

Measure Theoretic Probability

P.J.C. Spreij

this version: November 12, 2010

Contents

1	σ-algebras and measures	1
1.1	σ -algebras	1
1.2	Measures	3
1.3	Null sets	4
1.4	π - and d -systems	5
1.5	Probability language	6
1.6	Exercises	8
2	Existence of Lebesgue measure	10
2.1	Outer measure and construction	10
2.2	A general extension theorem	13
2.3	Exercises	15
3	Measurable functions and random variables	17
3.1	General setting	17
3.2	Random variables	19
3.3	Independence	21
3.4	Exercises	23
4	Integration	25
4.1	Integration of simple functions	25
4.2	A general definition of integral	28
4.3	Integrals over subsets	32
4.4	Expectation and integral	33
4.5	\mathcal{L}^p -spaces of random variables	35
4.6	\mathcal{L}^p -spaces of functions	36
4.7	Exercises	40
5	Product measures	42
5.1	Product of two measure spaces	42
5.2	Application in Probability theory	45
5.3	Infinite products	47
5.4	Exercises	49
6	Derivative of a measure	52
6.1	Linear functionals on \mathbb{R}^n	52
6.2	Linear functionals on a Hilbert space	52
6.3	Real and complex measures	53
6.4	Absolute continuity and singularity	54
6.5	The Radon-Nikodym theorem	56
6.6	Additional results	57
6.7	Exercises	58

7	Convergence and Uniform Integrability	61
7.1	Modes of convergence	61
7.2	Uniform integrability	64
7.3	Exercises	66
8	Conditional expectation	68
8.1	Conditional expectation for $X \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$	69
8.2	Conditional probabilities	72
8.3	Exercises	74
9	Martingales and alike	76
9.1	Basic concepts and definition	76
9.2	Stopping times and martingale transforms	79
9.3	Doob's decomposition	82
9.4	Optional sampling	83
9.5	Exercises	85
10	Convergence theorems	87
10.1	Doob's convergence theorem	87
10.2	Uniformly integrable martingales and convergence	89
10.3	\mathcal{L}^p convergence results	91
10.4	The strong law of large numbers	93
10.5	Exercises	95
11	Weak convergence	99
11.1	Generalities	100
11.2	The Central Limit Theorem	107
11.3	Exercises	111
12	Characteristic functions	113
12.1	Definition and first properties	113
12.2	Characteristic functions and weak convergence	117
12.3	The Central Limit Theorem revisited	119
12.4	Exercises	121
13	Brownian motion	125
13.1	The space $C[0, \infty)$	125
13.2	Weak convergence on metric spaces	126
13.3	An invariance principle	127
13.4	The proof of Theorem 13.8	128
13.5	Exercises	130

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 computing 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 *limiting* 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 David Williams, but also other texts have been used. In particular we consulted *Convergence of Stochastic Processes* by David Pollard, *Real and Complex Analysis* by Walter Rudin and *Foundations of Modern Probability* by Olav Kallenberg.

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.

1 σ -algebras and measures

In this chapter we lay down the measure theoretic foundations for 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 $E_n \in \Sigma$ ($n = 1, 2, \dots$).

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 $(-\infty, x] = \bigcap_n (-\infty, x + \frac{1}{n}) \in \mathcal{B}$, 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 $(-\infty, x) = \bigcup_n (-\infty, x - \frac{1}{n}] \in \sigma(\mathcal{I})$. 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 is $\varepsilon(x)$ such that $(x - \varepsilon(x), x + \varepsilon(x)) \subset G$. Hence $G = \cup_{x \in G} (x - \varepsilon(x), x + \varepsilon(x))$. Since \mathbb{Q} is dense in \mathbb{R} , we can replace this union by restricting the x to $G \cap \mathbb{Q}$, which gives a countable union open intervals, of which we just proved that they are elements of $\sigma(\mathcal{I})$. Thus $\mathcal{O} \subset \sigma(\mathcal{I})$, but then also $\mathcal{B} \subset \sigma(\mathcal{I})$. \square

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}_α , 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}_β are defined for all $\beta < \alpha$. Put $\mathcal{E}_\alpha^0 = \cup_{\beta < \alpha} \mathcal{E}_\beta$. We define \mathcal{E}_α 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}_ω 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 $E^c = \cup_{n=1}^\infty E_n^c$ with $E_n = E_n^c$, 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}_ω is a σ -algebra.

The next thing to show is that the cardinality of \mathcal{E}_ω 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}_β is less than or equal to 2^{\aleph_0} for all $1 \leq \beta < \alpha$, a property which then also holds for \mathcal{E}_α^0 . The construction of \mathcal{E}_α from \mathcal{E}_α^0 yields, by the same argument as used above for \mathcal{E}_1 , that also \mathcal{E}_α 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}_ω 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. We can take \mathcal{E} as the set of intervals (a, b) with $a, b \in \mathbb{Q}$, augmented with the empty set. We 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 $\mu_0 : \Sigma_0 \rightarrow [0, \infty]$ and $\mu : \Sigma \rightarrow [0, \infty]$. Note that ∞ is allowed as a possible value.

We call μ_0 *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 $\mu : \Sigma \rightarrow [0, \infty]$ 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 $\mathbb{N} = \cup_n \{1, \dots, n\}$.

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

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

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

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. \square

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 $n \rightarrow \infty$.

(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 $n \rightarrow \infty$.

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. \square

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. \square

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 $F_n = \{n, n+1, \dots\}$, 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 $N \in S$ 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}_μ 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, \dots 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 $d(\mathcal{I})$. But, by construction, \mathcal{D}_2 is a π -system. So we conclude that $d(\mathcal{I})$ is a π -system, as desired. \square

Sometimes another version of Lemma 1.13 is useful.

Corollary 1.14 *The assertion of Proposition 1.12 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 Proposition 1.12. 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 $\sigma(\mathcal{I})$. \square

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 $F \setminus E \in \mathcal{D}$. \square

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 $G_n = \{n, n+1, \dots\}$ ($n \in \mathbb{N}$).

1.5 Probability language

In Probability Theory, one usually writes $(\Omega, \mathcal{F}, \mathbb{P})$ instead of (S, Σ, μ) . On one hand this is merely change of notation. 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, $\mathbb{P}(\Omega) = 1$. 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 $(\omega_1, \omega_2, \dots)$ 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, \dots and define

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

Note that the sets $F_m = \bigcap_{n \geq m} E_n$ form an increasing sequence and the sets $D_m = \bigcup_{n \geq m} E_n$ 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, which are defined as follows. 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 $p : \Omega \rightarrow [0, 1]$ 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 $\mu : \Sigma_0 \rightarrow [0, \infty]$ is countably additive if $\mu(H_n) \rightarrow 0$ 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 $\mu(H_n) \rightarrow 0$ 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 $\mu^* : 2^S \rightarrow [0, \infty]$ 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 $\mu : \Sigma_\mu \rightarrow [0, \infty]$ 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

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

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 $F \subset S$

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

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

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

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

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

Combined with $\mu^*(F) \leq \mu^*(F \cap E) + \mu^*(F \cap E^c)$, which again is the result of subadditivity, we see that $E \in \Sigma_\mu$. It follows that Σ_μ is a σ -algebra. Take then $F = S$ in (2.2) to see that μ^* restricted to Σ_μ is a measure. \square

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 $E \subset \mathbb{R}$

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

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, \dots 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

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

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 $\varepsilon \rightarrow 0$.

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

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

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 , $\{I_1, \dots, I_n\}$ 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 $\lambda_0(I) = (b - a_k) + (a_k - a) \leq (b_k - a_k) + (a_k - a) \leq \lambda_0(I_k) + \lambda_0(I \setminus I_k) \leq \sum_j \lambda_0(I_j)$. \square

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 $\lambda^*(E) \leq \sum_{I \in \mathcal{I}(E)} \lambda^*(I) - \varepsilon$, 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 $\varepsilon \downarrow 0$. \square

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

Theorem 2.6 *The (restricted) function $\lambda : \mathcal{B} \rightarrow [0, \infty]$ 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

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

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

The sets in Σ_λ are also called *Lebesgue-measurable* sets. A function $f : \mathbb{R} \rightarrow \mathbb{R}$ 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 $C_0 = [0, 1]$ by 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. 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.

Theorem 2.7 *Let Σ_0 be an algebra on a set S and let $\mu_0 : \Sigma_0 \rightarrow [0, \infty]$ 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 $\{E_1, E_2, \dots\}$ 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 $\varepsilon \rightarrow 0$, 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, \dots 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 choose 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 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 $\mu^*(A) = \inf\{\sum_{i=1}^{\infty} \mu(E_i) : \{E_1, E_2, \dots\} \in \mathcal{M}(A)\}$. 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 associated outer measure on S . 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 $\lambda^*((0, 1) \setminus E) = 1$.

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 $h : S \rightarrow \mathbb{R}$ 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 Σ/\mathcal{B} -measurable functions, depending on the situation. If S is 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 $E \subset S$. The indicator function of E is defined by $\mathbf{1}_E(s) = 1$ if $s \in E$ and $\mathbf{1}_E(s) = 0$ if $s \notin E$. Note 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 $h : S \rightarrow [-\infty, \infty]$ 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 $h : S \rightarrow \mathbb{R}$.

(i) If \mathcal{C} is 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 $f : \mathbb{R} \rightarrow \mathbb{R}$ 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 $(f \circ h)^{-1}[B] = h^{-1}[f^{-1}[B]] \in \Sigma$. \square

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 $h : S \rightarrow [-\infty, \infty]$ 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 ($i = 1, 2, 3$), $h : S_1 \rightarrow S_2$ is Σ_1/Σ_2 -measurable and $f : S_2 \rightarrow S_3$ 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 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 $E \in \Sigma$.

