\mathbb{E}_{\mathbb{Q}}[|\hat{M}^{\mathbb{Q}}_n|\,|\mathcal{F}_{n-1}] < \infty$ $\mathbb{P}$-a.s. and therefore $\hat{M}^{\mathbb{Q}}$ is a local martingale under $\mathbb{Q}$. □

Remark 11.13 Under the assumptions of Theorems 11.11 and 11.12 one computes

$$\mathbb{E}_{\mathbb{Q}}[\Delta M_n|\mathcal{F}_{n-1}] = \frac{\mathbf{1}_{\{Z_{n-1}>0\}}}{Z_{n-1}} \mathbb{E}[\Delta M_n \Delta Z_n|\mathcal{F}_{n-1}].$$

Hence, in this case the processes $\widetilde{M}^{\mathbb{Q}}$ and $\hat{M}^{\mathbb{Q}}$ are the same under $\mathbb{Q}$.

In the remainder of this section we assume that $\mathbb{P}$ and $\mathbb{Q}$ are locally equivalent. Then all Z_n are positive under $\mathbb{P}$. Define the process μ by $\mu_0 = 1$ and

$$\Delta\mu_n := \frac{\Delta Z_n}{Z_{n-1}} > -1.$$

Obviously, μ is a a martingale under $\mathbb{P}$. One immediately checks that

$$Z = 1 + Z_- \cdot \mu.$$

Hence for any process X, we find $[Z, X] = Z_- \cdot [\mu, X]$, and if $\langle Z, X\rangle$ exists, we even have $\langle Z, X\rangle = Z_- \cdot \langle \mu, X\rangle$. Under the assumptions of Theorem 11.11 (ii) we then find the simple expression, of course meaningful under both $\mathbb{P}$ and $\mathbb{Q}$,

$$\widetilde{M}^{\mathcal{Q}} = M - \langle \mu, M\rangle. \tag{11.12}$$

11.4 Exercises

11.1 Consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Suppose $X \geq 0$ and that $\mathcal{G}$ is a sub-σ-algebra of $\mathcal{F}$. Then there exists a $\mathcal{G}$-measurable random variable $\hat{X}$ such that $\mathbb{E}X\mathbf{1}_G = \mathbb{E}\hat{X}\mathbf{1}_G$ for all $G \in \mathcal{G}$.

11.2 Consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and let $\{A_n, n \geq 1\}$ be a measurable partition of Ω with $\mathbb{P}(A_n) = 2^{-n}$. Let (Z_n) be a sequence of random variables, independent of the A_n, with $\mathbb{P}(Z_n = \pm 2^n) = \frac{1}{2}$. Put $Y_n = \sum_{i=1}^n \mathbf{1}_{A_j} Z_j$ for $n \geq 1$, Y_∞, the a.s. limit of the Y_n, as well as $X_0 = 0$ and $X_n = Y_\infty$ if $n \geq 1$, $T^n = \infty \cdot \mathbf{1}_{\{\cup_{1\leq k\leq n} A_k\}}$.

(a) Show that Y_∞ is well defined and finite a.s. What is it?
(b) Show that $X_k^{T^n} = Y_n$ for $k \geq 1$.
(c) Show that X^{T^n} is a martingale for every n.
(d) Show that X_1 is not integrable (and hence X is not a martingale).

11.3 Prove Equation (11.9) for local martingales X and Y satisfying $\mathbb{E}[\Delta X_n^2 | \mathcal{F}_{t-1}] < \infty$ and $\mathbb{E}[\Delta Y_n^2 | \mathcal{F}_{t-1}] < \infty$.

11.4 Prove Lemma 11.9.

11.5 Prove Proposition 11.5.

11.6 Show that the process $\{X_{T^n}(Z_k - Z_{T^n})\mathbf{1}_{\{k \geq T^n\}},\ k \geq 0\}$ in the proof of Proposition 11.10 is a martingale under $\mathbb{P}$.

11.7 Let $Y_1, Y_2, \ldots : \Omega \to \mathbb{R}$ be *iid* random variables that have a normal $N(\theta_0, \sigma)$ distribution under a probability measure $\mathbb{P}$. We use the filtration of σ-algebras $\mathcal{F}_n = \sigma(Y_1, \ldots, Y_n)$ and $\mathcal{F} = \mathcal{F}_\infty$. Clearly, M defined by $M_n = \sum_{k=1}^n (Y_k - \theta_0)$ is a martingale, even square integrable. Let for some $\theta_1 \in \mathbb{R}$ and $n \geq 1$

$$Z_n = \exp\Big(\frac{\theta_1 - \theta_0}{\sigma^2} \sum_{k=1}^n Y_k + \frac{n}{2\sigma^2}(\theta_0^2 - \theta_1^2)\Big).$$

(a) Show that $Z = (Z_n)$ is martingale. What is $\mathbb{E}Z_n$?
(b) The Z_n can be taken as densities, $Z_n = \frac{\mathrm{d}\mathbb{Q}_n}{\mathrm{d}\mathbb{P}_n}$. Use Theorem 11.11 or Equation (11.12) to determine $\widetilde{M}^{\mathbb{Q}}$.
(c) The $\mathbb{Q}^n$ can be seen as restriction of a probability measure $\mathbb{Q}$ on $(\Omega, \mathcal{F})$ to $\mathcal{F}_n$. Is $\mathbb{Q} \ll \mathbb{P}$?

12 Weak convergence

In this chapter we encounter yet another convergence concept for random variables, weak convergence. Although the origin of the terminology is in functional analysis, fortunately weak convergence is also weaker than the other kinds of convergence for random variables that we have seen sofar. First we sketch a functional analytic background of weak convergence. After that we go back to our probabilistic environment.

Consider a complete normed vector space $(X, ||\cdot||)$, a Banach space, and let X^* be the vector space of all continuous linear functionals $T : X \to \mathbb{R}$, also called the (strong) dual space of X. The operator norm of $T \in X^*$ is defined as

$$||T|| = \sup\{\frac{|Tx|}{||x||} : ||x|| \neq 0\}.$$

It is known that a linear functional is continuous iff it has finite norm. Note that we use the same symbol $||\cdot||$ to denote both the norm on X and the one on X^*. It follows that for all $x \in X$ one has $|Tx| \leq ||T||\,||x||$. One can show that $(X^*, ||\cdot||)$ is a Banach space. Let (x_n) be a sequence in X that converges to x in norm, $||x_n - x|| \to 0$. If $T \in X^*$, we then have

$$|Tx_n - Tx| \leq ||T||\,||x_n - x|| \to 0. \tag{12.1}$$

If a sequence (x_n) satisfies (12.1) for some $x \in X$ and all $T \in X^*$, we say that x_n converges to x *weakly*.

Now we mimic the above by taking $Y = X^*$ as the basic space and along with Y we consider Y^*, also denoted by X^{**}. A sequence of operators $(T_n) \subset Y$ then strongly converges to $T \in Y$, if $||T_n - T|| \to 0$. Of course, we then also have for all $y^* \in Y^*$ that

$$|y^*T_n - y^*T| \to 0. \tag{12.2}$$

Parallelling the above, we say that T_n converges weakly to T if (12.2) holds for all $y^* \in Y^*$.

Let's have a look at some special linear operators in $Y^* = X^{**}$. Let $T \in X^*$ and $x \in X$. Define $x(T) = Tx$. We can view $x(\cdot)$ as an element of X^{**}, which is easy to check. A sequence $(T_n) \subset X^*$ which weakly converges to $T \in X^*$, then also satisfies (12.2) for $y^* = x(\cdot)$ and we have

$$|T_nx - Tx| \to 0, \forall x \in X. \tag{12.3}$$

If (12.3) happens, we say that T_n converges to T in the *weak* sense*. This convergence is by construction weaker than weak convergence, and in general strictly weaker.

Let's look at a more concrete situation. Consider $C(\mathbb{R})$, the space of continuous functions and we take as a norm on this space, the sup-norm, $||f|| = \sup\{|f(x)| : x \in \mathbb{R}\}$. We take X as the subset of $C(\mathbb{R})$ consisting of functions

with finite norm. Note that every $f \in X$ is bounded. We write $C_b(\mathbb{R})$ for this space. One easily checks that $C_b(\mathbb{R})$ is a linear space and one can verify by a direct argument that it is complete. This is also known from Theorem 4.49, upon noticing that $C_b(\mathbb{R})$ is a closed subspace of $\mathcal{L}^\infty(\mathbb{R}, \mathcal{B}, \lambda)$.

Probability theory comes in when we look at special operators on $C_b(\mathbb{R})$, originating from probability measures on $(\mathbb{R}, \mathcal{B})$. Although in this chapter we will confine ourselves mainly to $(\mathbb{R}, \mathcal{B})$, occasionally we will hint at a more general context. If μ is such a probability measure, we view the integral $\mu : f \to \mu(f)$ as a linear functional. It is easy to check that every μ, viewed as a functional, has norm $||\mu|| = 1$. Notice that the space of probability measures on $(\mathbb{R}, \mathcal{B})$ is not a vector space, but only a convex set. The encompassing vector space is the set of all signed finite measures. Following the above scheme, weak* convergence for a sequence of probability measures (μ_n) means $\mu_n(f) \to \mu(f)$, for all $f \in C_b(\mathbb{R})$. However, in probability theory, it has become a *convention* to speak of *weak* convergence of a sequence of probability measures instead of weak* convergence, partly because the above notion of weak convergence turns out to be too strong and not useful for certain purposes, see further down in this chapter. One may view $C_b(\mathbb{R})$ as a space of *test functions*. Other choices for of test functions are also possible, for instance the space of continuous functions with compact support, $C_K(\mathbb{R})$, or the space of continuous functions that converge to zero at $\pm\infty$, $C_0(\mathbb{R})$. One can show that for convergence of *probability measures* on $(\mathbb{R}, \mathcal{B})$, the definition of convergence by using any of these collections yields the same convergence concept. To define convergence of a more general (signed) measures (possibly even defined on more general metric, or even topological spaces, with the Borel σ-algebra), the use of the different test function lead to different convergence concepts. To distinguish between these, different names are used. What we have just called weak convergence is then called narrow convergence, whereas the name weak convergence is then reserved for $C_0(\mathbb{R})$ as the space of test functions. In fact, a theorem by Riesz says that the dual space of $C_0(\mathbb{R})$ can be identified with the space of all signed measures on $(\mathbb{R}, \mathcal{B})$, which makes $C_0(\mathbb{R})$ in a sense more natural to work with, if one considers weak convergence. On the other hand, we shall characterize weak convergence of probability measures by their action on a relatively small but rich enough collection of functions that are not in $C_0(\mathbb{R})$, see Theorem 12.15.

Below we will adhere to the custom followed in probability theory.

12.1 Generalities

Here is the formal definition of weak convergence of probability measures on $(\mathbb{R}, \mathcal{B})$ and of a sequence of random variables.

Definition 12.1 Let $\mu, \mu_1, \mu_2, \ldots$ be probability measures on $(\mathbb{R}, \mathcal{B})$. It is said that μ_n converges weakly to μ, and we then write $\mu_n \xrightarrow{w} \mu$, if $\mu_n(f) \to \mu(f)$ for all $f \in C_b(\mathbb{R})$. If $X, X_1, X_2, \ldots$ are (real) random variables (possibly defined on different probability spaces) with distributions $\mu, \mu_1, \mu_2, \ldots$ then we say that

X_n converges weakly to X, and write $X_n \xrightarrow{w} X$ if it holds that $\mu_n \xrightarrow{w} \mu$. In this case, one also says that X_n converges to X in distribution.

Other accepted notation for weak convergence of a sequence of random variables is $X_n \xrightarrow{d} X$, one says that X_n converges to X *in distribution*. Later we will see an appealing characterization of weak convergence (convergence in distribution) in terms of distribution functions, which makes the definition less abstract. Look at the following example, that illustrates for a special case, that there is some reasonableness in Definition 12.1. Let (x_n) be a convergent sequence, suppose with $\lim x_n = 0$. Then for every $f \in C_b(\mathbb{R})$ one has $f(x_n) \to f(0)$. Let μ_n be the Dirac measure concentrated on $\{x_n\}$ and μ the Dirac measure concentrated in the origin. Since $\mu_n(f) = f(x_n)$, we see that $\mu_n \xrightarrow{w} \mu$.

One could naively think of another definition of convergence of (probability) measures, for instance by requiring that $\mu_n(B) \to \mu(B)$ for every $B \in \mathcal{B}$, or even by requiring that the integrals $\mu_n(f)$ converge to $\mu(f)$ for every bounded measurable function. It turns out that each of these requirements is too strong to get a useful convergence concept. One drawback of such a definition can be illustrated by the above example with the Dirac measures. Take $B = (-\infty, x]$ for some $x > 0$. Then for all $n > 1/x$, we have $\mu_n(B) = \mathbf{1}_B(\frac{1}{n}) = 1$ and $\mu(B) = \mathbf{1}_B(0) = 1$. For $x < 0$, we get that all measures of B are equal to zero. But for $B = (-\infty, 0]$, we have $\mu_n(B) = 0$ for all n, whereas $\mu(B) = 1$. Hence convergence of $\mu_n(B) \to \mu(B)$ doesn't hold for this choice of B. If F_n is the distribution function of μ_n and F that of μ, then we have seen that $F_n(x) \to F(x)$, for all $x \in \mathbb{R}$ *except for* $x = 0$.

Suppose that a random variable X has distribution μ. Recall from Proposition 4.27 that $\mathbb{E}f(X) = \mathbb{E}f \circ X = \int f \, \mathrm{d}\mu$. It follows that $X_n \xrightarrow{w} X$ iff $\mathbb{E}f(X_n) \to \mathbb{E}f(X)$ for all $f \in C_b(\mathbb{R})$. Note that X and the X_n need not be defined on the same probability space, but that an expectation doesn't depend on the underlying probability space, only on the law of the random variable involved.

Next we give the relation between weak convergence and other types of convergence. The proposition says that weak convergence is indeed weaker than other types of convergence that we encountered.

Proposition 12.2 *Let the random variables $X, X_1, X_2, \ldots$ be defined on a single probability space. If $X_n \xrightarrow{\text{a.s.}} X$ or if $X_n \xrightarrow{\mathbb{P}} X$, then $X_n \xrightarrow{w} X$. If $X_n \xrightarrow{w} X$ and $g : \mathbb{R} \to \mathbb{R}$ is continuous, then also $g(X_n) \xrightarrow{w} g(X)$. Finally, if $X_n \xrightarrow{w} x$, where $x \in \mathbb{R}$ is a constant random variable, then also $X_n \xrightarrow{\mathbb{P}} x$.*

Proof Assume $X_n \xrightarrow{\text{a.s.}} X$. If $f \in C_b(\mathbb{R})$, then also $f(X_n) \xrightarrow{\text{a.s.}} f(X)$. Since f is bounded, we can apply the Dominated Convergence Theorem to get the assertion. The remainder of the proof is left as Exercise 12.1. □

One could guess that checking weak convergence of a sequence of random variables may be a hard job, one needs to work with all functions in $C_b(\mathbb{R})$. Fortunately there is a fine characterization in terms of distribution functions, and

these are the objects that we are quite familiar with. As a first result, we have the following proposition. Recall the notation $F(x-) = \lim_{y \uparrow x} F(y)$.

Proposition 12.3 *Let $\mu, \mu_1, \mu_2 \ldots$ be a sequence of probability measures on $(\mathbb{R}, \mathcal{B})$ and let $F, F_1, F_2, \ldots$ be their distribution functions. Assume that $\mu_n \xrightarrow{w} \mu$. Then one has $\limsup F_n(x) \leq F(x)$ for all $x \in \mathbb{R}$ and $\liminf F_n(x) \geq F(x-)$ for all $x \in \mathbb{R}$. In particular, $\lim F_n(x) = F(x)$ for all $x \in C_F$, the set of points where F is continuous.*

Proof By definition, $F(x) = \mu((-\infty, x]) = \int \mathbf{1}_{(-\infty,x]} \, \mathrm{d}\mu$, an integral of the discontinuous function $\mathbf{1}_{(-\infty,x]}$. In order to connect to the definition of weak convergence, we approximate the indicator with continuous functions as follows. Let $x \in \mathbb{R}$, $\varepsilon > 0$ and define g by $g(y) = 1$ if $y \leq x$, $g(y) = 0$, if $y \geq x + \varepsilon$ and by linear interpolation on $(x, x+\varepsilon)$. Then $g \in C_b(\mathbb{R})$ and $\mathbf{1}_{(-\infty,x]} \leq g \leq \mathbf{1}_{(-\infty,x+\varepsilon]}$. Therefore we have

$$\begin{aligned} F_n(x) \leq \mu_n(g) \leq F_n(x+\varepsilon), \\ F(x) \leq \mu_(g) \leq F(x+\varepsilon). \end{aligned}$$

Hence, from the postulated weak convergence we have

$$\limsup F_n(x) \leq \limsup \mu_n(g) = \mu(g) \leq F(x+\varepsilon).$$

Letting $\varepsilon \downarrow 0$, we get by right-continuity of F that $\limsup F_n(x) \leq F(x)$.