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

Proposition 3.5 (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 $h : S \rightarrow [-\infty, \infty]$ is measurable. The assertion follows from $L = \{\limsup h_n < \infty\} \cap \{\liminf h_n > -\infty\} \cap \{h = 0\}$. \square

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, Σ -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

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

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 $\{i2^{-n} \leq f(s) < (i+1)2^{-n}\} \in \Sigma$, since f is measurable. But then this set is also an element of \mathcal{D} , since $\Sigma \subset \mathcal{D}$ (see above) and hence $\mathbf{1}_{\{i2^{-n} \leq f(s) < (i+1)2^{-n}\}} \in \mathcal{H}$. 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 $f \in \mathcal{H}$. \square

3.2 Random variables

We return to the setting of Section 1.5 and so we consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. In this setting Definition 3.1 takes the following form.

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

By definition, random variables are nothing else but measurable functions. Following the tradition, we denote them by X (or other capital letters), rather than by h , as in the previous sections. We often need the σ -algebra $\sigma(X)$, the smallest σ -algebra on Ω that makes X a random variable. In other words, it is the intersection of all σ -algebras \mathcal{F} , such that X is a random variable in the sense of Definition 3.7. Of course, X is \mathcal{F} -measurable iff $\sigma(X) \subset \mathcal{F}$.

If we have a collection of mappings $X := \{X_i : \Omega \rightarrow \mathbb{R}\}_{i \in I}$, 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 $\mu : \mathcal{B} \rightarrow [0, 1]$ by

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

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 $F : \mathbb{R} \rightarrow [0, 1]$, 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 $\lim_{x \rightarrow \infty} F(x) = 1$ and $\lim_{x \rightarrow -\infty} F(x) = 0$. 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 $U : \Omega \rightarrow [0, 1]$ 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 $X : \Omega \rightarrow \mathbb{R}$ 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^- is a 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 $\mathbb{P}(X^- \leq c) = \mathbb{P}([0, F(c)]) = \lambda([0, F(c)]) = F(c)$. \square

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

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

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

The above definition also applies to finite sequences. For instance, a finite sequence of σ -algebras $\mathcal{F}_1, \dots, \mathcal{F}_n$ is called independent if the infinite sequence $\mathcal{F}_1, \mathcal{F}_2, \dots$ 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 $\nu_I(H) = \mathbb{P}(H)\mathbb{P}(I)$. 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 $H \in \mathcal{H}$. \square

Lemma 3.13 (Borel-Cantelli) *Let E_1, E_2, \dots 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 $N \rightarrow \infty$, 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 $\mathbb{P}(\liminf E_n^c) \leq \sum_{n=1}^\infty \mathbb{P}(D_n^\infty) = 0$. \square

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 the marginal distributions of each member is prescribed. This is the content of Theorem 3.15 below. It turns out 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 $b_1(x), b_2(x), \dots$ 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 $x \mapsto b_k(x)$, $k = 1, 2, \dots$ are Borel-measurable and $x = \sum_{k=1}^\infty 2^{-k} b_k(x)$.

Lemma 3.14 (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 $f_k : [0, 1) \rightarrow [0, 1)$ 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 $x_1, \dots, x_n \in \{0, 1\}$, one easily computes the joint probability $\mathbb{P}(X_1 = x_1, \dots, X_n = x_n) = 2^{-n}$. 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} , $j, k = 1, 2, \dots$ 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!). \square

Here is the result we are after.

Theorem 3.15 *Let μ_1, μ_2, \dots 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 3.14 and define $Y_k = X_k \circ Z_k$, $k \geq 1$. These are easily seen to have the desired properties \square

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 $\Pi(X) := \{X^{-1}(-\infty, x] : x \in \mathbb{R}\}$ 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 $\mathcal{X}_n = \sigma\{X_1, \dots, X_n\}$ ($n \in \mathbb{N}$) 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 $X : \Omega \rightarrow \mathbb{R}$ 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 having a uniform distribution on $[0, 1]$ and put $X = F^{-1}(U)$. Show that X has distribution F .

3.7 Let F be a distribution function and put $X^+(\omega) = \inf\{z \in \mathbb{R} : F(x) > \omega\}$. 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 $\mathcal{G}_1, \mathcal{G}_2, \dots$ be σ -algebras and let $\mathcal{G} = \sigma(\mathcal{G}_1 \cup \mathcal{G}_2 \cup \dots)$.

- (a) Show that $\Pi = \{G_{i_1} \cap G_{i_2} \cap \dots \cap G_{i_k} : k \in \mathbb{N}, i_k \in \mathbb{N}, G_{i_j} \in \mathcal{G}_{i_j}\}$ is a π -system that generates \mathcal{G} .
- (b) Assume that $\mathcal{G}_1, \mathcal{G}_2, \dots$ 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 *iid* sequence X_1, X_2, \dots . Let $\mathcal{F}_n = \sigma(X_1, \dots, X_n)$ and $\mathcal{T}_n = \sigma(X_{n+1}, X_{n+2}, \dots)$, $n \geq 1$. Let \mathcal{I} be the collection of events of the type $\{X_1 \in B_1, \dots, X_n \in B_n\}$, 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 $(\omega_1, \omega_2, \dots)$ 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 $E_p = \{\omega \in \Omega : \lim_{n \rightarrow \infty} \frac{1}{n} S_n(\omega) = p\}$. 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 $\mathbb{E}X = \int x f(x) dx$. 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 $f : S \rightarrow [0, \infty]$ is called a nonnegative simple function, if it has a representation as a finite sum

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

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

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

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

when f has representation (4.1).

Other notations that we often use for this integral are $\int f(s) \mu(ds)$ 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 $\phi : S \rightarrow \{0, 1\}^n$ by $\phi(s) = (\mathbf{1}_{A_1}(s), \dots, \mathbf{1}_{A_n}(s))$. Let $\{0, 1\}^n = \{u_1, \dots, u_m\}$ where $m = 2^n$ and put $U_k = \phi^{-1}(u_k)$. Then the collection $\{U_1, \dots, U_m\}$ 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

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

Now apply the definition of $\mu(f)$ by using the representation of f given by (4.3). This gives $\mu(f) = \sum_k \left(\sum_{i \in T_k} a_i \right) \mu(U_k)$. Interchanging the summation order, we see that this is equal to $\sum_i a_i \left(\sum_{k \in S_i} \mu(U_k) \right) = \sum_i a_i \mu(A_i)$, 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 representation 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. \square

Example 4.4 Here is the most elementary 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) . If $f \in \mathcal{S}^+$, only finitely many of the f_i are nonzero. Let us assume that $f_i = 0$ for $i > n$. Then f can be represented in a somewhat cumbersome (but not necessarily unique) way by

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

Since $\tau(\{i\}) = 1$, we get $\tau(f) = \sum_{i=1}^n f_i$, not else but the sum of the f_i . In this case, integration is just summation. Of course, a different representation would yield the same answer.

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 4.5 *Let $f, g \in \mathcal{S}^+$ 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 $g = \sum_j b_j \mathbf{1}_{B_j}$ say. 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 $A_i \cap B_j$. \square

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

Notice that for functions $f \in \mathcal{S}^+$, Definition 4.6 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 . The set of nonnegative measurable functions will be denoted by Σ^+ . We immediately have some extensions of results in the previous section.

Proposition 4.7 Let $f, g \in \Sigma^+$. 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 $N = \{f > g\}$. Take $h \in \mathcal{S}^+$ with $h \leq f$. Then also $h\mathbf{1}_N, h\mathbf{1}_{N^c} \in \mathcal{S}^+$ and by Proposition 4.5(iii), 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. \square

Example 4.8 We extend the situation of Example 4.4, by allowing infinitely many f_i to be positive. The result will be $\tau(f) = \sum_{i=1}^{\infty} f_i$, classically defined as $\lim_{n \rightarrow \infty} \sum_{i=1}^n f_i$. Check that this is in agreement with Definition 4.6. See also Exercise 4.1.

Example 4.9 Let $(S, \Sigma, \mu) = ([0, 1], \mathcal{B}([0, 1]), \lambda)$ and f the indicator of the rational numbers in $[0, 1]$. Then $f = 0$ a.e. and, consequently, $\lambda(f) = 0$. This is a nice example of a function that is not Riemann integrable.

The following will frequently be used.

Lemma 4.10 *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 $h_n \in \mathcal{S}^+$ 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. \square

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

Theorem 4.11 *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 4.7, 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 $h \in \mathcal{S}^+$ 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

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

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 $n \rightarrow \infty$ 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. \square

Example 4.12 Here is a nice application of Theorem 4.11. Let $f \in \Sigma^+$ and, for each $n \in \mathbb{N}$, put $E_{n,i} = \{i2^{-n} \leq f < (i+1)2^{-n}\}$ ($i \in I_n := \{0, \dots, n2^n - 1\}$),

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 \mathcal{S}^+ , with limit f . Theorem 4.11 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 4.13 *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 4.14 *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 4.11, $\mu(g_n) \uparrow \mu(\liminf f_n)$ 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 4.15 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 4.14.