To prove the statement concerning the liminf we use

$$\liminf F_n(x+\varepsilon) \geq \liminf \mu_n(g) = \mu(g) \geq F(x).$$

Since this holds for all $x \in \mathbb{R}$, we rename $x + \varepsilon$ as x to obtain $\liminf F_n(x) \geq F(x-\varepsilon)$. Let $\varepsilon \downarrow 0$. The final statement then also follows. □

If in Proposition 12.3 the limit function F is continuous, then convergence of F_n to F is even uniform, see Exercise 12.13. This proposition also has a converse statement, which can be directly proved, see Exercise 12.5, but we prefer to let it follow from Skorohod's representation theorem, Theorem 12.4, as a trivial consequence. Note that this theorem is a subtle kind of converse to Proposition 12.2.

Theorem 12.4 *Let $F, F_1, F_2, \ldots$ be distribution functions satisfying $F_n(x) \to F(x)$ for all $x \in C_F$. Then there exists a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and random variables $X, X_1, X_2, \ldots$ defined on it such that X has distribution function F, the X_n have distribution functions F_n and $X_n \xrightarrow{\text{a.s.}} X$.*

Proof Look back at Theorem 3.10 and its proof. We take $([0,1], \mathcal{B}, \lambda)$ as our probability space and we consider the random variables X^- and X_n^- $(n \geq 1)$, the definition of them should be clear. Along with these random variables we also need $X^+(\omega) = \inf\{z : F(z) > \omega\}$ and the similarly defined $X_n^+(\omega)$. Fix $\omega \in (0,1)$ and take $z \in C_F$ with $z > X^+(\omega)$. Then $F(z) > \omega$ and by the assumed

convergence, eventually all $F_n(z) > \omega$. It follows that $z \geq \limsup X_n^+(\omega)$. Since this holds for all $z \in C_F$, we can choose a sequence of them (there is plenty of choice, since the complement of C_F is at most countable) that decreases to $X^+(\omega)$ to obtain $X^+(\omega) \geq \limsup X_n^+(\omega)$. Similar reasoning yields $X^-(\omega) \leq \liminf X_n^-(\omega)$. Since $X_n^-(\omega) \leq X_n^+(\omega)$, we deduce $X^-(\omega) \leq \liminf X_n^+(\omega) \leq \limsup X_n^+(\omega) \leq X^+(\omega)$. By Exercise 3.7, we know that $\lambda(\{X^- < X^+\}) = 0$ and we thus conclude that $\liminf X_n^+ = \limsup X_n^+ = X$ on $\{X^- = X^+\}$, which has probability one. Hence the X_n and X in the assertion can be taken as $X_n = X_n^+$ and $X = X^+$ (or as $X_n = X_n^-$ and $X = X^-$). □

Corollary 12.5 *Let $\mu, \mu_1, \mu_2, \ldots$ be probability measures on $(\mathbb{R}, \mathcal{B})$ with distribution functions $F, F_1, F_2, \ldots$. Equivalence holds between*

(i) $F_n(x) \to F(x)$ for all $x \in C_F$ and

(ii) $\mu_n \xrightarrow{w} \mu$.

Proof In view of Proposition 12.3 it is sufficient to establish (i) ⇒ (ii). This implication follows by combining Proposition 12.2 and Theorem 12.4. □

Here is a result that gives an appealing sufficient condition for weak convergence, when the random variables involved admit densities.

Theorem 12.6 *Consider real random variables $X, X_1, X_2, \ldots$ having densities $f, f_1, f_2, \ldots$ w.r.t. Lebesgue measure λ. Suppose that $f_n \to f$ λ-a.e. Then $X_n \xrightarrow{w} X$.*

Proof We apply Scheffé's lemma, Lemma 4.20, to conclude that $f_n \to f$ in $\mathcal{L}^1(\mathbb{R}, \mathcal{B}, \lambda)$. Let $g \in C_b(\mathbb{R})$. Since g is bounded, we also have $f_n g \to fg$ in $\mathcal{L}^1(\mathbb{R}, \mathcal{B}, \lambda)$ and hence $X_n \xrightarrow{w} X$. □

The Bolzano-Weierstraß theorem states that every bounded sequence of real numbers has a convergent subsequence. The theorem easily generalizes to sequences in $\mathbb{R}^n$, but fails to hold for uniformly bounded sequences in $\mathbb{R}^\infty$. But if extra properties are imposed, there can still be an affirmative answer. Something like that happens in the next theorem that is known as Helly's selection theorem. It is convenient to introduce the concept of *defective distribution function*. Such a function, F say, has values in $[0, 1]$ is by definition right-continuous, increasing but at least one of the two properties $\lim_{x\to\infty} F(x) = 1$ and $\lim_{x\to-\infty} F(x) = 0$ fails to hold. The measure μ corresponding to F on $(\mathbb{R}, \mathcal{B})$ will then be a *subprobability measure*, $\mu(\mathbb{R}) < 1$.

Theorem 12.7 *Let (F_n) be a sequence of distribution functions. Then there exists a, possibly defective, distribution function F and a subsequence (F_{n_k}) such that $F_{n_k}(x) \to F(x)$, for all $x \in C_F$.*

Proof The proof's main ingredients are an infinite repetition of the Bolzano-Weierstraß theorem combined with a Cantor diagonalization. First we restrict ourselves to working on $\mathbb{Q}$, instead of $\mathbb{R}$, and exploit the countability of $\mathbb{Q}$.

Write $\mathbb{Q} = \{q_1, q_2, \ldots\}$ and consider the F_n restricted to $\mathbb{Q}$. Then the sequence $(F_n(q_1))$ is bounded and along some subsequence (n_k^1) it has a limit, $\ell(q_1)$ say. Look then at the sequence $F_{n_k^1}(q_2)$. Again, along some subsequence of (n_k^1), call it (n_k^2), we have a limit, $\ell(q_2)$ say. Note that along the thinned subsequence, we still have the limit $\lim_{k\to\infty} F_{n_k^2}(q_1) = \ell(q_1)$. Continue like this to construct a *nested* sequence of subsequences (n_k^j) for which we have that $\lim_{k\to\infty} F_{n_k^j}(q_i) = \ell(q_i)$ holds for every $i \le j$. Put $n_k = n_k^k$, then (n_k) is a subsequence of (n_k^i) for every $i \le k$. Hence for any fixed i, eventually $n_k \in (n_k^i)$. It follows that for arbitrary i one has $\lim_{k\to\infty} F_{n_k}(q_i) = \ell(q_i)$. In this way we have constructed a function $\ell : \mathbb{Q} \to [0,1]$ and by the monotonicity of the F_n this function is increasing.

In the next step we extend this function to a function F on $\mathbb{R}$ that is right-continuous, and still increasing. We put

$$F(x) = \inf\{\ell(q) : q \in \mathbb{Q},\, q > x\}.$$

Note that in general $F(q)$ is not equal to $\ell(q)$ for $q \in \mathbb{Q}$, but the inequality $F(q) \ge \ell(q)$ always holds true. Obviously, F is an increasing function and by construction it is right-continuous. An explicit verification of the latter property is as follows. Let $x \in \mathbb{R}$ and $\varepsilon > 0$. There is $q \in \mathbb{Q}$ with $q > x$ such that $\ell(q) < F(x) + \varepsilon$. Pick $y \in (x, q)$. Then $F(y) < \ell(q)$ and we have $F(y) - F(x) < \varepsilon$. Note that it may happen that for instance $\lim_{x\to\infty} F(x) < 1$, F can be defective.

The function F is of course the one we are aiming at. Having verified that F is a (possibly defective) distribution function, we show that $F_{n_k}(x) \to F(x)$ if $x \in C_F$. Take such an x and let $\varepsilon > 0$ and q as above. By left-continuity of F at x, there is $y < x$ such that $F(x) < F(y) + \varepsilon$. Take now $r \in (y, x) \cap \mathbb{Q}$, then $F(y) \le \ell(r)$, hence $F(x) < \ell(r) + \varepsilon$. So we have the inequalities

$$\ell(q) - \varepsilon < F(x) < \ell(r) + \varepsilon.$$

Then $\limsup F_{n_k}(x) \le \lim F_{n_k}(q) = \ell(q) < F(x) + \varepsilon$ and $\liminf F_{n_k}(x) \ge \liminf F_{n_k}(r) = \ell(r) > F(x) - \varepsilon$. The result follows since ε is arbitrary. □

Here is an example for which the limit is not a true distribution function. Let μ_n be the Dirac measure concentrated on $\{n\}$. Then its distribution function is given by $F_n(x) = \mathbf{1}_{[n,\infty)}(x)$ and hence $\lim_{n\to\infty} F_n(x) = 0$. Hence any limit function F in Theorem 12.7 has to be the zero function, which is clearly defective.

Translated into terms concerning probability measures on $(\mathbb{R}, \mathcal{B})$, the proposition seems to say that every sequence of probability measures has a weakly converging subsequence whose limit μ is a subprobability measure, $\mu(\mathbb{R}) \le 1$. In topological terms this would mean that the family of probability measure is relatively sequentially compact (w.r.t. topology generated by weak convergence). But look again at the example with the Dirac measures. The integral $\mu_n(f)$ is equal to $f(n)$, which has in general no limit for $n \to \infty$, if $f \in C_b(\mathbb{R})$,

although the zero measure is the only possible limit. There are a number of ways to circumvent the problem. One of them is to replace in the definition of weak convergence the space $C_b(\mathbb{R})$ with the smaller set $C_0(\mathbb{R})$. Another way-out is to look at probability measures on the Borel sets of $[-\infty, \infty]$. The space $C([-\infty, \infty])$ can be identified with $C[0, 1]$ and in this space every continuous function automatically has limits at the boundary points. For the sequence of Dirac measures, we would then have the Dirac measure concentrated on $\{\infty\}$ as the limit and weak convergence holds again.

The example with the Dirac measures also provides another insight why the limit is only a defective distribution function, the point masses at n 'disappear' from $\mathbb{R}$ as n tends to infinity. A possible way out to prevent this phenomenon is by requiring that all probability measures involved have probability one on a fixed bounded set. This is too stringent, because it rules out many useful distributions. Fortunately, a considerably weaker assumption suffices. For any probability measure μ on $(\mathbb{R}, \mathcal{B})$ it holds that $\lim_{M\to\infty} \mu([-M, M]) = 1$. If F is the distribution function of μ, we equivalently have $\lim_{M\to\infty}(F(M)-F(-M)) = 1$. Note that $F(M) - F(-M) = \mu((-M, M])$. The next condition, *tightness*, gives a uniform version of this.

Definition 12.8 A sequence of probability measures (μ_n) on $(\mathbb{R}, \mathcal{B})$ is called tight, if $\lim_{M\to\infty} \inf_n \mu_n([-M, M]) = 1$.

Remark 12.9 Note that a sequence (μ_n) is tight iff if every 'tail sequence' $(\mu_n)_{n\geq N}$ is tight. In order to show that a sequence is tight it is thus sufficient to show tightness from a certain suitably chosen index on. Tightness of a sequence is also a necessary condition for weak convergence, as we shall see later. Recall that a distribution function F has at most countably points of discontinuity and that $\mu(\{x\}) > 0$ iff x is a discontinuity point of F. In this case $\{x\}$ is called an atom of μ.

Proposition 12.10 *Weak convergence and tightness are related as follows.*

(i) Let (μ_n) be a sequence of probability measures that weakly converges to a probability measure μ on $(\mathbb{R}, \mathcal{B})$. Then (μ_n) is tight.

(ii) Let (F_n) be the distribution functions of a tight sequence of probability measures (μ_n) on $(\mathbb{R}, \mathcal{B})$. Then there exists a (proper) distribution function F and a subsequence (F_{n_k}) such that $F_{n_k}(x) \to F(x)$, for all $x \in C_F$. Equivalently, there exists a probability measure μ on $(\mathbb{R}, \mathcal{B})$ and a subsequence (μ_{n_k}) such that $\mu_{n_k} \xrightarrow{w} \mu$.

Proof (i) Fix $\varepsilon > 0$ and choose $M > 0$ such that $\mu([-M, M]) > 1 - \varepsilon$. Since the collection of all atoms of all μ_n is at most countable, we can choose M such that it is not an atom of any μ_n and not of μ. If F and F_n are the corresponding distribution functions, we thus have that $F_n(\pm M) \to F(\pm M)$. Hence, there is $N > 0$ such that $|F_n(\pm M) - F(\pm M)| < \varepsilon$ for $n > N$. For these n we then have by the triangle inequality that $\mu_n([-M, M]) > 1 - 3\varepsilon$. Hence the sequence $(\mu_n)_{n>N}$ is tight.

(ii) The tightness condition means that for all $\varepsilon > 0$ we can find $M > 0$ such that $\mu_n([-M, M]) > 1 - \varepsilon$ for all n. Again we may assume that the singletons M and $-M$ are no atoms of all the probability measures involved. Take a subsequence as in Theorem 12.7 and choose N such that $|F_{n_k}(\pm M) - F(\pm M)| < \varepsilon$ for $n_k > N$. Then $F(M) = (F(M) - F_{n_k}(M)) + F_{n_k}(M) > 1 - 2\varepsilon$ and likewise $F(-M) < 2\varepsilon$. It follows that F is not defective. □

All definitions and results so far generalize without difficulties to (weak) convergence of sequences of random vectors with values in $\mathbb{R}^n$, although some care must be taken in formulating the statements about convergence of distribution functions at points where the limit is continuous. Take this for granted or verify it, if you want. Here is a useful warning. If you know of two sequences of random variables that $X_n \xrightarrow{w} X$ and $Y_n \xrightarrow{w} Y$, it tells you a priori nothing about weak convergence of the random vectors (X_n, Y_n), simply because nothing is known of the *joint* distribution of the (X_n, Y_n). Under extra conditions something can be said though.

Proposition 12.11 *Assume for simplicity that all random variables below are defined on the same space.*

(i) If (X_n) and (Y_n) are independent sequences with $X_n \xrightarrow{w} X$ and $Y_n \xrightarrow{w} Y$, then also $(X_n, Y_n) \xrightarrow{w} (X, Y)$.

(ii) If (X_n) and (Y_n) are sequences of random variables with $X_n \xrightarrow{w} X$ and $Y_n \xrightarrow{w} y$, where $y \in \mathbb{R}$ a constant, then also $(X_n, Y_n) \xrightarrow{w} (X, y)$.

(iii) In any of the previous cases, one also has weak convergence of $(g(X_n, Y_n))$ for a continuous $g : \mathbb{R}^2 \to \mathbb{R}$. In particular there holds weak convergence for $X_n + Y_n$.

Proof Exercise 12.2 □

We close this section with presenting some results that apply to a rather general setting, probability measure defined on separable metric spaces (S, d) endowed with the Borel σ-algebra.

Weak convergence can be characterized by a great variety of properties, we present some of them in the next theorem, known as the *portmanteau theorem*. Recall that the boundary ∂E of a set E in a topological space is $\partial E = \operatorname{Cl} E \setminus \operatorname{Int} E$.

Theorem 12.12 *Let $\mu, \mu_1, \mu_2, \ldots$ be probability measures on a metric space (S, d). The following statements are equivalent.*

(i) $\mu_n \xrightarrow{w} \mu$.

(ii) $\limsup_{n\to\infty} \mu_n(F) \leq \mu(F)$ for all closed sets F.

(iii) $\liminf_{n\to\infty} \mu_n(G) \geq \mu(G)$ for all open sets G.

(iv) $\lim_{n\to\infty} \mu_n(E) = \mu(E)$ for all sets E with $\mu(\partial E) = 0$.

Proof We start with (i)⇒(ii), the proof of which is similar to the one of Proposition 12.3. We construct a function that is one on F, and zero just a bit away from it. Let $\delta > 0$. If $x \notin F$, then $d(x, F) > 0$. Define g by

$$g(x) = \begin{cases} 0 & \text{if } d(x,F) > \delta \\ 1 - d(x,F)/\delta & \text{if } d(x,F) \leq \delta. \end{cases}$$

The key observation is that $\mathbf{1}_F \leq g \leq \mathbf{1}_{F^\delta}$, where $F^\delta = \{x \in S : d(x, F) < \delta\}$ and so the rest of the proof is basically as before.

(ii)⇔(iii) is almost trivial. Knowing this, the implication (iii)⇒(iv) is easy.

(iv)⇒(i) The proof of this implication roughly follows a pattern needed in Exercise 12.5. Let $\varepsilon > 0$, $g \in C_b(S)$ and assume that $0 < g < B$ for some $B > 0$. Let $D = \{x \in \mathbb{R} : \mu(\{g = x\}) > 0\}$. So, D is the set of atoms of g and hence it is at most countable. Let $0 = x_0 < \ldots < x_m = B$ be a finite set of points not in D such that $\max\{x_k - x_{k-1} : k = 1, \ldots, m\} < \varepsilon$. Let $I_k = (x_{k-1}, x_k]$ and $J_k = g^{-1}[I_k]$. The continuity of g implies that $\partial J_k \subset g^{-1}[\{x_{k-1}, x_k\}]$. Hence $\mu(\partial J_k) = 0$. Let $\tilde{g} = \sum_{k=1}^m x_k \mathbf{1}_{J_k}$. Then $|\mu_n(\tilde{g}) - \mu(\tilde{g})| \leq \sum_{k=1}^m x_k |\mu_n(J_k) - \mu(J_k)|$, which tends to zero as $n \to \infty$ by assumption. Since $0 \leq \tilde{g} - g < \varepsilon$, we have $0 \leq \mu(\tilde{g}) - \mu(g) \leq \varepsilon$ and $0 \leq \mu_n(\tilde{g}) - \mu_n(g) \leq \varepsilon$ for all n. Use the triangle inequality twice to obtain

$$\limsup_{n\to\infty} |\mu_n(g) - \mu(g)| \leq 2\varepsilon.$$

This finishes the proof. □

12.2 The Central Limit Theorem

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. We will assume that all random variables that we encounter below are defined on this space and real valued. Let $X, X_1, X_2, \ldots$ be random variables and recall that $X_n \xrightarrow{w} X$ can be cast as

$$\mathbb{E}f(X_n) \to \mathbb{E}f(X), \, \forall f \in C_b(\mathbb{R}). \tag{12.4}$$

As a matter of fact one can show that weak convergence takes place, if (12.4) holds for all bounded *uniformly continuous* functions (Exercise 12.4). In the present section we take this as our characterization of weak convergence.

Our goal is to prove Theorem 12.16 for which we need a couple of preparatory results. The approach followed in this section is based on *smoothing by convolution* and *small disturbances*. In the sequel $||f||$ denotes the sup norm of a function f.

Lemma 12.13 *Let X and Y be random variables and f a bounded uniformly continuous function. Then, for all $\varepsilon > 0$, there exists $\delta > 0$ such that*

$$|\mathbb{E}f(X) - \mathbb{E}f(X + Y)| \leq \varepsilon + 2||f|| \, \mathbb{P}(|Y| \geq \delta). \tag{12.5}$$

Proof Let $\varepsilon > 0$ be given and choose $\delta > 0$ such that $|f(x)-f(y)| < \varepsilon$ whenever $|x-y| < \delta$. Then

$$\begin{aligned}|\mathbb{E}f(X) - \mathbb{E}f(X+Y)| &\le \mathbb{E}(\mathbf{1}_{\{|Y|<\delta\}}|f(X)-f(X+Y)|) \\ &\quad + \mathbb{E}(\mathbf{1}_{\{|Y|\ge\delta\}}(|f(X)|+|f(X+Y)|)) \\ &\le \varepsilon + 2||f||\,\mathbb{P}(|Y|\ge\delta).\end{aligned}$$

□

Lemma 12.14 *Let $Y, X, X_1, X_2, \ldots$ be random variables such that for all $\sigma > 0$ it holds that $X_n + \sigma Y \xrightarrow{w} X + \sigma Y$. Then also $X_n \xrightarrow{w} X$ (the $\sigma = 0$ case).*

Proof Let f be a bounded uniformly continuous function, $\varepsilon > 0$ be given and choose $\delta > 0$ as in the previous lemma. From (12.5) it follows that

$$|\mathbb{E}f(X) - \mathbb{E}f(X+\sigma Y)| \le \varepsilon + 2||f||\,\mathbb{P}(|Y| \ge \frac{\delta}{\sigma})$$

and

$$|\mathbb{E}f(X_n) - \mathbb{E}f(X_n+\sigma Y)| \le \varepsilon + 2||f||\,\mathbb{P}(|Y| \ge \frac{\delta}{\sigma}).$$

Now we consider

$$\begin{aligned}|\mathbb{E}f(X_n) - \mathbb{E}f(X)| &\le |\mathbb{E}f(X_n) - \mathbb{E}f(X_n+\sigma Y)| \\ &\quad + |\mathbb{E}f(X_n+\sigma Y) - \mathbb{E}f(X+\sigma Y)| \\ &\quad + |\mathbb{E}f(X) - \mathbb{E}f(X+\sigma Y)| \\ &\le 2\varepsilon + 4||f||\,\mathbb{P}(|Y| \ge \frac{\delta}{\sigma}) \\ &\quad + |\mathbb{E}f(X_n+\sigma Y) - \mathbb{E}f(X+\sigma Y)|.\end{aligned}$$

By assumption, the last term tends to zero for $n \to \infty$. Letting then $\sigma \downarrow 0$, we obtain $\limsup_n |\mathbb{E}f(X_n) - \mathbb{E}f(X)| \le 2\varepsilon$, which finishes the proof, since ε is arbitrary. □

For small σ, we view $X + \sigma Y$ as a perturbation of X. Let us take a standard normally distributed random variable Y, independent of X and the X_n. Notice that $Z := X + \sigma Y$ given $X = x$ has a $N(x, \sigma^2)$ distribution. Let f be bounded and uniformly continuous. Then $\mathbb{E}f(X+\sigma Y) = \mathbb{E}\mathbb{E}[f(Z)|X]$ and

$$\mathbb{E}[f(Z)|X=x] = \int_{-\infty}^{\infty} f(z)\frac{1}{\sigma\sqrt{2\pi}}\exp(-\frac{1}{2\sigma^2}(z-x)^2)\,\mathrm{d}z =: f_\sigma(x). \quad (12.6)$$

Hence

$$\mathbb{E}f(X+\sigma Y) = \mathbb{E}f_\sigma(X). \quad (12.7)$$

Let $p_\sigma(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{1}{2\sigma^2}x^2)$, the density of a $N(0,\sigma^2)$ distributed random variable. The function f_σ is obtained by convolution of f with the normal density p_σ. By the Dominated Convergence Theorem, one can show (Exercise 12.3) that f has derivatives of all orders given by

$$f_\sigma^{(k)}(x) = \int_{-\infty}^{\infty} f(z) p_\sigma^{(k)}(z-x)\,\mathrm{d}z. \tag{12.8}$$

Hence f_σ is a smooth function. Write $\mathcal{C}^\infty$ for the class of bounded functions that have *bounded* derivatives of all orders. Examples of such function are p_σ and f_σ. We have already weakened the requirement for weak convergence that convergence is assumed to hold for expectations involving uniformly continuous functions. The next step is to drastically reduce this class of functions.

Theorem 12.15 *Let $X, X_1, X_2, \ldots$ be random variables. The weak convergence $X_n \xrightarrow{w} X$ takes place iff $\mathbb{E}f(X_n) \to \mathbb{E}f(X)$, for all $f \in \mathcal{C}^\infty$.*

Proof Suppose that $\mathbb{E}f(X_n) \to \mathbb{E}f(X)$, for all $f \in \mathcal{C}^\infty$, then it holds in particular for any f_σ obtained from a uniformly continuous f as in (12.6). In view of (12.7), we have $\mathbb{E}f(X_n + \sigma Y) \to \mathbb{E}f(X + \sigma Y)$ for all bounded uniformly continuous f, so $X_n + \sigma Y \xrightarrow{w} X + \sigma Y$ for all $\sigma > 0$. Now Lemma 12.14 applies. □

As a final preparation for the proof of the Central Limit Theorem we proceed with some analytic technicalities that eventually lead to the crucial inequality (12.12). Let $f \in \mathcal{C}^\infty$ and put

$$R(x,y) = f(x+y) - f(x) - yf'(x) - \frac{1}{2}y^2 f''(x).$$

Replacing x and y above by independent random variables X and Y and taking expectations, then yields

$$\mathbb{E}f(X+Y) - \mathbb{E}f(X) - \mathbb{E}Y\,\mathbb{E}f'(X) - \frac{1}{2}\mathbb{E}Y^2\,\mathbb{E}f''(X) = \mathbb{E}R(X,Y).$$

Let W be another random variable, independent of X, and assume that $\mathbb{E}W = \mathbb{E}Y$ and $\mathbb{E}W^2 = \mathbb{E}Y^2$. Then a similar equality is valid and we then obtain by taking the difference the inequality

$$|\mathbb{E}f(X+Y) - \mathbb{E}f(X+W)| \le \mathbb{E}|R(X,Y)| + \mathbb{E}|R(X,W)|. \tag{12.9}$$

We are now going to find bounds on the remainder terms in this equation. The mean value theorem yields for any x and y that $R(x,y) = \frac{1}{6}y^3 f'''(\theta_1(x,y))$ for some $\theta_1(x,y)$ between x and $x+y$. Alternatively, we can express $R(x,y)$ by another application of the mean value theorem as

$$R(x,y) = f(x+y) - f(x) - yf'(x) - \frac{1}{2}y^2 f''(x) = \frac{1}{2}y^2(f''(\theta_2(x,y)) - f''(x)),$$

for some $\theta_2(x,y)$ between x and $x+y$. Let $C = \max\{||f''||, \frac{1}{6}||f'''||\}$. Then we have the estimate $|R(x,y)| \leq C|y|^3$, as well as for every $\varepsilon > 0$ the estimate

$$|R(x,y)| \leq C(|y|^3 \mathbf{1}_{\{|y|\leq\varepsilon\}} + y^2 \mathbf{1}_{\{|y|>\varepsilon\}}) \leq Cy^2(\varepsilon + \mathbf{1}_{\{|y|>\varepsilon\}}).$$

Hence we have the following bounds on $\mathbb{E}|R(X,Y)|$:

$$\mathbb{E}|R(X,Y)| \leq C\mathbb{E}|Y|^3 \tag{12.10}$$

and

$$\mathbb{E}|R(X,Y)| \leq C(\varepsilon\mathbb{E}Y^2 + \mathbb{E}Y^2\mathbf{1}_{\{|Y|>\varepsilon\}}). \tag{12.11}$$

The proof of the Central Limit Theorem is based on the following idea. Consider a sum of independent random variables $S = \sum_{j=1}^n \xi_j$, where n is 'big'. If we replace one of the ξ_j by another random variable, then we can think of a small perturbation of S and the expectation of $f(S)$ will hardly change. This idea will be repeatedly used, all the ξ_j that sum up to S will be step by step replaced with other, *normally distributed*, random variables. We assume that the ξ_j have finite second moments and expectation zero. Let $\eta_1, \ldots, \eta_n$ be independent normal random variables, also independent of the ξ_j, with expectation zero and $\mathbb{E}\eta_j^2 = \mathbb{E}\xi_j^2$. Put $Z = \sum_{j=1}^n \eta_j$ and notice that also Z has a normal distribution with variance equal to $\sum_{j=1}^n \mathbb{E}\xi_j^2$. We are interested in $\mathbb{E}f(S) - \mathbb{E}f(Z)$. The following notation is convenient. Put $X_j = \sum_{i=1}^{j-1} \xi_i + \sum_{i=j+1}^n \eta_i$. Notice that $S = X_n + \xi_n$ and $Z = X_1 + \eta_1$ and $X_j + \xi_j = X_{j+1} + \eta_{j+1}$ for $1 \leq j \leq n-1$. These facts and an application of (12.9) yield

$$\begin{aligned}
|\mathbb{E}f(S) - \mathbb{E}f(Z)| &= |\mathbb{E}f(X_n + \xi_n) - \mathbb{E}f(X_1 + \eta_1)| \\
&= |\mathbb{E}\big(\sum_{j=1}^n f(X_j + \xi_j) - \sum_{j=1}^{n-1} f(X_j + \xi_j) \\
&\qquad + \sum_{j=2}^n f(X_j + \eta_j) - \sum_{j=1}^n f(X_j + \eta_j)\big)| \\
&= |\mathbb{E}\big(\sum_{j=1}^n f(X_j + \xi_j) - \sum_{j=1}^n f(X_j + \eta_j)\big)| \\
&\leq \sum_{j=1}^n |\mathbb{E}f(X_j + \xi_j) - \mathbb{E}f(X_j + \eta_j)| \\
&\leq \sum_{j=1}^n \mathbb{E}|R(X_j, \xi_j)| + \mathbb{E}|R(X_j, \eta_j)|.
\end{aligned} \tag{12.12}$$

Theorem 12.16 (Lindeberg-Feller Central Limit Theorem) *Assume for each $n \in \mathbb{N}$ the random variables $\xi_{n1}, \ldots, \xi_{nk_n}$ are independent with $\mathbb{E}\xi_{nj} = 0$*

and $\sum_{j=1}^{k_n} \operatorname{Var} \xi_{nj} = 1$. *Let for every* $\varepsilon > 0$

$$L_n(\varepsilon) = \sum_{j=1}^{k_n} \mathbb{E}[\xi_{nj}^2 \mathbf{1}_{\{|\xi_{nj}|>\varepsilon\}}].$$

Suppose that the Lindeberg condition *holds:* $L_n(\varepsilon) \to 0$ *as* $n \to \infty$ *for every* $\varepsilon > 0$. *Then* $S_n := \sum_{j=1}^{k_n} \xi_{nj} \stackrel{w}{\to} Z$, *where* Z *has a N(0,1) distribution.*

Proof Let $S_n = \sum_{j=1}^{k_n} \xi_{nj}$ and let η_{nj} $(j = 1, \ldots, k_n,\ n \in \mathbb{N})$ be a double array of zero mean normal random variables, independent of all the ξ_{nj}, such that also for every n the η_{nj} $(j = 1, \ldots, k_n)$ are independent and such that $\mathbb{E}\eta_{nj}^2 = \mathbb{E}\xi_{nj}^2$. Let $Z_n = \sum_{j=1}^{k_n} \eta_{nj}$. Notice that the distributions of the Z_n are all standard normal and thus $\mathbb{E}f(Z_n) = \mathbb{E}f(Z)$ for every f in $\mathcal{C}^\infty$. Recall Theorem 12.15. Take such $f \in \mathcal{C}^\infty$ and apply (12.12) to get

$$\begin{aligned} |\mathbb{E}f(S_n) - \mathbb{E}f(Z)| &= |\mathbb{E}f(S_n) - \mathbb{E}f(Z_n)| \\ &\le \sum_{j=1}^{k_n} \mathbb{E}|R(X_{nj}, \xi_{nj})| + \mathbb{E}|R(X_{nj}, \eta_{nj})|, \end{aligned} \tag{12.13}$$

with an obvious meaning of the X_{nj}. For the first error terms in (12.13) we use the estimate of (12.11) which yields $\mathbb{E}\sum_{j=1}^{k_n} \mathbb{E}|R(X_{nj}, \xi_{nj})| \le C(\varepsilon + L_n(\varepsilon))$. In view of the Lindeberg condition, this term can be made arbitrarily small. We now focus on the second error term in (12.13). Let $\sigma_{nj}^2 = \mathbb{E}\xi_{nj}^2 = \mathbb{E}\eta_{nj}^2$ and use (12.10) to obtain

$$\sum_{j=1}^{k_n} \mathbb{E}|R(X_{nj}, \eta_{nj})| \le C \sum_{j=1}^{k_n} \mathbb{E}|\eta_{nj}|^3 = C \sum_{j=1}^{k_n} \sigma_{nj}^3 \mathbb{E}|N(0,1)|^3.$$

To finish the proof, we first observe that

$$\max_j \sigma_{nj}^2 = \max_j \mathbb{E}\xi_{nj}^2 = \max_j \mathbb{E}\xi_{nj}^2 (\mathbf{1}_{\{|\xi_{nj}|\le\varepsilon\}} + \mathbf{1}_{\{|\xi_{nj}|>\varepsilon\}}) \le \varepsilon^2 + L_n(\varepsilon).$$

Hence (use $\sum_{j=1}^{k_n} \sigma_{nj}^2 = 1$)

$$\sum_{j=1}^{k_n} \sigma_{nj}^3 \le \max_j \sigma_{nj} \sum_{j=1}^{k_n} \sigma_{nj}^2 \le (\varepsilon^2 + L_n(\varepsilon))^{1/2}.$$

And, again, this term can be made arbitrarily small, because of the Lindeberg condition. □

12.3 Exercises

12.1 Prove the remaining assertions of Proposition 12.2.

12.2 Prove Proposition 12.11.

12.3 Show, using the Dominated Convergence Theorem, that (12.8) holds. Show also that all the derivatives are bounded functions.

12.4 Show that $X_n \xrightarrow{w} X$ iff $\mathbb{E}f(X_n) \to \mathbb{E}f(X)$ for all bounded uniformly continuous functions f. *Hint:* for one implication the proof of Proposition 12.3 is instructive.

12.5 Show the implication $F_n(x) \to F(x)$ for all $x \in C_F \Rightarrow \mu_n \xrightarrow{w} \mu$ of Corollary 12.5 without referring to the Skorohod representation. *First you take for given $\varepsilon > 0$ a $K > 0$ such that $F(K) - F(-K) > 1 - \varepsilon$. Approximate a function $f \in C_b(\mathbb{R})$ on the interval $[-K, K]$ by a piecewise constant function, compute the integrals of this approximating function and use the convergence of the $F_n(x)$ at continuity points of F etc.*

12.6 Suppose that $X_n \xrightarrow{w} X$ and that the collection $\{X_n, n \geq 1\}$ is uniformly integrable (you make a minor change in the definition of this notion if the X_n are defined on different probability spaces). Use the Skorohod representation to show that $X_n \xrightarrow{w} X$ implies $\mathbb{E}X_n \to \mathbb{E}X$.