We now extend the notion of integral to (almost) arbitrary measurable functions. Let $f \in \Sigma$. For (extended) real numbers x on defines $x^+ = \max\{x, 0\}$ and $x^- = \max\{-x, 0\}$. Then, for $f : S \rightarrow [-\infty, \infty]$, 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 4.16 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 4.17 *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 $\mathcal{L}^1(S, \Sigma, \mu)$.*

Proof Exercise 4.3. □

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

Theorem 4.18 *Let $(f_n) \subset \Sigma$ and $f = \limsup f_n$. Assume that $f_n(s) \rightarrow f(s)$ for all s outside a set of measure zero. Assume there exists a function $g \in \Sigma^+$ such that $|f_n| \leq g$ a.e. and that $\mu(g) < \infty$. Then $\mu(|f_n - f|) \rightarrow 0$, and hence $\mu(f_n) \rightarrow \mu(f)$.*

Proof The second assertion easily follows from the first one, which we prove now for the case that $f_n \rightarrow f$ everywhere. One has the inequality $|f| \leq g$, whence $|f_n - f| < 2g$. 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. \square

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

Lemma 4.19 *Let $(f_n) \subset \Sigma^+$ and assume that $f_n \rightarrow f$ a.e. Assume that $\mu(f_n)$ is finite for all n and $\mu(f) < \infty$ as well. Then $\mu(|f_n - f|) \rightarrow 0$ iff $\mu(f_n) \rightarrow \mu(f)$.*

Proof The 'only if' part follows from Theorem 4.18. Assume then that $\mu(f_n) \rightarrow \mu(f)$. 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 $\mu((f_n - f)^-) \rightarrow 0$, by virtue of Theorem 4.18. \square

Example 4.20 Let $(S, \Sigma, \mu) = ([0, 1], \mathcal{B}(\mathbb{R}), \lambda)$, where λ is Lebesgue measure. Assume that $f \in C[0, 1]$. Exercise 4.6 tells us that $f \in \mathcal{L}^1([0, 1], \mathcal{B}(\mathbb{R}), \lambda)$ and that $\lambda(f)$ is equal to the Riemann integral $\int_0^1 f(x) dx$. 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 6.7. Let $f \in \Sigma^+$ and $E \in \Sigma$. Then we may define

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

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). 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 4.21 *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 $E \in \Sigma$

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

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 4.22 *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 4.22 states that $\int h \, d\nu = \int hf \, d\mu$, valid for all $h \in \mathcal{L}^1(S, \Sigma, \nu)$. By ‘erasing’ on both sides of this equality, one gets $d\nu = f \, d\mu$, and then, by ‘dividing’, one gets

$$\frac{d\nu}{d\mu} = f.$$

Although both of these operations are nonsense from a mathematical point of view, the notation in the last displayed identity, which looks like a derivative, turns out to be very convenient. We will return to this in Chapter 6, where we discuss Radon-Nikodym derivatives.

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 $X : \Omega \rightarrow \mathbb{R}$ 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}

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

provided that the integral is well defined, which is certainly the case if $\mathbb{P}(|X| < \infty) = 1$. 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 \mathbb{E} xpectation of X . Note 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 4.23 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 4.4 and 4.8, we get $\mathbb{E}X = \sum_{n=1}^{\infty} x_n p_n$. Let us switch to a different approach. Let ξ_1, ξ_2, \dots be the different elements of the set $\{x_1, x_2, \dots\}$ 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 $h : \mathbb{R} \rightarrow \mathbb{R}$ 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 by direct application of the definition of expectation to Y . But we also have

Proposition 4.24 *Let X be a random variable, and $h : \mathbb{R} \rightarrow \mathbb{R}$ 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*

$$\mathbb{E}h \circ X = \int_{\mathbb{R}} h \, d\mathbb{P}^X. \quad (4.8)$$

Proof Exercise 4.11. □

It follows from this proposition that one can also compute $\mathbb{E}Y$ as $\mathbb{E}Y = \int_{\mathbb{R}} y \mathbb{P}^Y(dy)$.

This definition of expectation yields the familiar formulas containing a sum for discrete random variables and integrals for random variables 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 4.18) with the function g equal to a constant. Here is a simple example of the application of the Monotone Convergence Theorem.

Example 4.25 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 4.26 *Let X be a real valued random variable and $g : \mathbb{R} \rightarrow [0, \infty]$ 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 4.26 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 $\text{Var } X \geq c^2\mathbb{P}(|X - \mathbb{E} X| \geq c)$.

We now turn to a result that is known as *Jensen's inequality*, Proposition 4.27 below. Recall that a function $g : G \rightarrow \mathbb{R}$ 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 $\text{Int } G$, the interior of G . Rewrite the above inequality as

$$\frac{g(y) - g(x)}{y - x} \leq \frac{g(z) - g(x)}{z - x}. \tag{4.9}$$

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

that both onesided derivatives are increasing. If one takes in (4.9) 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

$$g(z) - g(x) \geq d(x)(z - x). \quad (4.10)$$

The following proposition is now easy to prove.

Proposition 4.27 *Let $g : G \rightarrow \mathbb{R}$ 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 Since $\mathbb{P}(X \in G) = 1$, we have $\mathbb{E}X \in G$ and (4.10) with $x = \mathbb{E}X$ and z replaced with X holds a.s. So,

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

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

4.5 \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 4.28 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-Schwartz inequality.

Proposition 4.29 *Let $X, Y \in \mathcal{L}^2$. Then $\mathbb{E}|XY| \leq \|X\|_2 \|Y\|_2$.*

Proof If $\mathbb{E}X^2 = \mathbb{E}Y^2 = 0$, then $X = 0$ and $Y = 0$ a.s. (Lemma 4.10) 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 4.29 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 extend to a product rule for expectations.

Proposition 4.30 *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 $X \in \mathcal{S}^+$. 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 . \square

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

Proposition 4.31 *Let $1 \leq p \leq r$ and $X \in \mathcal{L}^r$, then $X \in \mathcal{L}^p$ 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 $X \in \mathcal{L}^p$. Observe that $x \rightarrow |x|^a$ is convex. We apply Jensen's inequality to get $(\mathbb{E}|X|^p)^a \leq \mathbb{E}(|X|^{pa})$, from which the result follows. \square

4.6 \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 4.28.

Definition 4.32 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 4.33 *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 4.10 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. \square

Notice that this result for $p = 2$ yields the Cauchy-Schwartz inequality $\|fg\|_1 \leq \|f\|_2 \|g\|_2$ for square integrable functions.

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

Theorem 4.34 *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

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

But $(p-1)q = p$ and after dividing by $(\mu(|f + g|^{(p-1)q})^{1/q}$, we obtain $\mu(|f + g|^p)^{1-1/q} \leq \|f\|_p + \|g\|_p$. The result follows, because $1 - 1/q = 1/p$. \square

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 4.35 *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 : $\|X_n - X_m\|_p \rightarrow 0$ for $n, m \rightarrow \infty$. Then there exists a limit $X \in \mathcal{L}^p$ such that $\|X_n - X\|_p \rightarrow 0$. 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 4.25) 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 $X_{k_n} \rightarrow X$ 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 $m \rightarrow \infty$ 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, $\limsup \mathbb{E} |X_l - X|^p \leq 2^{-np}$, for all n . Hence $\|X_l - X\|_p \rightarrow 0$. The proof for $p = \infty$ is left as Exercise 4.13. \square

Remark 4.36 Notice that it follows from Theorem 4.35 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 $L^p(S, \Sigma, \mu)$ ($p \in [1, \infty]$), 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 $\langle f, g \rangle := \int fg d\mu$. 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 an inner product space and norm $\|\cdot\|$. Recall that an orthogonal projection on a subspace L is a linear mapping $\pi : S \rightarrow L$ 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 4.37 *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 (ii) 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

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

to get

$$2\|x - \frac{1}{2}(x_m + x_n)\|^2 = \|x - x_m\|^2 + \|x - x_n\|^2 - \frac{1}{2}\|x_m - x_n\|^2. \quad (4.11)$$

Since the left hand side of (4.11) 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 $n \rightarrow \infty$ 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 (4.11) 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 $f(t) = \|x - \hat{x}\|^2 + 2t\langle x - \hat{x}, y \rangle + t^2\|y\|^2$. 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 $\|x - y\|^2 = \|x - \hat{x}\|^2 + 2\langle x - \hat{x}, \hat{x} - y \rangle + \|\hat{x} - y\|^2$, since the cross term vanishes by assumption. The implication (ii) \Rightarrow (i) is obvious. \square

Corollary 4.38 *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 4.35 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 4.37, 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

4.7 Exercises

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

4.2 Prove Proposition 4.13.

4.3 Prove Proposition 4.17 (assume Proposition 4.13). 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 4.18 by using the ‘everywhere’ version.

4.5 Here is another version of Scheffé’s lemma (Lemma 4.19). Let $(f_n) \subset \mathcal{L}^1(S, \Sigma, \mu)$ and assume that $f_n \rightarrow f$ a.e., where $f \in \mathcal{L}^1(S, \Sigma, \mu)$ too. Then $\mu(|f_n - f|) \rightarrow 0$ iff $\mu(|f_n|) \rightarrow \mu(|f|)$. Show this. (*Hint*: apply Lemma 4.19 to f_n^+ and f_n^- .)

4.6 In this exercise λ denotes Lebesgue measure on the Borel sets of $[0, 1]$. Let $f : [0, 1] \rightarrow \mathbb{R}$ be continuous. Then the Riemann integral $I := \int_0^1 f(x) dx$ 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 $f : [0, \infty) \rightarrow \mathbb{R}$ be given by $f(x) = \frac{\sin x}{x}$ for $x > 0$ and $f(0) = 1$. Show that $I := \int_0^{\infty} f(x) dx$ exists as an improper Riemann integral (i.e. the limit $\lim_{T \rightarrow \infty} \int_0^T f(x) dx$ exists and is finite), but that $f \notin \mathcal{L}^1([0, \infty), \mathcal{B}([0, \infty)), \lambda)$. In Exercise 5.10 you compute that $I = \frac{\pi}{2}$.

4.8 Prove Proposition 4.21 by means of the standard machinery.

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

4.10 Prove Proposition 4.22.

4.11 Prove Proposition 4.24. *Hint*: Use the standard machinery for h .

4.12 Give the details for Example 4.25.

4.13 Give the proof of Theorem 4.35 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 $f : [0, 1] \rightarrow \mathbb{R}$ be Riemann integrable and let $I = \int_0^1 f(x) dx$. 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 \mathcal{S}^+ and an increasing sequence (ℓ_n) in \mathcal{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 4.18.* Let $(f_n) \subset \Sigma$ and $f = \limsup f_n$ and assume that $f_n(s) \rightarrow f(s)$ 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 $g_n(s) \rightarrow g(s)$ for all s outside a set of measure zero and that $\mu(g_n) \rightarrow \mu(g) < \infty$. Show that $\mu(|f_n - f|) \rightarrow 0$.

5 Product measures

Sofar we have considered measure spaces (S, Σ, μ) and we have looked at integrals of the type $\mu(f) = \int f \, d\mu$. 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 $f : S_1 \times S_2 \rightarrow \mathbb{R}$. 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 $f : S_1 \times S_2 \rightarrow \mathbb{R}$ and assume any good notion of measurability and integrability. Then $\mu(f(\cdot, s_2)) := \int f(\cdot, s_2) \, d\mu_1$ 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 $\sum_n \sum_m a_{n,m} = -1$. Similarly one shows that $\sum_m \sum_n a_{n,m} = +1$. In order to justify interchanging the summation order, additional conditions have to be imposed.

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 $\Sigma := \sigma(\mathcal{R})$.

Alternatively, one can consider the projections $\pi_i : S \rightarrow S_i$, 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 $e_{s_1} : S_2 \rightarrow S$ 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 $f : S \rightarrow \mathbb{R}$ 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. \square

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

$$\begin{aligned} I_1^f(s_1) &= \int f(s_1, s_2) \mu_2(ds_2) \\ I_2^f(s_2) &= \int f(s_1, s_2) \mu_1(ds_1). \end{aligned}$$

Lemma 5.2 *Let f be a bounded Σ -measurable function. Then the mappings $I_i^f : S_i \rightarrow \mathbb{R}$ are Σ_i -measurable ($i = 1, 2$). Moreover we have the identity*

$$\mu_1(I_1^f) = \mu_2(I_2^f), \quad (5.1)$$

or, in a more appealing notation,

$$\int_{S_1} \left(\int_{S_2} f(s_1, s_2) \mu_2(ds_2) \right) \mu_1(ds_1) = \int_{S_2} \left(\int_{S_1} f(s_1, s_2) \mu_1(ds_1) \right) \mu_2(ds_2). \quad (5.2)$$

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 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 4.11 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. \square

Remark 5.3 The converse statement of Proposition 5.1 is in general not true. There are functions $f : S \rightarrow \mathbb{R}$ 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 function is given. See Exercise 5.9.

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 5.2 doesn't hold and therefore conclude that f cannot be measurable w.r.t. $\mathcal{B}([0, 1]) \times \mathcal{B}([0, 1])$.

It follows from Lemma 5.2 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 $\mu : \Sigma \rightarrow [0, \infty)$ by $\mu(E) = \mu_2(I_2^{\mathbf{1}_E})$ for $E \in \Sigma$.

In Theorem 5.5 (known as Fubini's theorem) below we assert that this defines a measure 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 (i) *The mapping μ of Definition 5.4 is a measure on (S, Σ) . Moreover, it is the only measure on (S, Σ) with the property that $\mu(E_1 \times E_2) = \mu_1(E_1)\mu_2(E_2)$. 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*

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

(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. \square

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 $S_1 = \cup_{i=1}^n S_1^i$ 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 not 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 measure fails in this example.

We close the section 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. The result is product spaces of the form $(S_1 \times \dots \times S_n, \Sigma_1 \times \dots \times \Sigma_n, \mu_1 \times \dots \times \mu_n)$ 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. 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.

5.2 Application in Probability theory

In this section we consider real valued random variables, as well as *random vectors*. The latter require a definition. Consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a map $X : \Omega \rightarrow E$, 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 recognize in the term random vector.

An obvious example of a vector space is \mathbb{R}^2 . Suppose we have two random variables $X_1, X_2 : \Omega \rightarrow \mathbb{R}$. We can consider the map $X = (X_1, X_2) : \Omega \rightarrow \mathbb{R}^2$, 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 5.6 *It holds that $\mathcal{B}(\mathbb{R}^2) = \mathcal{B}(\mathbb{R}) \times \mathcal{B}(\mathbb{R})$.*

Proof The projections $\pi_i : \mathbb{R}^2 \rightarrow \mathbb{R}$ 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. \square

Remark 5.7 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 5.8 *Let $X_1, X_2 : \Omega \rightarrow \mathbb{R}$ be given. The vector mapping $X = (X_1, X_2) : \Omega \rightarrow \mathbb{R}^2$ is a random vector iff the X_i are random variables.*

Proof Exercise 5.3. \square

Remark 5.9 Let X_1, X_2 be two random variables. We already knew that $X_1 + X_2$ is a random variable too. It also follows from the present results. Let $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ be a continuous function. Then it is also $\mathcal{B}(\mathbb{R}^2)$ -measurable, and by Corollary 5.8 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 $F = F_X : \mathbb{R}^2 \rightarrow [0, 1]$, 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, $F_{X_1}(x_1) = \lim_{x_2 \rightarrow \infty} F(x_1, x_2)$, also written as $F(x_1, \infty)$.

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 5.10 *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. □

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 5.13 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 $\omega = (\omega_1, \omega_2, \dots)$ a typical element of Ω . The construction of the product σ -algebra remains the same as before. The projection $\pi_n : \Omega \rightarrow \Omega_n$ 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 $\pi_{(1, \dots, n)} : \Omega \rightarrow \prod_{k=1}^n \Omega_k$ by $\pi_{(1, \dots, n)}(\omega) = (\omega_1, \dots, \omega_n)$. 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 $\mathcal{F}_1 \times \dots \times \mathcal{F}_n$ under the projection $\pi_{(1, \dots, n)}$. It follows that a cylinder is

of the type $B_n \times \prod_{k=n+1}^{\infty} \Omega_k$ (for some n), with $B_n \in \mathcal{F}_1 \times \dots \times \mathcal{F}_n$. 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 $\mathbb{P}_0 : \mathcal{C} \rightarrow [0, 1]$ 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 5.11 *There exists a unique probability measure \mathbb{P} on (Ω, \mathcal{F}) such that \mathbb{P} restricted to \mathcal{C} is equal to \mathbb{P}_0 . In particular,*

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

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

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 $\lim_{n \rightarrow \infty} \mathbb{P}_0(C^n) > 0$, one must have $C := \bigcap_{n=1}^{\infty} C^n \neq \emptyset$. 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 $B_n \times \prod_{k=N+1}^{\infty} \Omega_k$, with $B_n \in \mathcal{F}_1 \times \dots \times \mathcal{F}_N$, we are done, since in this case $\mathbb{P}_0(C_n) = \mathbb{P}_1 \times \dots \times \mathbb{P}_N(B_n)$. We already know that $\mathbb{P}_1 \times \dots \times \mathbb{P}_N$ 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 $\omega = (\omega_1, \dots, \omega_n, \omega'_n)$. 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 $C(\omega_1, \dots, \omega_n) = \{\omega'_n : (\omega_1, \dots, \omega_n, \omega'_n) \in C\}$. Then the probabilities $\mathbb{P}'_n(C(\omega_1, \dots, \omega_n))$ are well defined.

Since we have assumed for the decreasing sequence (C_n) that $\lim_{n \rightarrow \infty} \mathbb{P}_0(C_n) > 0$, 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 5.2 that $E_n^1 \in \mathcal{F}_1$. Then, using 'Fubini computations',

we obtain

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

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 $E^1 = \bigcap_n E_n^1$, 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 $\omega^* = (\omega_1^*, \omega_2^*, \dots)$ that belongs to all C_n and therefore the intersection $\bigcap_n C_n$ is not empty. \square

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. Prove also 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 $f \in m\Sigma^+$). 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 5.8.

5.4 Prove Proposition 5.10.

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

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

Define

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

Show that the distribution \mathbb{P}^X of X is absolutely continuous w.r.t. the Lebesgue measure on $\mathcal{B}(\mathbb{R})$ and that for all $B \in \mathcal{B}(\mathbb{R})$ one has

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

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 $F_Z(z) = \int_{\mathbb{R}} F_X(z - y) \mu(dy)$.

(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

$$f_Z(z) := \int_{\mathbb{R}} f_X(z - y) \mu(dy).$$

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

$$\mathbb{E} \sum_{k=1}^{\infty} Z_k = \sum_{k=1}^{\infty} \mathbb{E} Z_k. \quad (5.4)$$

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 $\mathbb{P}(\sum_{k=1}^{\infty} Z_k = \infty)$. Formulate a result similar to (5.4) for random variables Z_k that may assume negative values as well.