12.7 Show the following variation on Fatou's lemma: if $X_n \xrightarrow{w} X$, then $\mathbb{E}|X| \leq \liminf_{n\to\infty} \mathbb{E}|X_n|$.

12.8 Show that the weak limit of a sequence of probability measures is unique.

12.9 Consider the $N(\mu_n, \sigma_n^2)$ distributions, where the μ_n are real numbers and the σ_n^2 nonnegative. Show that this family is tight iff the sequences (μ_n) and (σ_n^2) are bounded. Under what condition do we have that the $N(\mu_n, \sigma_n^2)$ distributions converge to a (weak) limit? What is this limit?

12.10 The classical Central Limit Theorem says that $\frac{1}{\sigma\sqrt{n}} \sum_{j=1}^{n} (X_j - \mu) \xrightarrow{w} N(0,1)$, if the X_j are *iid* with $\mathbb{E}X_j = \mu$ and $0 < \operatorname{Var} X_j = \sigma^2 < \infty$. Show that this follows from the Lindeberg-Feller Central Limit Theorem.

12.11 For each n we have a sequence $\xi_{n1}, \ldots, \xi_{nk_n}$ of independent random variables with $\mathbb{E}\xi_{nj} = 0$ and $\sum_{j=1}^{k_n} \operatorname{Var} \xi_{nj} = 1$. If

$$\sum_{j=1}^{k_n} \mathbb{E}|\xi_{nj}|^{2+\delta} \to 0 \text{ as } n \to \infty \text{ for some } \delta > 0, \tag{12.14}$$

then $\sum_{j=1}^{k_n} \xi_{nj} \xrightarrow{w} N(0,1)$. Show that this follows from the Lindeberg-Feller Central Limit Theorem 12.16. Condition (12.14) is called Lyapunov's condition.

12.12 Let $X_1, X_2, \ldots, X_n$ be an *iid* sequence having a distribution function F, a continuous density (w.r.t. Lebesgue measure) f. Let m be such that $F(m) = \frac{1}{2}$. Assume that $f(m) > 0$ and that n is odd, $n = 2k - 1$, say $(k = \frac{1}{2}(n+1))$.

(a) Show that m is the unique solution of the equation $F(x) = \frac{1}{2}$. We call m the *median* of the distribution of X_1.

(b) Let $X_{(1)} = \min\{X_1, \ldots, X_n\}$, $X_{(2)} = \min\{X_1, \ldots, X_n\} \setminus \{X_{(1)}\}$, etc. The resulting $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ is called the ordered sample. The *sample median* M_n of $X_1, \ldots, X_n$ is by definition $X_{(k)}$. Show that with $U_{nj} = \mathbf{1}_{\{X_j \le m + n^{-1/2}x\}}$ we have

$$\mathbb{P}(n^{1/2}(M_n - m) \le x) = \mathbb{P}(\sum_j U_{nj} \ge k).$$

(c) Let $p_n = \mathbb{E}U_{nj} = \mathbb{P}(X_j \le m + n^{-1/2}x)$, $b_n = (np_n(1-p_n))^{1/2}$, $\xi_{nj} = (U_{nj} - p_n)/b_n$, $Z_n = \sum_{j=1}^n \xi_{nj}$, $t_n = (k - np_n)/b_n$. Note that $p_n \in (0,1)$ for n large enough. Rewrite the probabilities in part (b) as $\mathbb{P}(Z_n \ge t_n)$ and show that $t_n \to t := -2xf(m)$.

(d) Show that $\mathbb{P}(Z_n \ge t) \to 1 - \Phi(t)$, where Φ is the standard normal distribution.

(e) Show that $\mathbb{P}(Z_n \ge t_n) \to \Phi(2f(m)x)$ and conclude that the *Central Limit Theorem for the sample median* holds:

$$2f(m)n^{1/2}(M_n - m) \xrightarrow{w} N(0,1).$$

12.13 Consider the situation of Proposition 12.3 and assume that F is continuous. For given $n \in \mathbb{N}$, select x_k such that $F(x_k) = \frac{k}{n}$ with $k \in \{1, \ldots, n-1\}$ and take $x_0 = -\infty$, $x_n = \infty$.

(a) Show that for all $x \in \mathbb{R}$ it holds that

$$|F_m(x) - F(x)| \le \sup_{0 \le k \le n} |F_m(x_k) - F(x_k)| + \frac{1}{n}.$$

(b) Deduce that $\sup_{x \in \mathbb{R}} |F_m(x) - F(x)| \to 0$ for $m \to \infty$.

(c) Give an example (with discontinuous F) such that the convergence of F_m to F is not even uniform on the complement of C_F.

12.14 Prove the following generalization of Proposition 12.10. A sequence (μ_n) of probability measures on $(\mathbb{R}, \mathcal{B})$ is tight iff it is weakly sequentially compact, i.e. any subsequence of (μ_n) has a further subsequence that is weakly converging to a probability measure μ. (This equivalence is a special case of Prohorov's theorem for probability measures on complete separable metric spaces.)

13 Characteristic functions

'Characteristic functions are characteristic'. In this chapter we will explain what this statement means and why it is true. We develop some theory for characteristic functions, primarily with the goal to apply it to prove by other means a Central Limit Theorem.

13.1 Definition and first properties

Let X be a random variable defined on $(\Omega, \mathcal{F}, \mathbb{P})$. As we have seen in Chapter 3, X induces a probability measure on $(\mathbb{R}, \mathcal{B})$, the law or distribution of X, denoted e.g. by $\mathbb{P}^X$ or by μ. This probability measure in turn, determines the distribution function F of X. We have seen in Theorem 3.10 that, conversely, F also determines $\mathbb{P}^X$. Hence distribution functions on $\mathbb{R}$ and probability measures on $(\mathbb{R}, \mathcal{B})$ are in bijective correspondence. In this chapter we develop another such correspondence. We start with a definition.

Definition 13.1 Let μ be a probability measure on $(\mathbb{R}, \mathcal{B})$. Its characteristic function $\phi : \mathbb{R} \to \mathbb{C}$ is defined by

$$\phi(u) = \int_{\mathbb{R}} e^{\mathrm{i}ux} \mu(\mathrm{d}x). \tag{13.1}$$

Whenever needed, we write ϕ_μ instead of ϕ to express the dependence on μ.

Note that in this definition we integrate a complex valued function. By splitting a complex valued function $f = g + \mathrm{i}h$ into its real part g and imaginary part h, we define $\int f \,\mathrm{d}\mu := \int g \,\mathrm{d}\mu + \mathrm{i} \int h \,\mathrm{d}\mu$. For integrals of complex valued functions, previously shown theorems are, mutatis mutandis, true. For instance, one has $|\int f \,\mathrm{d}\mu| \leq \int |f| \,\mathrm{d}\mu$, where $|\cdot|$ denotes the norm of a complex number.

If X is a random variable with distribution μ, then it follows from Proposition 4.27 applied to $h(x) = \exp(\mathrm{i}ux)$, that ϕ_μ can alternatively be expressed by $\phi(u) = \mathbb{E} \exp(\mathrm{i}uX)$. There are many random variables sharing the same distribution μ, they can even be defined on different underlying probability spaces. We also adopt the notation ϕ_X to indicate that we are dealing with the characteristic function of the random variable X.

Before we give some elementary properties of characteristic functions, we look at a special case. Suppose that X admits a density f with respect to Lebesgue measure. Then

$$\phi_X(u) = \int_{\mathbb{R}} e^{\mathrm{i}ux} f(x) \,\mathrm{d}x. \tag{13.2}$$

Analysts define for $f \in \mathcal{L}^1(\mathbb{R}, \mathcal{B}, \lambda)$ (λ Lebesgue measure) the Fourier transform $\hat{f}$ by

$$\hat{f}(u) = \int_{\mathbb{R}} e^{-\mathrm{i}ux} f(x) \,\mathrm{d}x,$$

sometimes also by dividing this expression by $\sqrt{2\pi}$. What we thus see, is the equality $\phi_X(u) = \hat{f}(-u)$.

Proposition 13.2 *Let $\phi = \phi_X$ be the characteristic function of some random variable X. The following hold true.*

(i) $\phi(0) = 1$, $|\phi(u)| \leq 1$, *for all* $u \in \mathbb{R}$

(ii) ϕ *is uniformly continuous on* $\mathbb{R}$.

(iii) $\phi_{aX+b}(u) = \phi_X(au)e^{\mathrm{i}ub}$.

(iv) ϕ *is real valued and symmetric around zero, if* X *and* $-X$ *have the same distribution.*

(v) If X *and* Y *are independent, then* $\phi_{X+Y}(u) = \phi_X(u)\phi_Y(u)$.

(vi) If $\mathbb{E}|X|^k < \infty$, *then* $\phi \in C^k(\mathbb{R})$ *and* $\phi^{(k)}(0) = \mathrm{i}^k\mathbb{E}X^k$.

Proof Properties (i), (iii) and (iv) are trivial. Consider (ii). Fixing $u \in \mathbb{R}$, we consider $\phi(u+t) - \phi(u)$ for $t \to 0$. We have

$$\begin{aligned} |\phi(u+t) - \phi(u)| &= |\int (\exp(\mathrm{i}(u+t)x) - \exp(\mathrm{i}ux))\,\mu(\,\mathrm{d}x)| \\ &\leq \int |\exp(\mathrm{i}tx) - 1|\,\mu(\,\mathrm{d}x). \end{aligned}$$

The functions $x \mapsto \exp(\mathrm{i}tx) - 1$ converge to zero pointwise for $t \to 0$ and are bounded by 2. The result thus follows from dominated convergence.

Property (v) follows from the product rule for expectations of independent random variables, Proposition 4.43.

Finally, property (vi) for $k = 1$ follows by an application of the Dominated Convergence Theorem and the inequality $|e^{\mathrm{i}x} - 1| \leq |x|$, for $x \in \mathbb{R}$. The other cases can be treated similarly. □

Remark 13.3 The implication that ϕ_{X+Y} is the product of ϕ_X and ϕ_Y for independent random variables X and Y cannot simply be reversed, as shown by the following example. Let X have a standard Cauchy distribution. From Exercise 13.4 one sees that $\phi_X(u) = \exp(-|u|)$. With $Y = X$ one computes that $\exp(-2|u|) = \phi_{2X}(u) = \phi_{X+Y}(u) = \phi_X(u)\phi_Y(u)$, although in this case X and Y are obviously not independent. Similarly, if $X_1 = \ldots = X_n = X$, one has $\phi_{nX}(u) = \phi_X(u)^n$. In fact, the property that $\phi_{nX} = \phi_X^n$ for all $n \geq 1$ characterizes the (scaled) Cauchy distributions, see Exercise 13.5.

For a distribution function F we define $\tilde{F}$ by $\tilde{F}(x) = \frac{1}{2}(F(x) + F(x-))$, where $F(x-) = \lim_{y \uparrow x} F(y)$. Note that $\tilde{F}$ coincides at those x where F is continuous. At the points where F is not (left-)continuous, $\tilde{F}$ is neither left- nor right-continuous. One always has $F(x-) \leq \tilde{F}(x) \leq F(x)$, $F(x) = \tilde{F}(x+)$ and $F(x-) = \tilde{F}(x-)$. In particular, $\tilde{F}$ completely determines F. The following inversion theorem for characteristic functions is similar to Fourier inversion. Note that the integration interval in (13.3) is symmetric around zero. This is essential.

Theorem 13.4 *Let F be a distribution function and ϕ its characteristic function. Then, for all $a < b$*

$$\lim_{T\to\infty} \frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-\mathrm{i}ua} - e^{-\mathrm{i}ub}}{\mathrm{i}u} \phi(u)\,\mathrm{d}u = \tilde{F}(b) - \tilde{F}(a). \tag{13.3}$$

Proof Let $a < b$. We compute, using Fubini's theorem which we will justify below,

$$\begin{aligned}
\Phi_T &:= \frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-\mathrm{i}ua} - e^{-\mathrm{i}ub}}{\mathrm{i}u} \phi(u)\,\mathrm{d}u \qquad (13.4)\\
&= \frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-\mathrm{i}ua} - e^{-\mathrm{i}ub}}{\mathrm{i}u} \int_{\mathbb{R}} e^{\mathrm{i}ux}\,\mu(\mathrm{d}x)\,\mathrm{d}u\\
&= \frac{1}{2\pi} \int_{\mathbb{R}} \int_{-T}^{T} \frac{e^{-\mathrm{i}ua} - e^{-\mathrm{i}ub}}{\mathrm{i}u} e^{\mathrm{i}ux}\,\mathrm{d}u\,\mu(\mathrm{d}x)\\
&= \frac{1}{2\pi} \int_{\mathbb{R}} \int_{-T}^{T} \frac{e^{\mathrm{i}(x-a)u} - e^{\mathrm{i}(x-b)u}}{\mathrm{i}u}\,\mathrm{d}u\,\mu(\mathrm{d}x) \qquad (13.5)\\
&=: \int_{\mathbb{R}} E_T(x)\,\mu(\mathrm{d}x)
\end{aligned}$$

Application of Fubini's theorem is justified as follows. First, the integrand in (13.5) is bounded by $b - a$, because $|e^{\mathrm{i}x} - e^{\mathrm{i}y}| \le |x - y|$ for all $x, y \in \mathbb{R}$. Second, the product measure $\lambda \times \mu$ on $[-T, T] \times \mathbb{R}$ is finite.

By splitting the integrand of $E_T(x)$ into its real and imaginary part, we see that the imaginary part vanishes and we are left with the real expression

$$\begin{aligned}
E_T(x) &= \frac{1}{2\pi} \int_{-T}^{T} \frac{\sin(x-a)u - \sin(x-b)u}{u}\,\mathrm{d}u\\
&= \frac{1}{2\pi} \int_{-T}^{T} \frac{\sin(x-a)u}{u}\,\mathrm{d}u - \frac{1}{2\pi} \int_{-T}^{T} \frac{\sin(x-b)u}{u}\,\mathrm{d}u\\
&= \frac{1}{2\pi} \int_{-T(x-a)}^{T(x-a)} \frac{\sin v}{v}\,\mathrm{d}v - \frac{1}{2\pi} \int_{-T(x-b)}^{T(x-b)} \frac{\sin v}{v}\,\mathrm{d}v.
\end{aligned}$$

The function g given by $g(s,t) = \int_s^t \frac{\sin y}{y}\,\mathrm{d}y$ is continuous in (s,t). Hence it is bounded on any compact subset of $\mathbb{R}^2$. Moreover, $g(s,t) \to \pi$ as $s \to -\infty$ and $t \to \infty$. Hence g, as a function on $\mathbb{R}^2$, is bounded in s, t. We conclude that also $E_T(x)$ is bounded, as a function of T and x, a first ingredient to apply the Dominated Convergence Theorem to (13.5), since μ is a finite measure. The second ingredient is $E(x) := \lim_{T\to\infty} E_T(x)$. From Exercise 5.9 we deduce that

$$\int_0^\infty \frac{\sin \alpha x}{x}\,\mathrm{d}x = \mathrm{sgn}(\alpha)\frac{\pi}{2}.$$

By comparing the location of x relative to a and b, we use the value of the latter

integral to obtain

$$E(x) = \begin{cases} 1 & \text{if } a < x < b, \\ \frac{1}{2} & \text{if } x = a \text{ or } x = b, \\ 0 & \text{else.} \end{cases}$$

We thus get, using the Dominated Convergence Theorem again,

$$\Phi_T \to \mu(a,b) + \frac{1}{2}\mu(\{a,b\}) = \tilde{F}(b) - \tilde{F}(a).$$

□

Corollary 13.5 *If μ and ν are two probability measures on $(\mathbb{R}, \mathcal{B})$ whose characteristic functions are the same, then they coincide.*

Proof Exercise 13.1. □

The content of Corollary 13.5 explains why characteristic functions are called characteristic, there is a bijective correspondence between probability measures and characteristic functions.