5.8 Let the vector of random variables (X, Y) have a joint probability density function f . Let f_X and f_Y be the (marginal) probability density functions of X and Y respectively. Show that X and Y are independent iff $f(x, y) = f_X(x)f_Y(y)$ for all x, y except in a set of $\lambda \times \lambda$ -measure zero.

5.9 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 and such that for all $b \in \mathbb{R}$ the function $x \mapsto f(x, b)$ is continuous. 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 . Let $a_i^n = i/n$, $i \in \mathbb{Z}, n \in \mathbb{N}$. Define

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

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

5.10 Show that for $t > 0$

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

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

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

5.11 Let $F, G : \mathbb{R} \rightarrow \mathbb{R}$ 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 $\int f dF$ instead of $\int f d\mu_F$. Use Fubini's theorem to show the integration by parts formula, valid for all $a < b$,

$$F(b)G(b) - F(a)G(a) = \int_{(a,b]} F(s-) dG(s) + \int_{(a,b]} G(s) dF(s).$$

Hint: integrate $\mathbf{1}_{(a,b)^2}$ and split the square into a lower and an upper triangle.

5.12 Let F be the distribution function of a nonnegative random variable X and assume that $\mathbb{E} X^\alpha < \infty$ for some $\alpha > 0$. Use Exercise 5.11 to show that

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

5.13 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 .

(i) 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.

(ii) 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 $E \in \mathcal{F}_i$.

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 4.22, known as the Radon-Nikodym theorem, Theorem 6.7. In the proof it, we need that a continuous map on a Hilbert space can be represented by the inner product with a fixed element in that space. Therefore we start with proving this.

6.1 Linear functionals on \mathbb{R}^n

Let $H = \mathbb{R}^n$. It is well known that every linear map $T : H \rightarrow \mathbb{R}^m$ can uniquely be represented by an $m \times n$ matrix $M = M(T)$ via $Tx = Mx$, which we will prove below for the case $m = 1$. Take the result for granted, 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. Let $y = M^\top \in \mathbb{R}^n$, then we have

$$Tx = \langle x, y \rangle. \tag{6.1}$$

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 . 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

$$\lambda = \frac{\langle x, z \rangle}{\langle z, z \rangle} \text{ and } Tx = \lambda Tz. \tag{6.2}$$

Let $y = \frac{Tz}{\langle z, z \rangle} z$. Then $\langle x, y \rangle = \frac{Tz}{\langle z, z \rangle} \langle x, z \rangle = Tx$, as follows from (6.2). Uniqueness of y is shown as follows. Let $y' \in E$ be such that $Tx = \langle x, y' \rangle$. Then $\langle x, y - y' \rangle$ is zero for all $x \in E$, 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.

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 be a continuous linear functional on 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 *There exists a unique element $y \in H$ such that $Tx = \langle x, y \rangle$.*

Proof We exclude the trivial case in which $T = 0$. Let K be the kernel of T . Since T is linear, K is a closed subspace of H . Take an element w with $Tw \neq 0$. Since K is closed, the orthogonal projection u of w on K exists and we have $w = u + z$, where z belongs to the orthogonal complement of K . Obviously $z \neq 0$. The rest of the proof is exactly the same as for the finite dimensional case. \square

This theorem can be summarized as follows. The dual space H^* of H (the linear space of all continuous linear functionals on 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 are associate to H 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^* . 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 $\|T_n - T\|^* = \|y_n - y\| \rightarrow 0$. 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-Schwartz inequality that this norm $\|\cdot\|^*$ is the same as the one in the previous paragraph.

6.3 Real and complex measures

Consider a measurable space (S, Σ) . A function $\mu : \Sigma \rightarrow \mathbb{C}$ 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 now be called a *positive measure*. In these notes a measure is 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, \dots 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

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

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

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

It can be shown (and this is quite some work) that $|\mu|$ is a (positive) measure on (S, Σ) with $|\mu|(S) < \infty$ and 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

$$\mu = \mu^+ - \mu^-.$$

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

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 6.2 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 6.5 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 6.2.

Proposition 6.3 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.

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

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

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

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

Proof Exercise 6.2. □

Proposition 6.4 Let μ be a positive measure and ν_a and ν_s be arbitrary measures on (S, Σ) . Assume that $\nu_a \ll \mu$ and $\nu_s \perp \mu$. Put

$$\nu = \nu_a + \nu_s. \tag{6.4}$$

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 6.3), and hence both are zero (Proposition 6.3 again). □

The content of Proposition 6.4 is that the decomposition (6.4) 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 . Recall

Proposition 6.5 Let μ be a positive measure on (S, Σ) and h a nonnegative measurable function on S . Then the map $\nu : \Sigma \rightarrow [0, \infty]$ defined by

$$\nu(E) = \mu(\mathbf{1}_E h) \tag{6.5}$$

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 measure ν that is absolutely continuous w.r.t. μ is of the form (6.5). We will use in that case the notation

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

In the next section we use

Lemma 6.6 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

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

where E runs through the collection of sets 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. Suppose that $\mu(E) > 0$. Then we would have

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

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

6.5 The Radon-Nikodym theorem

As an appetizer for the Radon-Nikodym theorem below 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

$$\nu(E) = \mu(\mathbf{1}_E h), \quad \forall E \subset S. \tag{6.6}$$

Observe that we have obtained an expression like (4.7). The principal theorem on absolute continuity (and singularity) is the following.

Theorem 6.7 *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 $h \in \mathcal{L}^1(S, \Sigma, \mu)$ (the Radon-Nikodym derivative of ν_a w.r.t. μ) such that $\nu_a(E) = \mu(\mathbf{1}_E h)$ for all $E \in \Sigma$. 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 often written as*

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

Proof Uniqueness of the decomposition $\nu = \nu_a + \nu_s$ is the content of Proposition 6.4. 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 Schwartz inequality 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 the pre-Hilbert space $\mathcal{L}^2(S, \Sigma, \phi)$. Hence there exists, by virtue of the Riesz-Fréchet Theorem 6.1, a $g \in \mathcal{L}^2(S, \Sigma, \phi)$ such that for all f

$$\nu(f) = \phi(fg). \tag{6.7}$$

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 $[0, 1]$. From Lemma 6.6 we obtain that

$\phi(\{g \notin [0, 1]\}) = 0$. Replacing g with $g\mathbf{1}_{\{0 \leq g \leq 1\}}$, we see that (6.7) still holds and hence we may assume that $0 \leq g \leq 1$.

Rewrite (6.7) as

$$\nu(f(1-g)) = \mu(fg) \quad (6.8)$$

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

$$h = \frac{g}{1-g} \mathbf{1}_{\{g < 1\}}, \quad (6.9)$$

and the measure ν_a on Σ by $\nu_a(E) = \mu(\mathbf{1}_E h)$. By Proposition 6.5 this indeed defines a measure and obviously $\nu_a \ll \mu$. Take $f = \frac{\mathbf{1}_{E \cap \{g < 1\}}}{1-g}$ in (6.8) 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 (6.9) 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 trivial details are omitted. \square

Remark 6.8 If we take ν 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 take $h \geq 0$.

Remark 6.9 The function h of Theorem 6.7, 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 $F(x) := \nu((-\infty, x]) = \int_{(-\infty, x]} f d\lambda$, where $f = \frac{d\nu}{d\lambda}$. 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 the tradition. We often follow the accepted terminology by merely saying that X admits a density f , if $f = \frac{d\nu}{d\lambda}$ without mentioning that its law is absolutely continuous w.r.t. Lebesgue measure.

6.6 Additional results

Proposition 6.10 *Let μ be a complex measure. Then $\mu \ll |\mu|$ and the Radon-Nikodym derivative $h = \frac{d\mu}{d|\mu|}$ may be taken such that $|h| = 1$.*

Proof Let h be any function as in the Radon-Nikodym theorem. Since $|\mu|(h\mathbf{1}_E)| = |\mu(E)| \leq |\mu|(E)$, it follows from Lemma 6.6 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 $h\mathbf{1}_{\{|h|=1\}} + \mathbf{1}_{\{|h| \neq 1\}}$. \square

Corollary 6.11 *Let μ be a real measure, $h = \frac{d\mu}{d|\mu|}$. 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 $\mu_1 \leq \mu^+$ and $\mu_2 \leq \mu^-$. 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 $\mu^+(E) = \mu(E \cap \{h = 1\}) \leq \mu_1(E \cap \{h = 1\}) \leq \mu_1(E)$. \square

Proposition 6.12 *If μ is a positive measure and ν a complex measure such that $\nu \ll \mu$, then $|\nu| \ll \mu$ and*

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

Proof Exercise 6.8. \square

6.7 Exercises

6.1 Let μ be a real measure on a space (S, Σ) . Define $\nu : \Sigma \rightarrow [0, \infty)$ 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 6.3.

6.3 Prove a version of Proposition 6.5 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. Show that the laws \mathcal{L}_θ ($\theta > 0$) of XY are not absolutely continuous w.r.t. Lebesgue measure on \mathbb{R} . 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, \dots 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

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

Find the distribution of X . 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 $F : [0, 1] \rightarrow \mathbb{R}$ 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 $x < \frac{1}{4}$. 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 $F(x) = \int_{[0,x]} f(u) du$, $x \in [0, 1]$).