Theorem 13.6 *If the characteristic function ϕ of a probability measure μ on $(\mathbb{R}, \mathcal{B})$ belongs to $\mathcal{L}^1(\mathbb{R}, \mathcal{B}, \lambda)$, then μ admits a density f w.r.t. the Lebesgue measure λ. Moreover, f is continuous.*

Proof Define

$$f(x) = \frac{1}{2\pi} \int_{\mathbb{R}} e^{-\mathrm{i}ux} \phi(u) \, \mathrm{d}u. \tag{13.6}$$

Since $|\phi|$ has a finite integral, f is well defined for every x. Observe that f is real valued, because $\overline{\phi(u)} = \phi(-u)$. An easy application of the Dominated Convergence Theorem shows that f is continuous. Note first that the limit of the integral expression in (13.3) is equal to the (Lebesgue) integral $\frac{1}{2\pi} \int \frac{e^{-\mathrm{i}ua} - e^{-\mathrm{i}ub}}{\mathrm{i}u} \phi(u) \, \mathrm{d}u$, again because of dominated convergence. We use Fubini's theorem to compute for $a < b$

$$\begin{aligned} \int_a^b f(x) \, \mathrm{d}x &= \frac{1}{2\pi} \int_a^b \int_{\mathbb{R}} e^{-\mathrm{i}ux} \phi(u) \, \mathrm{d}u \, \mathrm{d}x \\ &= \frac{1}{2\pi} \int_{\mathbb{R}} \phi(u) \int_a^b e^{-\mathrm{i}ux} \, \mathrm{d}x \, \mathrm{d}u \\ &= \frac{1}{2\pi} \int_{\mathbb{R}} \phi(u) \frac{e^{-\mathrm{i}ua} - e^{-\mathrm{i}ub}}{\mathrm{i}u} \, \mathrm{d}u \\ &= F(b) - F(a), \end{aligned}$$

for every a and b, because of Theorem 13.4 and because $\int_a^b f(x) \, \mathrm{d}x$ is continuous in a and b. It also follows that f must be nonnegative and so it is a density. □

Remark 13.7 Note the duality between the expressions (13.2) and (13.6). Apart from the presence of the minus sign in the integral and the factor 2π in the denominator in (13.6), the transformations $f \mapsto \phi$ and $\phi \mapsto f$ are similar.

The characteristic function ϕ of a probability measure μ on $(\mathbb{R}^k, \mathcal{B}(\mathbb{R}^k))$ is defined by the k-dimensional analogue of (13.1). We have with $u, x \in \mathbb{R}^k$, $\langle \cdot, \cdot \rangle$ the standard inner product,

$$\phi(u) = \int_{\mathbb{R}^k} e^{\mathrm{i}\langle u,x\rangle} \mu(\mathrm{d}x).$$

Like in the real case, also here probability measures are uniquely determined by their characteristic functions. The proof of this statement can be given as a multi-dimensional version of Exercise 13.2. As a consequence we have the following characterization of independent random variables.

Proposition 13.8 *Let $X = (X_1, \ldots, X_k)$ be a k-dimensional random vector. Then $X_1, \ldots, X_k$ are independent random variables iff $\phi_X(u) = \prod_{i=1}^k \phi_{X_i}(u_i)$, $\forall u = (u_1, \ldots, u_k) \in \mathbb{R}^k$.*

Proof If the X_i are independent, the statement about the characteristic functions is proved in the same way as Proposition 13.2 (v). If the characteristic function ϕ_X factorizes as stated, the result follows by the uniqueness property of characteristic functions. □

Remark 13.9 Let $k = 2$ in the above proposition. If $X_1 = X_2$ as in Remark 13.3, then we don't have $\phi_X(u) = \phi_{X_1}(u_1)\phi_{X_2}(u_2)$ for every u_1, u_2 (you check!), in agreement with the fact that X_1 and X_2 are not independent. But for the special choice $u_1 = u_2$ this product relation holds true.

13.2 Characteristic functions and weak convergence

The first result says that weak convergence of probability measures implies pointwise convergence of their characteristic functions.

Proposition 13.10 *Let $\mu, \mu_1, \mu_2, \ldots$ be probability measures on $(\mathbb{R}, \mathcal{B})$ and let $\phi, \phi_1, \phi_2, \ldots$ be their characteristic functions. If $\mu_n \xrightarrow{w} \mu$, then $\phi_n(u) \to \phi(u)$ for every $u \in \mathbb{R}$.*

Proof Consider for fixed u the function $f(x) = e^{\mathrm{i}ux}$. It is obviously bounded and continuous and we obtain straight from Definition 12.1 that $\mu_n(f) \to \mu(f)$. But $\mu_n(f) = \phi_n(u)$. □

Proposition 13.11 *Let $\mu_1, \mu_2, \ldots$ be probability measures on $(\mathbb{R}, \mathcal{B})$. Let $\phi_1, \phi_2, \ldots$ be the corresponding characteristic functions. Assume that the sequence (μ_n) is tight and that for all $u \in \mathbb{R}$ the limit $\phi(u) := \lim_{n\to\infty} \phi_n(u)$ exists. Then there exists a probability measure μ on $(\mathbb{R}, \mathcal{B})$ such that $\phi = \phi_\mu$ and $\mu_n \xrightarrow{w} \mu$.*

Proof Since (μ_n) is tight we use Proposition 12.10 to deduce that there exists a weakly converging subsequence (μ_{n_k}) with a probability measure as limit. Call this limit μ. From Proposition 13.10 we know that $\phi_{n_k}(u) \to \phi_\mu(u)$ for all u. Hence we must have $\phi_\mu = \phi$. We will now show that any convergent subsequence of (μ_n) has μ as a limit. Suppose that there exists a subsequence $(\mu_{n'_k})$ with limit μ'. Then $\phi_{n'_k}(u)$ converges to $\phi_{\mu'}(u)$ for all u. But, since $(\mu_{n'_k})$ is a subsequence of the original sequence, by assumption the corresponding $\phi_{n'_k}(u)$ must converge to $\phi(u)$ for all u. Hence we conclude that $\phi_{\mu'} = \phi_\mu$ and then $\mu' = \mu$.

Suppose that the whole sequence (μ_n) does not converge to μ. Then there must exist a function $f \in C_b(\mathbb{R})$ such that $\mu_n(f)$ does not converge to $\mu(f)$. So, there is $\varepsilon > 0$ such that for some subsequence (n'_k) we have

$$|\mu_{n'_k}(f) - \mu(f)| > \varepsilon. \tag{13.7}$$

Using Proposition 12.10, the sequence $(\mu_{n'_k})$ has a further subsequence $(\mu_{n''_k})$ that has a limit probability measure μ''. By the same argument as above (convergence of the characteristic functions) we conclude that $\mu''(f) = \mu(f)$. Therefore $\mu_{n''_k}(f) \to \mu(f)$, which contradicts (13.7). □

Characteristic functions are a tool to give a rough estimate of the tail probabilities of a random variable, useful to establish tightness of a sequence of probability measures. To that end we will use the following lemma. Check first that $\int_{-a}^{a}(1 - \phi(u))\,\mathrm{d}u \in \mathbb{R}$ for every $a > 0$.

Lemma 13.12 *Let a random variable X have distribution μ and characteristic function ϕ. Then for every $K > 0$*

$$P(|X| > 2K) \leq K \int_{-1/K}^{1/K} (1 - \phi(u))\,\mathrm{d}u. \tag{13.8}$$

Proof It follows from Fubini's theorem and $\int_{-a}^{a} e^{\mathrm{i}ux}\,\mathrm{d}u = 2\frac{\sin ax}{x}$ that

$$\begin{aligned}
K \int_{-1/K}^{1/K} (1 - \phi(u))\,\mathrm{d}u &= K \int_{-1/K}^{1/K} \int (1 - e^{\mathrm{i}ux})\,\mu(\mathrm{d}x)\,\mathrm{d}u \\
&= \int K \int_{-1/K}^{1/K} (1 - e^{\mathrm{i}ux})\,\mathrm{d}u\,\mu(\mathrm{d}x) \\
&= 2 \int (1 - \frac{\sin x/K}{x/K})\,\mu(\mathrm{d}x) \\
&\geq 2 \int_{|x/K|>2} (1 - \frac{\sin x/K}{x/K})\,\mu(\mathrm{d}x) \\
&\geq \mu([-2K, 2K]^c).
\end{aligned}$$

since $\frac{\sin x}{x} \leq \frac{1}{2}$ for $x > 2$. □

The following theorem is known as *Lévy's continuity theorem*.

Theorem 13.13 *Let $\mu_1, \mu_2, \ldots$ be a sequence of probability measures on $(\mathbb{R}, \mathcal{B})$ and $\phi_1, \phi_2, \ldots$ the corresponding characteristic functions. Assume that for all $u \in \mathbb{R}$ the limit $\phi(u) := \lim_{n\to\infty} \phi_n(u)$ exists. If ϕ is continuous at zero, then there exists a probability measure μ on $(\mathbb{R}, \mathcal{B})$ such that $\phi = \phi_\mu$ and $\mu_n \xrightarrow{w} \mu$.*

Proof We will show that under the present assumptions, the sequence (μ_n) is tight. We will use Lemma 13.12. Let $\varepsilon > 0$. Since ϕ is continuous at zero, the same holds for $\overline{\phi}$, and there is $\delta > 0$ such that $|\phi(u) + \phi(-u) - 2| < \varepsilon$ if $|u| < \delta$. Notice that $\phi(u) + \phi(-u)$ is real-valued and bounded from above by 2. Hence $2\int_{-\delta}^{\delta}(1 - \phi(u))\,\mathrm{d}u = \int_{-\delta}^{\delta}(2 - \phi(u) - \phi(-u))\,\mathrm{d}u \in [0, 2\delta\varepsilon)$.

By the convergence of the characteristic functions (which are bounded), the Dominated Convergence Theorem implies that for big enough n, $n \geq N$ say,

$$\int_{-\delta}^{\delta}(1 - \phi_n(u))\,\mathrm{d}u < \int_{-\delta}^{\delta}(1 - \phi(u))\,\mathrm{d}u + \delta\varepsilon.$$

Hence, by taking $n \geq N$ we have

$$\int_{-\delta}^{\delta}(1 - \phi_n(u))\,\mathrm{d}u < 2\delta\varepsilon.$$

It now follows from Lemma 13.12 that for $n \geq N$ and $K = 1/\delta$

$$\begin{aligned}\mu_n([-2K, 2K]^c) &\leq \frac{1}{\delta}\int_{-\delta}^{\delta}(1 - \phi_n(u))\,\mathrm{d}u \\ &< 2\varepsilon.\end{aligned}$$

We conclude that $(\mu_n)_{n\geq N}$ is tight. Apply Proposition 13.11. □

Corollary 13.14 *Let $\mu, \mu_1, \mu_2, \ldots$ be probability measures on $(\mathbb{R}, \mathcal{B})$ and ϕ, $\phi_1, \phi_2, \ldots$ be their corresponding characteristic functions. Then $\mu_n \xrightarrow{w} \mu$ if and only if $\phi_n(u) \to \phi(u)$ for all $u \in \mathbb{R}$.*

Proof If $\phi_n(u) \to \phi(u)$ for all $u \in \mathbb{R}$, then we can apply Theorem 13.13. Because ϕ, being a characteristic function, is continuous at zero. Hence there is a probability to which the μ_n weakly converge. But since the $\phi_n(u)$ converge to $\phi(u)$, the limiting probability measure must be μ. The converse statement we have encountered as Proposition 13.10. □

Remark 13.15 Corollary 13.14 has an obvious counterpart for probability measures on $(\mathbb{R}^d, \mathcal{B}^d)$. The formulation is left to the reader.

13.3 The Central Limit Theorem revisited

The proof of the Theorem 12.16 that we present in this section is based on an application of Lévy's continuity theorem and additional properties of characteristic functions, the first ones are contained in the following lemma.

Lemma 13.16 *Let X be a random variable with $\mathbb{E}X^2 < \infty$ and with characteristic function ϕ. Then*

$$|\phi(u) - 1| \le \mathbb{E}\min\{2, |uX|\},$$

$$|\phi(u) - 1 - \mathrm{i}u\mathbb{E}X| \le \mathbb{E}\min\{2|u||X|, \frac{1}{2}u^2X^2\}$$

and

$$|\phi(u) - 1 - \mathrm{i}u\mathbb{E}X + \frac{1}{2}u^2\mathbb{E}X^2| \le \mathbb{E}\min\{u^2X^2, \frac{1}{6}|u|^3|X|^3\}.$$

Proof Let $x \in \mathbb{R}$. Then $|e^{\mathrm{i}x} - 1| \le 2$ and $|e^{\mathrm{i}x} - 1| = |\int_0^x \mathrm{i}e^{\mathrm{i}y}\,\mathrm{d}y| \le |x|$. Hence $|e^{\mathrm{i}x} - 1| \le \min\{2, |x|\}$. Since

$$e^{\mathrm{i}x} - 1 - \mathrm{i}x = \mathrm{i}\int_0^x (e^{\mathrm{i}y} - 1)\,\mathrm{d}y,$$

and

$$e^{\mathrm{i}x} - 1 - \mathrm{i}x + \frac{1}{2}x^2 = -\int_0^x \int_0^y (e^{\mathrm{i}t} - 1)\,\mathrm{d}t\,\mathrm{d}y,$$

we arrive at the inequalities $|e^{\mathrm{i}x} - 1 - \mathrm{i}x| \le \min\{2|x|, \frac{1}{2}x^2\}$ and $|e^{\mathrm{i}x} - 1 - \mathrm{i}x + \frac{1}{2}x^2| \le \min\{x^2, |x|^3/6\}$. Replacing x with uX and taking expectations yields the assertions. □

We are now ready to give the announced alternative proof of Theorem 12.16.

Proof of Theorem 12.16 Let $\phi_{nj}(u) = \mathbb{E}\exp(\mathrm{i}u\xi_{nj})$, $\phi_n(u) = \mathbb{E}\exp(\mathrm{i}uS_n)$. Because of independence we have

$$\phi_n(u) = \prod_{j=1}^{k_n} \phi_{nj}(u).$$

First we show that

$$\sum_{j=1}^{k_n} (\phi_{nj}(u) - 1) \to -\frac{1}{2}u^2. \tag{13.9}$$

We write

$$\sum_{j=1}^{k_n} (\phi_{nj}(u) - 1) = \sum_{j=1}^{k_n} (\phi_{nj}(u) - 1 + \frac{1}{2}u^2\mathbb{E}\xi_{nj}^2) - \sum_{j=1}^{k_n} \frac{1}{2}u^2\mathbb{E}\xi_{nj}^2.$$

The last term gives the desired limit, so it suffices to show that the first term converges to zero. By virtue of Lemma 13.16, we can bound its absolute value by

$$\sum_{j=1}^{k_n} \mathbb{E}\min\{u^2\xi_{nj}^2, \frac{1}{6}|u|^3|\xi_{nj}|^3\}. \tag{13.10}$$

But

$$\mathbb{E}\min\{u^2\xi_{nj}^2, \frac{1}{6}|u|^3|\xi_{nj}|^3\} \leq \frac{1}{6}|u|^3\varepsilon\mathbb{E}\xi_{nj}^2\mathbf{1}_{\{|\xi_{nj}|\leq\varepsilon\}} + u^2\mathbb{E}\xi_{nj}^2\mathbf{1}_{\{|\xi_{nj}|>\varepsilon\}}.$$

Hence we get that the expression in (13.10) is majorized by

$$\frac{1}{6}|u|^3\varepsilon\sum_{j=1}^{k_n}\mathbb{E}\xi_{nj}^2 + u^2L_n(\varepsilon) = \frac{1}{6}|u|^3\varepsilon + u^2L_n(\varepsilon),$$

which tends to $\frac{1}{6}|u|^3\varepsilon$. Since ε is arbitrary, we have proved (13.9). It then also follows that

$$\exp(\sum_{j=1}^{k_n}(\phi_{nj}(u)-1)) \to \exp(-\frac{1}{2}u^2). \tag{13.11}$$

Recall that $u \mapsto \exp(-\frac{1}{2}u^2)$ is the characteristic function of $N(0,1)$. Hence, by application of Lévy's continuity theorem and (13.11), we are finished as soon as we have shown that

$$\prod_{j=1}^{k_n}\phi_{nj}(u) - \exp(\sum_{j=1}^{k_n}(\phi_{nj}(u)-1)) \to 0. \tag{13.12}$$

The displayed difference is in absolute value less than

$$\sum_{j=1}^{k_n}|\phi_{nj}(u) - \exp(\phi_{nj}(u)-1)|, \tag{13.13}$$

because of the following elementary result: if a_i and b_i are complex numbers with norm less than or equal to one, then

$$|\prod_{i=1}^{n}a_i - \prod_{i=1}^{n}b_i| \leq \sum_{i=1}^{n}|a_i - b_i|.$$

To apply this result we have to understand that the complex numbers involved indeed have norm less than or equal to one. For the $\phi_{nj}(u)$ this is one of the basic properties of characteristic functions. But it turns out that $\exp(\phi_{nj}(\cdot)-1)$ is a characteristic function as well (see Exercise 13.3).

Let $M_n(u) = \max_j|\phi_{nj}(u)-1|$. Now we use the inequality $|e^z - 1 - z| \leq |z|^2e^{|z|}$ (which easily follows from a Taylor expansion) with $z = \phi_{nj}(u)-1$ to bound (13.13) by

$$\sum_{j=1}^{k_n}|\phi_{nj}(u)-1|^2\exp(|\phi_{nj}(u)-1|) \leq M_n(u)e^{M_n(u)}\sum_{j=1}^{k_n}|\phi_{nj}(u)-1|.$$

From Lemma 13.16, second assertion, we get

$$\sum_{j=1}^{k_n} |\phi_{nj}(u) - 1| \leq \frac{1}{2}u^2 \sum_{j=1}^{k_n} \mathbb{E}\xi_{nj}^2 = \frac{1}{2}u^2.$$

On the other hand, we have $\max_j \mathbb{E}\xi_{nj}^2 \leq \varepsilon^2 + L_n(\varepsilon)$. Hence

$$\max_j \mathbb{E}\xi_{nj}^2 \to 0. \tag{13.14}$$

Then by Lemma 13.16 and Jensen's inequality

$$M_n(u) = \max_j |\phi_{nj}(u) - 1| \leq \max_j |u|\mathbb{E}|\xi_{nj}| \leq |u|(\max_j \mathbb{E}\xi_{nj}^2)^{1/2} \to 0.$$

This proves (13.12) and hence it completes the proof of the theorem. □

Remark 13.17 The Lindeberg condition in the theorem is almost necessary. One can show that if (13.14) holds and if the weak convergence as in the theorem takes place, then also the Lindeberg condition is satisfied.

13.4 Exercises

13.1 Prove Corollary 13.5.

13.2 Let μ and ν be probability measures on $(\mathbb{R}, \mathcal{B}))$ having corresponding characteristic functions ϕ_μ and ϕ_ν.

(a) Show that $\int_{\mathbb{R}} \exp(-\mathrm{i}uy)\phi_\mu(y)\,\nu(\mathrm{d}y) = \int_{\mathbb{R}} \phi_\nu(x-u)\,\mu(\mathrm{d}x)$.

(b) Assume that ν is the $N(0, \frac{1}{\sigma^2})$ distribution, then $\phi_\nu(u) = \exp(-\frac{1}{2}u^2/\sigma^2)$. Let f_{σ^2} be the density of the $N(0, \sigma^2)$ distribution. Show that

$$\frac{1}{2\pi}\int_{\mathbb{R}} \exp(-\mathrm{i}uy)\phi_\mu(y)\exp(-\frac{1}{2}\sigma^2 y^2)\,\mathrm{d}y = \int_{\mathbb{R}} f_{\sigma^2}(u-x)\,\mu(\mathrm{d}x),$$

and show that the right hand side gives the *density* of $X + Y$, where X has distribution μ, Y has the $N(0, \sigma^2)$ distribution and X and Y are independent (see also Exercise 5.6).

(c) Write $Y = \sigma Z$, with Z having the standard normal distribution (X and Z independent). It follows that ϕ_μ and σ determine the distribution of $X + \sigma Z$ for all $\sigma > 0$. Show that ϕ_μ uniquely determines μ. (This gives an alternative proof of the assertion of Corollary 13.5).

13.3 Let $X_1, X_2, \ldots$ be a sequence of *iid* random variables and N a Poisson(λ) distributed random variable, independent of the X_n. Put $Y = \sum_{n=1}^{N} X_n$. Let ϕ be the characteristic function of the X_n and ψ the characteristic function of Y. Show that $\psi = \exp(\lambda\phi - \lambda)$.

13.4 Verify the formulas for the characteristic functions in each of the following cases.

(a) $\phi_{N(0,1)}(u) = \exp(-\frac{1}{2}u^2)$. *Hint:* Show that $\phi_{N(0,1)}$ is a solution to $\dot{\phi}(u) = -u\phi(u)$.
(b) $\phi_{N(\mu,\sigma^2)}(u) = \exp(\mathrm{i}u\mu - \frac{1}{2}\sigma^2u^2)$.
(c) If X has an exponential distribution with parameter λ, then $\phi_X(u) = \lambda/(\lambda - \mathrm{i}u)$.
(d) If X has a Cauchy distribution, then $\phi_X(u) = \exp(-|u|)$. Show also that $\bar{X}_n := \frac{1}{n}\sum_{k=1}^n X_k$ has a Cauchy distribution, if the X_k are *iid* and Cauchy distributed. Hence, in this case the averages $\bar{X}_n$ have the same distribution as X_1 for every n.

13.5 Let ϕ be a real characteristic function with the property that $\phi(nu) = \phi(u)^n$ for all $u \in \mathbb{R}$ and $n \in \mathbb{N}$. Show that for some $\alpha \geq 0$ it holds that $\phi(u) = \exp(-\alpha|u|)$. Let X have characteristic function $\phi(u) = \exp(-\alpha|u|)$. If $\alpha > 0$, show that X admits the density $x \mapsto \frac{\alpha}{\pi}\frac{1}{\alpha^2+x^2}$. What is the distribution of X if $\alpha = 0$?

13.6 Let X_n have a $\mathrm{Bin}(n, \lambda/n)$ distribution (for $n > \lambda$). Show that $X_n \xrightarrow{w} X$, where X has a Poisson(λ) distribution.

13.7 Let X and Y be independent, assume that Y has a $N(0,1)$ distribution. Let $\sigma > 0$. Let ϕ be the characteristic function of X: $\phi(u) = \mathbb{E}\exp(\mathrm{i}uX)$.
(a) Show that $Z = X + \sigma Y$ has density $p(z) = \frac{1}{\sigma\sqrt{2\pi}}\mathbb{E}\exp(-\frac{1}{2\sigma^2}(z-X)^2)$.
(b) Show that $p(z) = \frac{1}{2\pi\sigma}\int \phi(-y/\sigma)\exp(\mathrm{i}yz/\sigma - \frac{1}{2}y^2)\,\mathrm{d}y$.

13.8 Let $X, X_1, X_2, \ldots$ be a sequence of random variables and Y a $N(0,1)$-distributed random variable independent of that sequence. Let ϕ_n be the characteristic function of X_n and ϕ that of X. Let p_n be the density of $X_n + \sigma Y$ and p the density of $X + \sigma Y$.
(a) If $\phi_n \to \phi$ pointwise, then $p_n \to p$ pointwise. Invoke Exercise 13.7 and the dominated convergence theorem to show this.
(b) Let $f : \mathbb{R} \to \mathbb{R}$ be bounded by B. Show that $|\mathbb{E}f(X_n+\sigma Y) - \mathbb{E}f(X+\sigma Y)| \leq 2B\int(p(z) - p_n(z))^+\,dz$.
(c) Show that $|\mathbb{E}f(X_n + \sigma Y) - \mathbb{E}f(X + \sigma Y)| \to 0$ (with f bounded) if $\phi_n \to \phi$ pointwise.
(d) Prove Corollary 13.14: $X_n \xrightarrow{w} X$ iff $\phi_n \to \phi$ pointwise.

13.9 Let Y be a random variable with a Gamma$(t,1)$ distribution, so it has density $\frac{1}{\Gamma(t)}y^{t-1}e^{-y}\mathbf{1}_{\{y>0\}}$, where $\Gamma(t) = \int_0^\infty y^{t-1}e^{-y}\,\mathrm{d}y$ for $t > 0$. Put $X_t = \frac{Y-t}{\sqrt{t}}$.
(a) Show that X_t has a density on $(-\sqrt{t}, \infty)$ given by

$$f_t(x) = \frac{\sqrt{t}}{\Gamma(t)}(x\sqrt{t} + t)^{t-1}e^{-(x\sqrt{t}+t)}.$$

(b) Show that the characteristic function $\phi_t(u) = \mathbb{E}e^{\mathrm{i}uX_t}$ of X_t is given by

$$\phi_t(u) = e^{-\mathrm{i}u\sqrt{t}}\frac{1}{(1 - \frac{\mathrm{i}u}{\sqrt{t}})^t}$$

and conclude that $\phi_t(u) \to e^{-\frac{1}{2}u^2}$ as $t \to \infty$.

(c) Show that

$$\frac{t^{t-\frac{1}{2}}e^{-t}}{\Gamma(t)} = \frac{1}{2\pi}\int_{-\infty}^{\infty} \phi_t(u)\,\mathrm{d}u.$$

(d) Prove Stirling's formula

$$\lim_{t\to\infty} \frac{\Gamma(t)}{\sqrt{2\pi}e^{-t}t^{t-\frac{1}{2}}} = 1.$$

13.10 Let X be a random variable defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and let $\mathcal{G}$ be a sub-σ-algebra of $\mathcal{F}$. Let for $u \in \mathbb{R}$ the random variable $\hat{\phi}(u)$ be a version of $\mathbb{E}[e^{\mathrm{i}uX}|\mathcal{G}]$. Note that we actually have a map $(u, \omega) \mapsto \hat{\phi}(u, \omega)$.

(a) Show that the map $u \mapsto \hat{\phi}(u)$ is continuous in $\mathcal{L}^1$ ($\mathbb{E}|\hat{\phi}(u+h) - \hat{\phi}(u)| \to 0$ for $h \to 0$).

(b) Show that we can take the $\hat{\phi}(u)$ such that $u \mapsto \hat{\phi}(u, \omega)$ is continuous on a set of probability one.

(c) Suppose that there exists a function $\phi : \mathbb{R} \to \mathcal{C}$ such that $\phi(u)$ is a version of $\mathbb{E}[e^{\mathrm{i}uX}|\mathcal{G}]$ for each $u \in \mathbb{R}$. Show that ϕ is the characteristic function of X and that $\mathcal{G}$ and $\sigma(X)$ are independent.

13.11 Let $X, X_1, X_2, \ldots$ be a sequence of d-dimensional random vectors. Show, using Remark 13.15, that $X_n \xrightarrow{w} X$ iff for all d-dimensional column vectors a one has $a^\top X_n \xrightarrow{w} a^\top X$. (*This equivalence is known as the Cramér-Wold device and it reduces d-dimensional weak convergence to 1-dimensional weak convergence.*) Suppose that X has a multivariate normal distribution with mean vector μ and covariance matrix Σ. Show that (the notation should be obvious) $X_n \xrightarrow{w} N(\mu, \Sigma)$ iff $a^\top X_n \xrightarrow{w} N(a^\top\mu, a^\top\Sigma a)$, for all $a \in \mathbb{R}^d$.

13.12 Prove the *Riemann-Lebesgue lemma*: If a random variable X admits a density w.r.t. the Lebesgue measure, then $\lim_{u\to\pm\infty} \phi_X(u) = 0$. Give a counterexample to this limit result when a density doesn't exist.

13.13 Let (X_n) be a sequence of random k-vectors such that $X_n \xrightarrow{\mathcal{L}^2} X$ for some random vector X.

(a) Show that $\mathbb{E}X_n \to \mathbb{E}X$ and $\mathrm{Cov}(X_n) \to \mathrm{Cov}(X)$.

(b) Assuming that all X_n are (multivariate) normal, show that also X is (multivariate) normal.

14 Brownian motion

This chapter is devoted to showing the existence of Brownian motion as a well defined mathematical object. By definition, a stochastic process $X = \{X_t : t \geq 0\}$ with time index $[0, \infty)$ on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is a collection of random variables $\{\omega \mapsto X_t(\omega) : t \geq 0\}$ on this probability space.

Definition 14.1 A stochastic process W is called Brownian motion or Wiener process, if it satisfies the following properties.

(i) $W_0 = 0$,

(ii) the *sample paths* $t \mapsto W_t(\omega)$ are continuous functions for every $\omega \in \Omega$,

(iii) the *increments* $W_t - W_s$ have a normal $N(0, t-s)$ distribution for all $t \geq s \geq 0$,

(iv) for any collection $0 \leq t_0 < \cdots < t_k$, the increments $W_{t_1} - W_{t_0}, \ldots, W_{t_k} - W_{t_{k-1}}$ are independent.

The topic of this chapter is thus to show that an appropriate $(\Omega, \mathcal{F}, \mathbb{P})$ exists and that one can define a Brownian motion on it. As it turns out, the choice $\Omega = \{\omega \in C[0, \infty) : \omega(0) = 0\}$ of real valued continuous functions on $[0, \infty)$ is a good one. If one defines $W_t(\omega) = \omega(t)$ for $\omega \in \Omega$, then automatically every $t \mapsto W_t(\omega) = \omega(t)$ is continuous, which already settles properties (i) and (ii).

Brownian motion is perhaps the most fundamental stochastic process with a continuous time set. The style of this chapter is to some extent expository. Some results that we have proved in previous chapters for finite dimensional spaces have a generalization to infinite dimensional spaces. In the present chapter these generalizations are sometimes presented without proof.

14.1 The space $C[0, \infty)$

In this section we summarize some facts concerning the space $C[0, \infty)$. For $x_1, x_2 \in C[0, \infty)$ we define

$$\rho(x_1, x_2) = \sum_{n \geq 1} 2^{-n} (\max\{|x_1(t) - x_2(t)| : 0 \leq t \leq n\} \wedge 1). \tag{14.1}$$

Then ρ defines a metric on $C[0, \infty)$ (which we use throughout this chapter) and one can show that the metric space $(C[0, \infty), \rho)$ is complete and separable.

Later on we need the relatively compact subsets of $C[0, \infty)$. To describe these we introduce the modulus of continuity m^T. For each $x \in C[0, \infty)$, $T, \delta > 0$ we define

$$m^T(x, \delta) = \max\{|x(t) - x(s)| : s, t \in [0, T], |s - t| \leq \delta\}. \tag{14.2}$$

It holds that $m^T(\cdot, \delta)$ is continuous and $\lim_{\delta \downarrow 0} m^T(x, \delta) = 0$ for each x and T. The following characterization is known as the Arzelà-Ascoli theorem.

Theorem 14.2 *A set A in $C[0,\infty)$ is relatively compact (has compact closure) iff (i) $\sup\{|x(0)| : x \in A\} < \infty$ and (ii) the functions in A are equicontinuous: $\lim_{\delta\downarrow 0} \sup\{m^T(x,\delta) : x \in A\} = 0$ for all $T > 0$.*

Proof Assume (i) and (ii). It is sufficient to prove the existence of a sequence of functions $(x_n) \subset A$ that is uniformly convergent on every interval $[0,T]$. To that end it suffices to show that for every $T \in \mathbb{N}$ the collection $A_T := \{x : [0,T] \to \mathbb{R} |\, x \in A\}$ contains a uniformly convergent sequence (note the subtle difference and verify that the last statement is correct). Under conditions (i) and (ii), A_T is uniformly bounded and we can apply the same technique as in the first part of the proof of Theorem 12.7. There exists a sequence $(x_n) \subset A_T$ such that $(x_n(q))$ converges for every rational $q \in [0,T]$. For any given $\varepsilon > 0$, there exists $N(q,\varepsilon)$ such that $|x_n(q) - x_m(q)| < \varepsilon/3$ for $m, n \geq N$. Since $\lim_{\delta\downarrow 0} m^T(x,\delta) = 0$ for each x and T, there also exist $\delta > 0$ such that $|x(s) - x(t)| < \varepsilon/3$ if $|s-t| < \delta$ and $s, t \in [0,T]$. For given $t \in [0,T]$, choose rational $q = q(t)$ in $[0,m]$ such that $|t - q| < \delta$. Since $[0,T]$ is bounded we can select this q from a finite set of rationals q_i, $i \in I$. Then $N := \max\{N(q_i,\varepsilon) : i \in I\}$ is finite and for $n, m \geq N$ one has

$$|x_n(t) - x_m(t)| \leq |x_n(t) - x_n(q)| + |x_n(q) - x_m(q)| + |x_m(q) - x_m(t)| < \varepsilon,$$

from which the assertion follows.

Conversely, we assume that A has compact closure $\bar{A}$. Since the map $x \mapsto x(0)$ is continuous, it follows that $\{x(0) : x \in A\}$ is bounded, property (i).

Fix $\varepsilon > 0$. By continuity of $m^T(\cdot,\delta)$, the sets $E_\delta := \{x \in C[0,\infty) : m^T(x,\delta) \geq \varepsilon\}$ $(\delta, \varepsilon > 0)$ are closed and hence the intersections $K_\delta = \bar{A} \cap E_\delta$ are all compact. Since $\bigcap_{\delta>0} E_\delta = \emptyset$, the finite intersection property of the nested compact sets K_δ implies that there exists $\delta_0 > 0$ such that $K_{\delta_0} = \emptyset$, which shows property (ii). □

Cylinder sets of $C[0,\infty)$ have the typical form $\{x : (x(t_1), \ldots, x(t_k)) \in A\}$, where $A \in \mathcal{B}(\mathbb{R}^k)$ for some $k \geq 1$ and $t_1, \ldots, t_k \in [0,\infty)$. A finite dimensional projection on $(C[0,\infty), \rho)$ is by definition of the following type: $\pi_{t_1,\ldots,t_k}(x) = (x(t_1), \ldots, x(t_k))$, where the t_i are nonnegative real numbers. It is easy to see that any finite dimensional projection is continuous ($\mathbb{R}^k$ is endowed with the ordinary metric). Note that cylinder sets are inverse images under finite dimensional projections of Borel sets of $\mathbb{R}^k$ $(k \geq 1)$. Let $\mathcal{C}$ be the collection of all cylinder sets and $\mathcal{B}$ the Borel σ-algebra on $C[0,\infty)$ induced by the metric ρ.