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

$$\frac{d\phi}{d\mu} = \frac{d\phi}{d\nu} \frac{d\nu}{d\mu}.$$

6.8 Prove Proposition 6.12.

6.9 Let ν and μ be positive σ -finite measures on (S, Σ) with $\nu \ll \mu$. Let $h = \frac{d\nu}{d\mu}$. Show that $\nu(\{h = 0\}) = 0$. Show that $\mu(\{h = 0\}) = 0$ iff $\mu \ll \nu$. What is $\frac{d\mu}{d\nu}$ 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.)

$$\frac{d\phi_a}{d\nu} = \frac{h}{k} \mathbf{1}_{\{k > 0\}}.$$

6.11 Consider the measurable space (Ω, \mathcal{F}) and a measurable map $X : \Omega \rightarrow \mathbb{R}^n$ (\mathbb{R}^n is endowed with the usual Borel σ -algebra \mathcal{B}^n). Consider two probability measure \mathbb{P} and \mathbb{Q} on (Ω, \mathcal{F}) and let $P = \mathbb{P}^X$ and $Q = \mathbb{Q}^X$ be the corresponding distributions (laws) on $(\mathbb{R}^n, \mathcal{B}^n)$. Assume that P and Q 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 $f, g : \mathbb{R}^n \rightarrow [0, \infty)$. 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*

$$\omega \mapsto \frac{d\mathbb{P}}{d\mathbb{Q}}(\omega) = \frac{f(X(\omega))}{g(X(\omega))}.$$

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, \dots 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 (i) If $\mathbb{P}(\omega : X_n(\omega) \rightarrow X(\omega)) = 1$, then we say that X_n converges to X almost surely (a.s.).

(ii) If $\mathbb{P}(|X_n - X| > \varepsilon) \rightarrow 0$ for all $\varepsilon > 0$, then we say that X_n converges to X in probability.

(iii) If $\mathbb{E}|X_n - X|^p \rightarrow 0$ (equivalently, $\|X_n - X\|_p \rightarrow 0$) 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: $X_n \xrightarrow{\text{a.s.}} X$, $X_n \xrightarrow{\mathbb{P}} X$ and $X_n \xrightarrow{\mathcal{L}^p} X$ 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)

$$\text{for all } \varepsilon > 0: \mathbb{P}(\exists N : \forall n \geq N : |X_n - X| < \varepsilon) = 1. \quad (7.1)$$

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 $X_n \xrightarrow{\text{a.s.}} X$ and $X_n \xrightarrow{\text{a.s.}} X'$. Let Ω_0 be the set of probability one on which $X_n(\omega) \rightarrow X(\omega)$ and Ω'_0 be the set of probability one on which $X_n(\omega) \rightarrow X'(\omega)$. 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 $\mathbb{P}(X = X') = 1$. \square

The following relations hold between the types of convergence introduced in Definition 7.1.

Proposition 7.3 (i) If $X_n \xrightarrow{a.s.} X$, 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 $X_n \xrightarrow{a.s.} X$.

(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 $X_n \xrightarrow{a.s.} X$, fix $\varepsilon > 0$ and let $A_n = \{|X_n - X| \geq \varepsilon\}$. From (7.1) we know that $\mathbb{P}(\liminf A_n^c) = 1$, or that $\mathbb{P}(\limsup A_n) = 0$. 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 $X_n \xrightarrow{\mathbb{P}} X$.

(ii) Fix $\varepsilon > 0$ and let $E_n = \{|X_n - X| > \varepsilon\}$. The first part of the Borel-Cantelli lemma (Lemma 3.13) 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) Apply Proposition 4.31 to get $(\mathbb{E}|X_n - X|^q)^{1/q} \leq (\mathbb{E}|X_n - X|^p)^{1/p}$. \square

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 $n \rightarrow \infty$ 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. \square

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*

- (a) $X_n \xrightarrow{\mathbb{P}} X$ and
- (b) every subsequence of (X_n) contains a further subsequence that is almost surely convergent to X .

Proof Assume that (a) holds, then for any $\varepsilon > 0$ and any subsequence we also have $\mathbb{P}(|X_{n_k} - X| > \varepsilon) \rightarrow 0$. Hence for every $p \in \mathbb{N}$, there is $k_p \in \mathbb{N}$ such that $\mathbb{P}(|X_{n_{k_p}} - X| > \varepsilon) \leq 2^{-p}$. Now we apply part (ii) of Proposition 7.3, which gives us (b). Conversely, assume that (b) holds. We reason by contradiction. Suppose that (a) doesn't hold. Then there exist an $\varepsilon > 0$ and a level $\delta > 0$ such that along some subsequence (n_k) one has

$$\mathbb{P}(|X_{n_k} - X| > \varepsilon) > \delta, \text{ for all } k. \quad (7.2)$$

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). \square

The following result cannot be a surprise.

Proposition 7.6 *Let X, X_1, X_2, \dots be random variables and $g : \mathbb{R} \rightarrow \mathbb{R}$ be continuous. If $X_n \xrightarrow{a.s.} X$, we also have $g(X_n) \xrightarrow{a.s.} g(X)$ and if $X_n \xrightarrow{\mathbb{P}} X$, then also $g(X_n) \xrightarrow{\mathbb{P}} g(X)$.*

Proof Exercise 7.2. \square

Convergence, almost surely or in probability, of random vectors is defined similarly. For instance, if X, X_1, X_2, \dots 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, \dots and Y, Y_1, Y_2, \dots 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 $\overset{*}{\rightarrow}$ denote any of the three types of convergence $\xrightarrow{a.s.}$, $\xrightarrow{\mathbb{P}}$ and $\xrightarrow{L^p}$. If $X_n \overset{*}{\rightarrow} X$ and $Y_n \overset{*}{\rightarrow} Y$, then $Z_n \overset{*}{\rightarrow} Z$.*

Proof Exercise 7.3 \square

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 3.13 states that $\mathbb{P}(F) = 0$. At the same time, we deduce from Fatou's lemma for sets (Exercise 1.4) the contradiction that $\nu(F) \geq \limsup \nu(F_n) \geq \varepsilon$. \square

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*

$$\lim_{K \rightarrow \infty} \mathbb{E}|X|\mathbf{1}_{\{|X| \geq K\}} = 0.$$

Proof Exercise 7.4. \square

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

$$\lim_{K \rightarrow \infty} \sup\{\mathbb{E}|X|\mathbf{1}_{\{|X| > K\}} : X \in \mathcal{C}\} = 0.$$

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 $K \rightarrow \infty$ 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 $p = 1$.

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 $\mathbb{E}|X_n - X| \rightarrow 0$) iff

(i) $X_n \xrightarrow{\mathbb{P}} X$ and

(ii) (X_n) is uniformly integrable.

Proof Assume that $\mathbb{E}|X_n - X| \rightarrow 0$. 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 $N \in \mathbb{N}$

$$\sup_{n \geq N} \mathbb{E}|X_n|\mathbf{1}_{\{|X_n|>K\}} \leq \sup_{n \geq N} \mathbb{E}|X_n - X| + \sup_{n \geq N} \mathbb{E}|X|\mathbf{1}_{\{|X_n|>K\}}. \quad (7.3)$$

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, 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 the collection $(X_n)_{n \geq N}$ is UI. The proof is concluded by invoking the two assertions of Proposition 7.14.

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 result follows from Proposition 7.4. If (X_n) is merely UI, we use *truncation functions* ϕ_K defined by $\phi_K(x) = x\mathbf{1}_{\{|x| \leq K\}} + \text{sgn}(x)K\mathbf{1}_{\{|x| > K\}}$.

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. This concludes the proof. \square

7.3 Exercises

7.1 Let $X_1, Y_1, X_2, Y_2, \dots$ be an i.i.d. sequence whose members have a uniform distribution on $[0, 1]$ and let $f : [0, 1] \rightarrow [0, 1]$ be continuous. Define $Z_i = \mathbf{1}_{\{f(X_i) > Y_i\}}$.

- (a) Show that $\frac{1}{n} \sum_{i=1}^n Z_i \rightarrow \int_0^1 f(x) dx$ 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 $\mathcal{C}_1, \dots, \mathcal{C}_n$ 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 $G : \mathbb{R}^+ \rightarrow \mathbb{R}^+$ such that $\lim_{t \rightarrow \infty} \frac{G(t)}{t} = \infty$ 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 $\lim_{n \rightarrow \infty} g_n = \infty$. Define $g(t) = \sum_{n=0}^{\infty} \mathbf{1}_{[n, n+1)}(t) g_n$ and $G(t) = \int_0^t g(s) ds$. Check that $\lim_{t \rightarrow \infty} \frac{G(t)}{t} = \infty$.

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 $\int_{|X| \geq k} |X| d\mathbb{P} \geq \sum_{m=k}^{\infty} a_m(X)$. Pick for every n a constant $c_n \in \mathbb{N}$ such that $\int_{|X| \geq c_n} |X| d\mathbb{P} \leq 2^{-n}$. 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 $G : \mathbb{R}^+ \rightarrow \mathbb{R}^+$ such that $\lim_{t \rightarrow \infty} \frac{G(t)}{t} = \infty$ 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, \dots be random variables. Then X_n converges to X in \mathcal{L}^p iff the following two conditions are satisfied.

- (a) $X_n \rightarrow X$ in probability,
 (b) 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, \dots 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 $X_n \xrightarrow{\mathcal{L}^1} X$.