Let $(\Omega, \mathcal{F})$ be a measurable space. A map $X : \Omega \to C[0,\infty)$ is called a random element of $C[0,\infty)$ if it is $\mathcal{F}/\mathcal{B}$-measurable. It then follows that $\pi_{t_1,\ldots,t_k} \circ X$ is a random vector in $\mathbb{R}^k$, for any finite dimensional projection $\pi_{t_1,\ldots,t_k}$, and it is usually denoted by $(X_{t_1}, \ldots, X_{t_k})$. One can prove that $\mathcal{B} = \sigma(\mathcal{C})$ and thus that X is a random element of $C[0,\infty)$, if all X_t are real random variables. Moreover, if $\mathbb{P}$ is a probability measure on $(\Omega, \mathcal{F})$ and X a random element of $C[0,\infty)$, then the distribution $\mathbb{P}^X$ of X on $(C[0,\infty), \mathcal{B})$ is completely determined by the distributions of all k-tuples $(X_{t_1}, \ldots, X_{t_k})$ on $\mathbb{R}^k$ $(k \geq 1, t_i \in [0,\infty))$.

14.2 Weak convergence on $C[0,\infty)$

Let (S,ρ) be a metric space and $\mu,\mu^1,\mu^2,\ldots$ be probability measures on the Borel σ-algebra $\mathcal{B}(S)$. Like in the real case we say that μ^n converges weakly to μ (notation $\mu^n \xrightarrow{w} \mu$) iff for all $f \in C_b(S)$ one has $\lim \mu^n(f) = \mu(f)$. If $X, X^1, X^2, \ldots$ are random variables defined on probability spaces $(\Omega,\mathcal{F},\mathbb{P})$ and $(\Omega^n,\mathcal{F}^n,\mathbb{P}^n)$ $(n \geq 1)$ with values in one and the same (S,ρ), we say that X^n converges in distribution to X ($X^n \xrightarrow{w} X$) if the laws μ^n of X^n converge weakly to the law μ of X, equivalently, iff $\mathbb{E}f(X^n) \to \mathbb{E}f(X)$ for all $f \in C_b(S)$.

We need a generalization of tightness as given in Definition 12.8 that is applicable in the present context. Since any interval $[-M,M] \subset \mathbb{R}$ is compact, the following is reasonable. A family of probability measures Π on $\mathcal{B}(S)$ is called tight if for every $\varepsilon > 0$, there is a compact subset K of S such that $\inf\{\mu(K) : \mu \in \Pi\} > 1-\varepsilon$. One can show that any single probability measure on $\mathcal{B}(S)$ is tight if (S,ρ) is a separable and complete metric space (a Polish space). A family of random variables with values in a metric space is called tight if the family of their distributions is tight. Like in the real case, see Proposition 12.10 and Exercise 12.14, (but much harder to prove here) there is equivalence between relative compactness (in this context it means that every sequence in a set of probability measures has a weakly converging subsequence) and tightness, known as Prohorov's theorem.

Theorem 14.3 *A family Π of probability measures on a complete separable metric space is tight iff it is relatively (weakly) compact.*

We will also need the following perturbation result.

Proposition 14.4 *Let $X^1, X^2, \ldots$ and $Y^1, Y^2, \ldots$ be random sequences in a metric space (S,ρ) and defined on a single probability space $(\Omega,\mathcal{F},\mathbb{P})$. If $X^n \xrightarrow{w} X$ and $\rho(Y^n, X^n) \xrightarrow{\mathbb{P}} 0$, then $Y^n \xrightarrow{w} X$.*

If we take $S = C[0,\infty)$ with the metric ρ of the previous section, we get the following 'stochastic version' of the Arzelà-Ascoli theorem.

Theorem 14.5 *Let $\mu^1, \mu^2, \ldots$ be a sequence of probability measures on the space $(C[0,\infty),\mathcal{B})$. This sequence is tight iff*

$$\lim_{\lambda\uparrow\infty} \sup\{\mu^n(\{x : |x(0)| > \lambda\}) : n \geq 1\} = 0 \tag{14.3}$$

and

$$\lim_{\delta\downarrow 0} \sup\{\mu^n(\{x : m^T(x,\delta) > \varepsilon\}) : n \geq 1\} = 0, \forall T, \varepsilon > 0. \tag{14.4}$$

Proof If the sequence is tight, the result is a straightforward application of Theorem 14.2. For every $\varepsilon > 0$ we can find a compact K such that $\inf_n \mu^n(K) > 1-\varepsilon$. But then we can find $\lambda > 0$ such that for all $x \in K$ we have $|x(0)| < \lambda$ and we can similarly find for given $T > 0$ and $\eta > 0$ a $\delta_0 > 0$ such that for all $0 < \delta < \delta_0$ we have on K that $m^T(x,\delta) < \eta$.

Conversely, assume (14.3) and (14.4) and let $\varepsilon, T > 0$, T integer, be given. Choose λ_T such that $\sup_n \mu^n(\{x : |x(0)| > \lambda_T\}) \leq \varepsilon 2^{-T-1}$. For each $k \geq 1$ we can also find δ_k such that $\sup_n \mu^n(\{x : m^T(x, \delta_k) > 1/k\}) \leq \varepsilon 2^{-T-k-1}$. Notice that the sets $A_{T,k} = \{x : m^T(x, \delta_k) \leq 1/k\}$ and $A_{T,0} = \{x : |x(0)| \leq \lambda_T\}$ are closed and so is their intersection over both k and (integer) T, call it K. From Theorem 14.2 we obtain that K has compact closure and it is thus compact itself. Finally we compute $\mu^n(K^c) \leq \sum_{T\geq 1} \mu^n(A^c_{T,0}) + \sum_{T\geq 1}\sum_{k\geq 1} \mu^n(A^c_{T,k}) \leq \varepsilon$. □

We have seen that any finite dimensional projection is continuous. Hence, if $X, X^1, X^2, \ldots$ are random elements of $(C[0,\infty), \mathcal{B})$ and if we assume that $X^n \xrightarrow{w} X$, then also $(X^n_{t_1}, \ldots, X^n_{t_k})$ considered as random elements in $\mathbb{R}^k$ converge in distribution to $(X_{t_1}, \ldots, X_{t_k})$. This is then true for any finite set of t_i's and we say that all finite dimensional distributions of the X^n converge weakly. The converse does not hold in general, unless one assumes tightness.

Theorem 14.6 *Let $X^1, X^2, \ldots$ be random elements of $C[0,\infty)$. Assume that their collection $\{\mu^1, \mu^2, \ldots\}$ of distributions is tight and that all finite dimensional distributions of the X^n converge weakly. Then there exists a probability measure μ on $(C[0,\infty), \mathcal{B})$ such that $\mu^n \xrightarrow{w} \mu$.*

Proof Every subsequence of (μ^n) is tight as well and thus has a convergent subsequence. Different subsequences have to converge to the same limit, call it μ, since the finite dimensional distributions corresponding to these sequences converge. Hence, if (μ^n) has a limit, it must be μ. Suppose therefore that the μ^n don't converge. Then there is bounded and continuous f and an $\varepsilon > 0$ such that $|\mu^{n_k}(f) - \mu(f)| > \varepsilon$ along a subsequence (μ^{n_k}). No further subsequence of this can have μ as a limit which contradicts what we just showed. □

14.3 An invariance principle

Throughout this section we work with a real valued *iid* sequence $\xi_1, \xi_2, \ldots$ with zero mean and variance $\sigma^2 \in (0,\infty)$ defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let $S_k = \sum_{i=1}^k \xi_i$ and for each integer n and $t \geq 0$

$$X^n_t = \frac{1}{\sigma\sqrt{n}}(S_{[nt]} + (nt - [nt])\xi_{[nt]+1}). \tag{14.5}$$

The processes X^n have continuous paths and can be considered as random elements of $C[0,\infty)$. Notice that the increments $X^n_t - X^n_s$ of each X^n over non-overlapping intervals (s,t) with $s = \frac{k}{n}, t = \frac{l}{n}$, $k < l$ integers, are independent. Since for these values of t and s we also have $\mathrm{Var}\,(X^n_t - X^n_s) = t - s$, the central limit theorem should be helpful to understand the limit behavior.

Theorem 14.7 *Let $0 = t_0 < t_1 < \cdots < t_k$. Then the k-vectors of increments $(X^n_{t_1} - X^n_{t_0}, \ldots, X^n_{t_k} - X^n_{t_{k-1}})$ converge in distribution to a random vector $(N_1, \ldots, N_k)$ with independent elements N_j having a $N(0, t_j - t_{j-1})$ distribution.*

Proof Since the term in (14.5) with the $\xi_{[nt]}$ tends to zero in probability, we can ignore it as a consequence of Proposition 14.4. But then the conclusion for each of the increments separately follows from the classical Central Limit Theorem of Exercise 12.10. Check that for each n the $S_{[nt_j]} - S_{[nt_{j-1}]}$ $(j = 1, \ldots, k)$ are independent to complete the proof. □

Denote by μ^n the law of X^n. Note that $\mu^n(\{x \in C[0,\infty) : x(0) = 0\}) = 0$. We have the following important result, whose proof is deferred to the next section.

Theorem 14.8 *The sequence of probability measures (μ^n) is tight.*

Combining Theorems 14.6 and 14.7 one obtains

Theorem 14.9 *There exists a probability measure μ^* on $(C[0,\infty), \mathcal{B})$ such that $\mu^n \xrightarrow{w} \mu^*$.*

Let $\Omega = \{\omega \in C[0,\infty) : \omega(0) = 0\}$, then Ω belongs to $\mathcal{B}$ and $\mu^*(\Omega) = 1$. Let W denotes the *coordinate mapping process* on Ω i.e. W is defined by $W_t(\omega) = \omega(t)$ for all $t \geq 0$. Under the measure μ^* (μ^* is the law of W) this process has independent increments over non-overlapping intervals (s,t) and these increments have a $N(0, t-s)$ distribution. Since by definition W is a random element of $(C[0,\infty), \mathcal{B})$, W is thus a Wiener process and the measure μ^* is called Wiener measure.

We can rephrase Theorem 14.9 as

Theorem 14.10 *The processes X^n of this section converge in distribution to a Wiener process W.*

Both Theorems 14.9 and 14.10 are known as Donsker's invariance principle. What we have done in this section can be summarized by saying that we have shown the existence of a Wiener process and we have given a *Functional* Central Limit Theorem.

14.4 The proof of Theorem 14.8

Consider the process S_n of Section 14.3. To prove Theorem 14.8 we use the following results.

Lemma 14.11 *Let $\gamma > 0$, $n \geq 1$, $N \geq n$ and $\eta \geq \sigma\sqrt{2(n-1)}$. The following inequalities are valid.*

$$\mathbb{P}(\max_{j \leq n} |S_j| > \gamma) \quad \leq \quad 2\mathbb{P}(|S_n| > \gamma - \eta) \tag{14.6}$$

$$\mathbb{P}(\max_{\substack{1 \leq j \leq n \\ 0 \leq k \leq N}} |S_{j+k} - S_k| > \gamma) \quad \leq \quad (\frac{N}{n} + 2)\mathbb{P}(\max_{j \leq n} |S_j| > \gamma/3). \tag{14.7}$$

Proof Assume $\eta < \gamma$. Let $\tau = \min\{j : |S_j| > \gamma\}$. Then we have to consider $\mathbb{P}(\tau \leq n)$. Split this probability up as

$$\mathbb{P}(\tau \leq n, |S_n| > \gamma - \eta) + \mathbb{P}(\tau \leq n, |S_n| \leq \gamma - \eta) \tag{14.8}$$

and work on the second probability. It can be written as $\sum_{j=1}^{n-1} \mathbb{P}(\tau = j, |S_n| \leq \gamma - \eta)$ and each of the probabilities in the sum is less than or equal to $\mathbb{P}(\tau = j, |S_n - S_j| > \eta) = \mathbb{P}(\tau = j)\mathbb{P}(|S_n - S_j| > \eta)$. The second factor is by Chebychev's inequality less than $\frac{1}{\eta^2}(n-1)\sigma^2 \leq \frac{1}{2}$, by the assumption on η. Therefore $\mathbb{P}(\tau \leq n, |S_n| \leq \gamma - \eta) \leq \frac{1}{2}\mathbb{P}(\tau \leq n-1)$. From (14.8), we then get $\mathbb{P}(\tau \leq n) \leq \mathbb{P}(|S_n| > \gamma - \eta) + \frac{1}{2}\mathbb{P}(\tau \leq n)$ and the inequality (14.6) follows.

To prove (14.7) we argue as follows. Let $m = [\frac{N}{n}]$ and consider the 'intervals' $\{pn, \ldots, (p+1)n - 1\}$, for $p = 0, \ldots, m$. N belongs to the last one. Consider j and k for which the maximum is bigger than γ. If $k + j$ belongs to the same interval as k, the one starting with pn, say, we certainly have $|S_{np} - S_k| > \gamma/3$ or $|S_{np} - S_{k+j}| > \gamma/3$ and so in this case there is $p \leq m$ such that $\max_{j \leq n} |S_{np} - S_j| > \gamma/3$. If $k + j$ lies in the interval starting with $(p+1)n$, we must have $|S_{np} - S_k| > \gamma/3$ or $|S_{n(p+1)} - S_{k+j}| > \gamma/3$ or $|S_{n(p+1)} - S_{np}| > \gamma/3$. Both cases are contained in the event $\bigcup_{0 \leq p \leq m+1}\{\max_{j \leq n} |S_{np} - S_{np+j}| > \gamma/3\}$, whose probability is less than or equal to $\sum_{p=0}^{m+1} \mathbb{P}(\max_{j \leq n} |S_{np} - S_{np+j}| > \gamma/3)$. By the *iid* assumption all probabilities in this sum are equal to the first one and thus the sum is equal to $(m+2)\mathbb{P}(\max_{j \leq n} |S_j| > \gamma/3)$, which yields the result. □

With the aid of this lemma we prove Theorem 14.8 as follows. According to Theorem 14.5 it is sufficient to show that

$$\lim_{\delta \downarrow 0} \sup_{n \geq 1} \mathbb{P}\Big(\max_{\substack{|s-t| \leq \delta \\ 0 \leq t,s \leq T}} |X_t^n - X_s^n| > \varepsilon \Big) = 0 \text{ for all } T, \varepsilon > 0. \tag{14.9}$$

But since we only need tightness for all but finitely many n, we can as well replace the 'sup' by a 'lim sup'. Let $Y_t = \sigma\sqrt{n} X_{t/n}^n$. Each of the probabilities in (14.9) is less than

$$\mathbb{P}\Big(\max_{\substack{|s-t| \leq [n\delta]+1 \\ 0 \leq t,s \leq [nT]+1}} |Y_t - Y_s| > \varepsilon\sigma\sqrt{n} \Big).$$

But, since Y is piecewise linear between the integer values of its arguments, the max is attained at integer numbers. Hence we consider

$$\mathbb{P}\Big(\max_{\substack{0 \leq j \leq [n\delta]+1 \\ 0 \leq k \leq [nT]+1}} |S_{j+k} - S_k| > \varepsilon\sigma\sqrt{n} \Big).$$

Now we apply inequality (14.7) and bound this probability by

$$\Big(\frac{[nT]+1}{[n\delta]+1} + 2\Big)\mathbb{P}\Big(\max_{j \leq [n\delta]+1} |S_j| > \varepsilon\sigma\sqrt{n}/3 \Big). \tag{14.10}$$

In view of (14.6) (take $\eta = \sigma\sqrt{2[n\delta]}$) the probability in (14.10) is less than

$$\mathbb{P}(|S_{[n\delta]+1}| > \varepsilon\sigma\sqrt{n}/3 - \sigma\sqrt{2[n\delta]}).$$

Now we apply the central limit theorem: $\frac{1}{\sigma\sqrt{[n\delta]}} S_{[n\delta]+1} \stackrel{w}{\to} Z$, where Z has a $N(0,1)$ distribution. So for $n \to \infty$ the last probability tends to $\mathbb{P}(|Z| > \frac{\varepsilon}{3\sqrt{\delta}} - \sqrt{2})$ which is less than $\frac{\delta^2}{(\varepsilon/3-\sqrt{2\delta})^4}\mathbb{E}Z^4$. Hence the lim sup in (14.10) for $n \to \infty$ is less than $\frac{T\delta+2\delta^2}{(\varepsilon/3-\sqrt{2\delta})^4}\mathbb{E}Z^4$, from which we obtain (14.9). □

14.5 Another proof of existence of Brownian motion

In this section we prove the existence of Brownian motion by a different method. The technique that is used in the existence proof is based on linear interpolation properties for continuous functions.