8 Conditional expectation

Let X be a random variable with values in $\{x_1, \dots, x_n\}$ and Y a random variable with values in $\{y_1, \dots, y_m\}$. 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

$$\begin{aligned} \mathbb{E} \hat{X} \mathbf{1}_{E_j} &= \mathbb{E} \hat{x}_j \mathbf{1}_{E_j} = \hat{x}_j \mathbb{P}(E_j) \\ &= \sum_i x_i \mathbb{P}(\{X = x_i\} | E_j) \mathbb{P}(E_j) \\ &= \sum_i x_i \mathbb{P}(\{X = x_i\} \cap E_j) \\ &= \mathbb{E} X \mathbf{1}_{E_j}. \end{aligned}$$

Every event $E \in \sigma(Y)$ is a finite union of events E_j . It then follows that

$$\mathbb{E} \hat{X} \mathbf{1}_E = \mathbb{E} X \mathbf{1}_E, \forall E \in \sigma(Y). \quad (8.1)$$

The random variable \hat{X} is 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.

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.

8.1 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

$$\mathbb{E} \hat{X} \mathbf{1}_G = \mathbb{E} X \mathbf{1}_G, \forall G \in \mathcal{G}. \quad (8.2)$$

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 (Ω, \mathcal{G}) . Moreover, $\nu^+ \ll \mathbb{P}$ and $\nu^- \ll \mathbb{P}$ on this space. According to the Radon-Nikodym Theorem 6.7 there exist nonnegative \mathcal{G} -measurable functions \hat{X}^+ and \hat{X}^- such that $\nu^+(G) = \mathbb{E} \hat{X}^+ \mathbf{1}_G$ and $\nu^-(G) = \mathbb{E} \hat{X}^- \mathbf{1}_G$. These functions are a.s. unique. Then $\hat{X} = \hat{X}^+ - \hat{X}^-$ is a version of $\mathbb{E}[X|\mathcal{G}]$. \square

Remark 8.3 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 8.4 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 8.5 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. \square

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 8.6 *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 $\liminf_{n \rightarrow \infty} \hat{X}_n \geq \hat{\ell}$ a.s., where ℓ is a version of the conditional expectation of $\ell := \liminf_{n \rightarrow \infty} X_n$. (Fatou's lemma for conditional expectations)*

(iii) *If (X_n) is a sequence of random variables such that for some X one has $X_n \rightarrow X$ 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 $\hat{X}_n \rightarrow \hat{X}$ 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. \square

Theorem 8.7 *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 \mathcal{G} -measurable, let Y be independent of \mathcal{G} and 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 $\hat{h}(X)$ is a version of $\mathbb{E}[h(X, Y)|\mathcal{G}]$.*

(v) *If $c : \mathbb{R} \rightarrow \mathbb{R}$ 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 $\mathbb{E}\mathbf{1}_H\tilde{X} = \mathbb{E}\mathbf{1}_HX$, 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. 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 $Z_n \uparrow Z$. Apply the Monotone Convergence Theorem to both sides of $\mathbb{E}\mathbf{1}_GZ_n\hat{X} = \mathbb{E}\mathbf{1}_GZ_nX$ 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 $\mathbb{E}\mathbf{1}_G\mathbf{1}_HX = \mathbb{E}\mathbf{1}_GX\mathbb{P}(H)$ and $\mathbb{E}\mathbf{1}_G\mathbf{1}_H\hat{X} = \mathbb{E}\mathbf{1}_G\hat{X}\mathbb{P}(H)$. It follows that $\mathbb{E}\mathbf{1}_G\mathbf{1}_HX = \mathbb{E}\mathbf{1}_G\mathbf{1}_H\hat{X}$, since \hat{X} is a version of $\mathbb{E}[X|\mathcal{G}]$. Recall 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 $E \mapsto \mathbb{E}\mathbf{1}_EX$ 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. Let V be the collection of all bounded measurable functions for which the statement holds true. 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 $\mathbb{E}\mathbf{1}_G\hat{h}_n(X) = \mathbb{E}\mathbf{1}_Gh_n(X, Y)$. Invoking the Monotone Convergence Theorem again results in $\mathbb{E}\mathbf{1}_G\hat{h}(X) = \mathbb{E}\mathbf{1}_Gh(X, Y)$. Since all $\hat{h}_n(X)$ are \mathcal{G} -measurable, the same holds for $\hat{h}(X)$ and we conclude that $h \in V$.

(v) Since c is convex, there are sequences (a_n) and (b_n) in \mathbb{R} such that $c(x) = \sup\{a_nx + b_n : n \in \mathbb{N}\}$, $\forall x \in \mathbb{R}$. Hence for all n we have $c(X) \geq a_nX + b_n$ 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. \square

Here is an illustrative example.

Example 8.8 Let X_1, X_2, \dots be an *iid* sequence of independent 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 $j \in \{1, \dots, n\}$. 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, $\mathbb{E}[X_1|S_n]$ is also equal to $\mathbb{E}[X_1|\mathcal{G}_n]$, where $\mathcal{G}_n = \sigma(S_n, S_{n+1}, \dots)$. To see this, observe first that $\mathcal{G}_n = \sigma(S_n) \vee \mathcal{H}_n$, where $\mathcal{H}_n = \sigma(X_{n+1}, X_{n+2}, \dots)$ and that \mathcal{H}_n is independent of $\sigma(X_1, S_n)$. Now apply (iii) of Theorem 8.7.

Let $P : L^1(\Omega, \mathcal{F}, \mathbb{P}) \rightarrow L^1(\Omega, \mathcal{G}, \mathbb{P})$ 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 8.9 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 4.38.

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.

8.2 Conditional probabilities

Let $F \in \mathcal{F}$ and \mathcal{G} a sub- σ -algebra of \mathcal{F} . We define $\mathbb{P}(F|\mathcal{G}) := \mathbb{E}[\mathbf{1}_F|\mathcal{G}]$. So a version of $\mathbb{P}(F|\mathcal{G})$ is a \mathcal{G} -measurable random variable $\hat{\mathbb{P}}(F)$ that satisfies

$$\mathbb{P}(F \cap G) = \mathbb{E}[\hat{\mathbb{P}}(F)\mathbf{1}_G], \forall G \in \mathcal{G}.$$

Of course all versions of $\mathbb{P}(F|\mathcal{G})$ are almost surely equal. Moreover, if F_1, F_2, \dots 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

$$\hat{\mathbb{P}}(\cup_{n=1}^{\infty} F_n) = \sum_{n=1}^{\infty} \hat{\mathbb{P}}(F_n). \quad (8.3)$$

Of course the set N in general depends on the sequence of events F_1, F_2, \dots . 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 if $N \neq \emptyset$, 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 8.10 *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 $\{q_1, q_2, \dots\}$ be the set of rationals strictly bigger than q and let $r_n = \inf\{q_1, \dots, q_n\}$. Then $r_n \downarrow q$, as $n \rightarrow \infty$. 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 $\lim_{q \rightarrow \infty} G(q) < 1$ or $\lim_{q \rightarrow -\infty} G(q) > 0$. 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, $\mathcal{C} = \mathcal{B}(\mathbb{R})$. \square

Proposition 8.11 *Let X be a random variable and $h : \mathbb{R} \rightarrow \mathbb{R}$ 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*

$$\int h(x) \hat{\mathbb{P}}^X(dx) \tag{8.4}$$

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 8.10 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 converge 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. \square

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 8.11. The good news is that in most common situations Proposition 8.11 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 .

8.3 Exercises

8.1 Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $\mathcal{A} = \{A_1, \dots, A_n\}$ 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 8.9.

8.4 Prove the conditional version of Fatou's lemma, Theorem 8.6(ii).

8.5 Prove the conditional Dominated Convergence theorem, Theorem 8.6(iii).

8.6 Let (X, Y) have a bivariate normal distribution with $\mathbb{E}X = \mu_X$, $\mathbb{E}Y = \mu_Y$, $\text{Var}X = \sigma_X^2$, $\text{Var}Y = \sigma_Y^2$ and $\text{Cov}(X, Y) = c$. 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 $\mathbb{E}[X|Y] = \mathbb{E}[Y|X]$ (or rather, version of them are a.s. equal). Show that $\mathbb{P}(X = Y) = 1$. *Hint:* Start to work on $\mathbb{E}(X - Y)\mathbf{1}_{\{X > z, Y \leq z\}} + \mathbb{E}(X - Y)\mathbf{1}_{\{X \leq z, Y > z\}}$ 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 $\hat{h}(y) = \int_{\mathbb{R}} h(x)\hat{f}(x|y) dx$. Show that $\hat{h}(Y)$ is a version of $\mathbb{E}[h(X)|Y]$. Show also that

$$\hat{\mathbb{P}}(E) = \int_E \hat{f}(x|Y) dx$$

defines a regular conditional probability on $\mathcal{B}(\mathbb{R})$ given Y . What is the exceptional set N of section 8.2?

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 $\hat{\mathbb{P}}$ is a probability measure on \mathcal{F} . Show that there exists no regular conditional probability $\hat{\mathbb{P}}$ on \mathcal{F} given \mathcal{B} .

9 Martingales and alike

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 $\{0, 1, 2, \dots\}$ in which case we have e.g. a sequence of random variables X_0, X_1, X_2, \dots . 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 $\mathcal{F}_n^X := \sigma(X_0, \dots, X_n)$. 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 $\mathcal{F} = \mathcal{F}^X$ for some process X , then another process Y is \mathbb{F}^X -adapted, iff for all n , there exists a Borel function $f_n : \mathbb{R}^{n+1} \rightarrow \mathbb{R}$ such that $Y_n = f_n(X_0, \dots, X_n)$, 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, \dots, 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

$$\mathbb{E}[M_{n+1} | \mathcal{F}_n] = M_n \text{ a.s.} \quad (9.1)$$

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 *iid* random variables X_n , with $n \geq 1$ and assume that $X_1 \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$. 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 8.5 and Theorem 8.7) 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 $\mathbb{E}X_1 = 0$. We conclude that a martingale is an extension of the partial sum process generated by a sequence of *iid* 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 *iid* random variables X_n , with $n \geq 1$ and assume that $X_1 \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$. 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 S_n belongs to $\mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$, because of the product rule for products of independent random variables, see Proposition 4.30. 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 $\mathbb{E}X_1 = 1$.