Let a continuous function $f : [0,1] \to \mathbb{R}$ be given with $f(0) = 0$. We will construct an approximation scheme of f, consisting of continuous piecewise linear functions. To that end we make use of the dyadic numbers in $[0,1]$. Let for each $n \in \mathbb{N}$ the set D_n be equal to $\{k2^{-n} : k = 0, \ldots, 2^n\}$. The dyadic numbers in $[0,1]$ are then the elements of $\cup_{n=1}^{\infty} D_n$. To simplify the notation we write t_k^n for $k2^{-n} \in D_n$.

The interpolation starts with $f_0(t) \equiv tf(1)$ and then we define the other f_n recursively. Suppose f_{n-1} has been constructed by prescribing the values at the points t_k^{n-1} for $k = 0, \ldots, 2^{n-1}$ and by linear interpolation between these points. Look now at the points t_k^n for $k = 0, \ldots, 2^n$. For the even integers $2k$ we take $f_n(t_{2k}^n) = f_{n-1}(t_k^{n-1})$. Then for the odd integers $2k-1$ we define $f_n(t_{2k-1}^n) = f(t_{2k-1}^n)$. We complete the construction of f_n by linear interpolation between the points t_k^n. Note that for $m \geq n$ we have $f_m(t_k^n) = f(t_k^n)$.

The above interpolation scheme can be represented in a more compact way by using the so-called *Haar* functions H_k^n. These are defined as follows. $H_1^0(t) \equiv 1$ and for each n we define H_k^n for $k \in I(n) = \{1, \ldots, 2^{n-1}\}$ by

$$H_k^n(t) = \begin{cases} \frac{1}{2\sigma_n} & \text{if } t_{2k-2}^n \leq t < t_{2k-1}^n \\ -\frac{1}{2\sigma_n} & \text{if } t_{2k-1}^n \leq t < t_{2k}^n \\ 0 & \text{elsewhere} \end{cases} \tag{14.11}$$

where $\sigma_n = 2^{-\frac{1}{2}(n+1)}$. Next we put $S_k^n(t) = \int_0^t H_k^n(u)\,\mathrm{d}u$. Note that the support of S_k^n is the interval $[t_{2k-2}^n, t_{2k}^n]$ and that the graphs of the S_k^n are tent shaped with peaks of height σ_n at t_{2k-1}^n.

Next we will show how to cast the interpolating scheme in such a way that the Haar functions, or rather the *Schauder* functions S_k^n, are involved. Observe that not only the S_k^n are tent shaped, but also the consecutive differences $f_n - f_{n-1}$ on each of the intervals $(t_{k-1}^{n-1}, t_k^{n-1})$! Hence they are multiples of each other and to express the interpolation in terms of the S_k^n we only have to determine the multiplication constant. The height of the peak of $f_n - f_{n-1}$ on $(t_{k-1}^{n-1}, t_k^{n-1})$ is

the value η_k^n at the midpoint t_{2k-1}^n. So $\eta_k^n = f(t_{2k-1}^n) - \frac{1}{2}(f(t_{k-1}^{n-1}) + f(t_k^{n-1}))$. Then we have for $t \in (t_{2k-2}^n, t_{2k}^n)$ the simple formula

$$f_n(t) - f_{n-1}(t) = \frac{\eta_k^n}{\sigma_n} S_k^n(t),$$

and hence we get for all t

$$f_n(t) = f_{n-1}(t) + \sum_{k \in I(n)} \frac{\eta_k^n}{\sigma_n} S_k^n(t). \tag{14.12}$$

Summing Equation (14.12) over n leads with $I(0) = \{1\}$ to the following representation of f_n on the whole interval $[0, 1]$:

$$f_n(t) = \sum_{m=0}^{n} \sum_{k \in I(m)} \frac{\eta_k^m}{\sigma_m} S_k^m(t). \tag{14.13}$$

Proposition 14.12 *Let f be a continuous function on $[0, 1]$. With the f_n defined by (14.13) we have $||f - f_n|| \to 0$, where $|| \cdot ||$ denotes the sup norm.*

Proof Let $\varepsilon > 0$ and choose N such that we have $|f(t) - f(s)| \leq \varepsilon$ as soon as $|t - s| < 2^{-N}$. It is easy to see that then $|\eta_k^n| \leq \varepsilon$ if $n \geq N$. On the interval $[t_{2k-2}^n, t_{2k}^n]$ we have that

$$|f(t) - f_n(t)| \leq |f(t) - f(t_{2k-1}^n)| + |f_n(t_{2k-1}^n) - f_n(t)| \leq \varepsilon + \eta_k^n \leq 2\varepsilon.$$

This bound holds on any of the intervals $[t_{2k-2}^n, t_{2k}^n]$. Hence $||f - f_n|| \to 0$. □

Corollary 14.13 *For arbitrary $f \in C[0, 1]$ we have*

$$f = \sum_{m=0}^{\infty} \sum_{k \in I(m)} \frac{\eta_k^m}{\sigma_m} S_k^m, \tag{14.14}$$

where the infinite sum converges in the sup norm.

The method we use to prove existence of Brownian motion (and of a suitable probability space on which it is defined) is a kind of converse of the interpolation scheme above. We will define what is going to be Brownian motion recursively on the time interval $[0, 1]$ by attributing values at the dyadic numbers in $[0, 1]$. A crucial part of the construction is the following fact. Supposing that we have shown that Brownian motion exists we consider the random variables $W(s)$ and $W(t)$ with $s < t$. Draw independent of these random variables a random variable ξ with a standard normal distribution and define $Z = \frac{1}{2}(W(s)+W(t))+\frac{1}{2}\sqrt{t-s}\,\xi$. Then Z also has a normal distribution, whose expectation is zero and whose variance can be shown to be $\frac{1}{2}(t + s)$ (this is Exercise 14.7). Hence Z has the same distribution as $W(\frac{1}{2}(t + s))$! This fact lies at the heart of the construction of Brownian motion by a kind of 'inverse interpolation' that we

will present below.

Let, as above, $I(0) = \{1\}$ and $I(n)$ be the set $\{1, \ldots, 2^{n-1}\}$ for $n \geq 1$. Take a double sequence of independent standard normally distributed random variables ξ_k^n that are all defined on some probability space Ω with $k \in I(n)$ and $n \in \mathbb{N} \cup \{0\}$, see Theorem 3.16 or Theorem 5.14. With the aid of these random variables we are going to construct a sequence of continuous processes W^n as follows. Let, also as above, $\sigma_n = 2^{-\frac{1}{2}(n+1)}$. Put

$$W^0(t) = t\xi_1^0.$$

For $n \geq 1$ we get the following recursive scheme

$$W^n(t_{2k}^n) = W^{n-1}(t_k^{n-1}) \tag{14.15}$$

$$W^n(t_{2k-1}^n) = \frac{1}{2}\left(W^{n-1}(t_{k-1}^{n-1}) + W^{n-1}(t_k^{n-1})\right) + \sigma_n \xi_k^n. \tag{14.16}$$

For other values of t we define $W^n(t)$ by linear interpolation between the values of W^n at the points t_k^n. We use the Schauder functions, similar to (14.13), for a compact expression of the random functions W^n. We have

$$W^n(t) = \sum_{m=0}^{n} \sum_{k \in I(m)} \xi_k^m S_k^m(t). \tag{14.17}$$

Theorem 14.14 *For almost all ω the functions $t \mapsto W^n(\omega, t)$ converge uniformly to a continuous function $t \mapsto W(\omega, t)$ and the process $W : (\omega, t) \to W(\omega, t)$ is Brownian motion on $[0, 1]$.*

Proof We start with the following result. If Z has a standard normal distribution and $x > 0$, then (Exercise 14.8)

$$\mathbb{P}(|Z| > x) \leq \sqrt{\frac{2}{\pi}} \frac{1}{x} \exp(-\frac{1}{2}x^2). \tag{14.18}$$

Let $\beta_n = \max_{k \in I(n)} |\xi_k^n|$. Then $b_n := \mathbb{P}(\beta_n > n) \leq 2^{n-1}\sqrt{\frac{2}{\pi}}\frac{1}{n}\exp(-\frac{1}{2}n^2)$. Observe that $\sum_n b_n$ is convergent and that hence in virtue of the Borel-Cantelli lemma $\mathbb{P}(\limsup\{\beta_n > n\}) = 0$. Hence we can find a subset $\tilde{\Omega}$ of Ω with $\mathbb{P}(\tilde{\Omega}) = 1$, such that for all $\omega \in \tilde{\Omega}$ there exists a natural number $n(\omega)$ with the property that all $|\xi_k^n(\omega)| \leq n$ if $n \geq n(\omega)$ and $k \in I(n)$. Consequently, for $p > n \geq n(\omega)$ we have

$$\sup_t |W^n(\omega, t) - W^p(\omega, t)| \leq \sum_{m=n+1}^{\infty} m\sigma_m < \infty. \tag{14.19}$$

This shows that the sequence $W^n(\omega, \cdot)$ with $\omega \in \tilde{\Omega}$ is Cauchy in $C[0, 1]$, so that it converges to a continuous limit, which we call $W(\omega, \cdot)$. For ω's not in $\tilde{\Omega}$ we

define $W(\omega,\cdot)=0$. So we now have continuous functions $W(\omega,\cdot)$ for all ω with the property $W(\omega,0)=0$.

As soon as we have verified properties (iii) and (iv) of Definition 14.1 we know that W is a Brownian motion. We will verify these two properties at the same time by showing that all increments $\Delta_j := W(t_j)-W(t_{j-1})$ with $t_j > t_{j-1}$ are independent $N(0,t_j-t_{j-1})$ distributed random variables. Thereto we will prove that the characteristic function $\mathbb{E}\exp(\mathrm{i}\sum_j \lambda_j\Delta_j)$ is equal to $\exp(-\frac{1}{2}\sum_j \lambda_j^2(t_j - t_{j-1}))$.

The Haar functions form a Complete Orthonormal System of $L^2[0,1]$ (see Exercise 14.9). Hence every function $f\in L^2[0,1]$ has the representation $f = \sum_{n,k}\langle f,H_k^n\rangle H_k^n = \sum_{n=0}^{\infty}\sum_{k\in I(n)}\langle f,H_k^n\rangle H_k^n$, where $\langle\cdot\cdot\rangle$ denotes the inner product of $L^2[0,1]$ and where the infinite sum is convergent in $L^2[0,1]$. As a result we have for any two functions f and g in $L^2[0,1]$ the Parseval identity $\langle f,g\rangle = \sum_{n,k}\langle f,H_k^n\rangle\langle g,H_k^n\rangle$. Taking the specific choice $f=1_{[0,t]}$ and $g=1_{[0,s]}$ results in $\langle 1_{[0,t]},H_k^n\rangle = S_k^n(t)$ and $t\wedge s = \langle 1_{[0,t]},1_{[0,s]}\rangle = \sum_{n,k} S_k^n(t)S_k^n(s)$.

We use this property to compute the limit of $\mathrm{Cov}(W^n(t),W^n(s))$. We have

$$\begin{aligned}\mathrm{Cov}(W^n(t),W^n(s)) &= \mathbb{E}\left(\sum_{m=0}^{n}\sum_{k\in I(m)}\xi_k^m S_k^m(t)\sum_{p=0}^{n}\sum_{l\in I(p)}\xi_l^p S_l^p(s)\right)\\ &= \sum_{m=0}^{n}\sum_{k\in I(m)} S_k^m(t)S_k^m(s)),\end{aligned}$$

which converges to $s\wedge t$ for $n\to\infty$. Since for all fixed t we have $W^n(t)\to W(t)$ a.s., we have $\mathbb{E}\exp(\mathrm{i}\sum_j\lambda_j\Delta_j) = \lim_{n\to\infty}\mathbb{E}\exp(\mathrm{i}\sum_j\lambda_j\Delta_j^n)$ with $\Delta_j^n = W^n(t_j)-W^n(t_{j-1})$. We compute

$$\begin{aligned}\mathbb{E}\exp(\mathrm{i}\sum_j\lambda_j\Delta_j^n) &= \mathbb{E}\exp(\mathrm{i}\sum_{m\le n}\sum_k\sum_j\lambda_j\xi_k^m(S_k^m(t_j)-S_k^m(t_{j-1})))\\ &= \prod_{m\le n}\prod_k\mathbb{E}\exp(\mathrm{i}\sum_j\lambda_j\xi_k^m(S_k^m(t_j)-S_k^m(t_{j-1})))\\ &= \prod_{m\le n}\prod_k\exp(-\frac{1}{2}(\sum_j\lambda_j(S_k^m(t_j)-S_k^m(t_{j-1}))^2)\end{aligned}$$

Working out the double product, we get in the exponential the sum over the four variables $i,j,m\le n,k$ of $-\frac{1}{2}\lambda_i\lambda_j$ times

$$S_k^m(t_j)S_k^m(t_i) - S_k^m(t_{j-1})S_k^m(t_i) - S_k^m(t_j)S_k^m(t_{i-1}) + S_k^m(t_{j-1})S_k^m(t_{i-1})$$

and this sum converges to $\exp(-\frac{1}{2}\sum_j\lambda_j^2(t_j-t_{j-1}))$ as $n\to\infty$. This completes the proof. □

Having constructed Brownian motion on $[0,1]$, we proceed to show that it also exists on $[0,\infty)$. Take for each $n\in\mathbb{N}$ a probability space $(\Omega_n,\mathcal{F}_n,\mathbb{P}_n)$ that supports a Brownian motion W^n on $[0,1]$. Consider then $\Omega=\prod_n\Omega_n$, $\mathcal{F}=$

$\prod_n \mathcal{F}_n$ and $\mathbb{P} = \prod_n \mathbb{P}_n$. Note that these Brownian motions are independent by construction. Let $\omega = (\omega_1, \omega_2, \ldots)$ and define then

$$W(\omega, t) = \sum_{n \geq 0} 1_{[n,n+1)}(t) \left(\sum_{k=1}^{n} W_k(\omega_k, 1) + W_{n+1}(\omega_{n+1}, t-n) \right). \quad (14.20)$$

This obviously defines a process with continuous paths and for all t the random variable $W(t)$ is the sum of independent normal random variables. It is not hard to see that the thus defined process has independent increments. It is immediate that $\mathbb{E}W(t) = 0$ and that $\operatorname{Var} W(t) = t$.

14.6 Exercises

14.1 Consider the sequence of 'tents' (X^n), where $X_t^n = nt$ for $t \in [0, \frac{1}{2n}]$, $X_t^n = 1 - nt$ for $t \in [\frac{1}{2n}, \frac{1}{n}]$, and zero elsewhere. (There is no real 'randomness' here, the tents don't depend on some ω, but one may think of the X^n defined on whatever probability space, such that $X_t^n(\omega)$ is constant in ω.) Show that all finite dimensional distributions of the X^n converge, but X^n does not converge in distribution.

14.2 Show that ρ as in (1.1) defines a metric.

14.3 Suppose that the ξ_i of Section 14.3 are *iid* normally distributed random variables. Use Doob's supremal inequality (10.5) to obtain $\mu(\max_{j \leq n} |S_j| > \gamma) \leq 3\gamma^{-4} n^2$.

14.4 Show that a finite dimensional projection on $C[0, \infty)$ (with the metric ρ) is continuous.

14.5 Consider $C[0, \infty)$ with the Borel σ-algebra $\mathcal{B}$ induced by ρ and some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. If $X : (\Omega, \mathcal{F}) \to (C[0, \infty), \mathcal{B})$ is measurable, then all maps $\omega \mapsto X_t(\omega)$ are random variables. Show this, as well as its converse. For the latter you need separability that allows you to say that the Borel σ-algebra $\mathcal{B}$ is a product σ-algebra. See also Remark 5.9.

14.6 Prove Proposition 14.4.

14.7 Show that the random variable $Z = \frac{1}{2}(W(s) + W(t)) + \frac{1}{2}\sqrt{t-s}\,\xi$ (see the paragraph following Corollary 14.13) has a normal distribution with mean zero and variance equal to $\frac{1}{2}(s+t)$.

14.8 Prove inequality (14.18).

14.9 The Haar functions form a Complete Orthonormal System in $L^2[0,1]$. Show first that the Haar functions are orthonormal. To prove that the system is complete, you argue as follows. Let f be orthogonal to all $H_{k,n}$ and set $F = \int_0^{\cdot} f(u)du$. Show that F is zero in all $t = k2^{-n}$, and therefore zero on the whole interval. Conclude that $f = 0$.

14.10 Consider the processes W^n of Section 14.5. Let $t_1, \ldots, t_k \in [0,1]$.

(a) Show that the sequence of random vectors $(W^n_{t_1}, \ldots, W^n_{t_k})$ in the $\mathcal{L}^2$-sense converges to $(W_{t_1}, \ldots, W_{t_k})$. (*Hint:* this sequence is Cauchy in L^2).

(b) Show that it follows that $(W_{t_1}, \ldots, W_{t_k})$ has a multivariate normal distribution (see also Exercise 13.13).

14.11 Show that for a random variable X with a $N(0, \sigma^2)$ distribution it holds that $\mathbb{P}(|X| \leq x) < x/\sigma$, for $x, \sigma > 0$.

Index