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 8.7(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 8.9. 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- en 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

$$\mathbb{E}[X_{n+1}|\mathcal{F}_n] \geq X_n \text{ a.s.} \quad (9.2)$$

Equation (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.

Equations (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

$$\Delta X_n = X_n - X_{n-1}, \quad n \geq 1.$$

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 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}^ξ -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 $S_t = \int_0^t Y_s dX_s$. Such an expression, for suitable 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 8.7(ii), and the definitions of martingale and sub-, supermartingale. \square

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*. A formal definition follows.

Definition 9.11 Let \mathbb{F} be a filtration. A mapping $T : \Omega \rightarrow \{0, 1, 2, \dots\} \cup \{\infty\}$ is called a stopping time if for all $n \in \{0, 1, 2, \dots\}$ 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 $n \in \{0, 1, 2, \dots\}$. 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 . \square

Let T be a stopping time and $n \in \{0, 1, 2, \dots\}$. 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 $\{\xi_1 = \dots = \xi_{n-1} = -1\}$ 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 $\mathbb{P}(T = n) = \mathbb{P}(\xi_1 = \dots = \xi_{n-1} = -1, \xi_n = 1) = 2^{-n}$. 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 not bounded.

Theorem 9.15 *Let X be supermartingale and T an a.s. finite stopping time. Assume either*

- (i) X 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 $\mathbb{E} T < \infty$

Then $\mathbb{E} X_T \leq \mathbb{E} X_0$.

If X is a martingale, then we get $\mathbb{E} X_T = \mathbb{E} X_0$ 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 $X_n^T \xrightarrow{a.s.} X_T$, since $T < \infty$ a.s.

Assume (i). By assumption, $X_n^T - Y \geq 0$. Hence we can apply Fatou's lemma and Proposition 9.13 to get $\mathbb{E}(X_T - Y) = \mathbb{E} \lim_{n \rightarrow \infty} (X_n^T - Y) \leq \liminf_{n \rightarrow \infty} \mathbb{E}(X_n^T - Y) = \liminf_{n \rightarrow \infty} \mathbb{E} X_n^T - \mathbb{E} Y \leq \mathbb{E} X_0 - \mathbb{E} Y$. 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 $\mathbb{E}(X^T - X_0) = \lim_{n \rightarrow \infty} \mathbb{E}(X_n^T - X_0) \leq 0$.

If X is a martingale, then the assertion follows as above under (ii) and (iii), whereas under (iv) it follows by dominated convergence. \square

9.3 Doob's decomposition

The celebrated decomposition theorem by Doob for submartingales is presented as part of the next proposition. It says that a submartingale can be decomposed as the sum of a predictable increasing *trend* and a martingale.

Proposition 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 as 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

Corollary 9.17 *Let M be a martingale with $\mathbb{E} M_n^2 < \infty$ for all n . Then there exists a unique (in the sense of Proposition 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.

$$\begin{aligned} \mathbb{E}[(M_n - M_{n-1})^2|\mathcal{F}_{n-1}] &= \mathbb{E}[M_n^2 - 2M_n M_{n-1} + M_{n-1}^2|\mathcal{F}_{n-1}] \\ &= \mathbb{E}[M_n^2|\mathcal{F}_{n-1}] - M_{n-1}^2 \\ &= \mathbb{E}[M_n^2 - M_{n-1}^2|\mathcal{F}_{n-1}], \end{aligned}$$

which is by the proof of Proposition 9.16 just the definition of $\Delta\langle M\rangle_n$. \square

Remark 9.18 The process $\langle M\rangle$ of Corollary 9.17 is also called the *predictable quadratic variation process* of M . 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 $X_T = \sum_{n=0}^{\infty} X_n 1_{\{T=n\}}$ and so $X_T = X_T 1_{\{T<\infty\}}$. 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!).

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*

$$X_T \geq \mathbb{E}[X_N|\mathcal{F}_T] \quad \text{a.s.} \tag{9.3}$$

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

$$\begin{aligned} \mathbb{E}[X_N 1_F] &= \sum_{n=0}^N \mathbb{E}[X_N 1_F 1_{\{T=n\}}] \\ &\geq \sum_{n=0}^N \mathbb{E}[X_n 1_F 1_{\{T=n\}}] \\ &= \sum_{n=0}^N \mathbb{E}[X_T 1_F 1_{\{T=n\}}] \\ &= \mathbb{E}[X_T 1_F], \end{aligned}$$

which is what we wanted to prove. \square

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 $\mathbb{E}|X_T| 1_{\{T=\infty\}} = \mathbb{E}|X_\infty| 1_{\{T=\infty\}} \leq \mathbb{E}|X_\infty| < \infty$. Next, because $|X|$ is a submartingale with last element $|X_\infty|$,

$$\begin{aligned} \mathbb{E}|X_T| 1_{\{T < \infty\}} &= \sum_{n=0}^{\infty} \mathbb{E}|X_n| 1_{\{T=n\}} \\ &\leq \sum_{n=0}^{\infty} \mathbb{E}|X_\infty| 1_{\{T=n\}} \\ &= \mathbb{E}|X_\infty| 1_{\{T < \infty\}} < \infty. \end{aligned}$$

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

$$\mathbb{E}[X_\infty 1_{F \cap \{T \leq n\}}] = \mathbb{E}[X_{T \wedge n} 1_{F \cap \{T \leq n\}}] = \mathbb{E}[X_T 1_{F \cap \{T \leq n\}}].$$

Let $n \rightarrow \infty$ and apply the Dominated convergence theorem to get

$$\mathbb{E}[X_\infty 1_F 1_{\{T < \infty\}}] = \mathbb{E}[X_T 1_F 1_{\{T < \infty\}}].$$

Together with the trivial identity $\mathbb{E}[X_\infty 1_F 1_{\{T=\infty\}}] = \mathbb{E}[X_T 1_F 1_{\{T=\infty\}}]$ this yields $\mathbb{E}[X_\infty 1_F] = \mathbb{E}[X_T 1_F]$ 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 $n = 0, 1, \dots$. 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 $0 \leq \mathbb{E}[-X_T 1_{\{T < \infty\}}] < \infty$.

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

$$\mathbb{E}[X_{S \wedge n} 1_E 1_{\{S \leq n\}}] \leq \mathbb{E}[X_{T \wedge n} 1_E 1_{\{S \leq n\}}] \leq \mathbb{E}[X_{T \wedge n} 1_E 1_{\{T \leq n\}}]$$

and hence

$$\mathbb{E}[X_S 1_E 1_{\{S \leq n\}}] \leq \mathbb{E}[X_T 1_E 1_{\{T \leq n\}}].$$

The Monotone convergence theorem yields $\mathbb{E}[X_S 1_E] \leq \mathbb{E}[X_T 1_E]$. □

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 $\omega = (\omega_1, \omega_2, \dots)$ a typical element of Ω . Let $X_n : \Omega \rightarrow \mathbb{R}$ be defined by $X_n(\omega) = \omega_n$ ($n \geq 1$) and put $\mathcal{F}_n = \sigma(X_1, \dots, X_n)$. 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. Show that $f(X)$ is again a submartingale, if $\mathbb{E}|f(X_n)| < \infty$ for all n . 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, \dots be independent random variables and $\mathcal{F}_n = \sigma(Z_1, \dots, Z_n)$. Show that (you may assume sufficient integrability) that $X_n = f_n(Z_1, \dots, Z_n)$ defines a martingale under the condition that $\mathbb{E}f_n(z_1, \dots, z_{n-1}, Z_n) = f_{n-1}(z_1, \dots, z_{n-1})$ 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 finite stopping time.

9.9 Let X_1, X_2, \dots be an *iid* sequence of Bernoulli random variables with probability of success equal to p . Put $\mathcal{F}_n = \sigma(X_1, \dots, X_n)$, $n \geq 1$. Let M be a martingale adapted to the generated 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, \dots 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 .

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 $x_n \rightarrow x \in \mathbb{R}$ 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 $U(a, b) = \lim_{N \rightarrow \infty} U_N(a, b)$. $U_N(a, b)$ 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 a random 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 $B_n, S_n : \Omega \rightarrow \{0, 1, 2, \dots\} \cup \{\infty\}$. 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}_∞ -measurable.*

Proof Exercise 10.1. □

Next 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 $Z_N \geq (b - a)U_N - (X_N - a)^-$.*

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 $Z_N = Z_{B_{U_{N+1}}} + (Z_N - Z_{B_{U_{N+1}}}) = Z_{S_N} + (X_N - X_{B_{U_{N+1}}}) \geq (b - a)U_N(a, b) + X_N - a \geq (b - a)U_N(a, b) - (X_N - a)^-$. Combining the two cases, we arrive at the assertion. \square

Proposition 10.4 *Let X be a supermartingale that is bounded in \mathcal{L}^1 (i.e. there is $M > 0$ 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 also 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 $(b - a)\mathbb{E}U(a, b) \leq |a| + M$. \square

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 $X_n \xrightarrow{a.s.} X_\infty$.*

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 now N be the event that X_n doesn’t have an a.s. 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. \square

Remark 10.6 Notice that the theorem only states that a.s. convergence hold, 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 would extend to $n = \infty$, i.e. $\mathbb{E}[X_\infty | \mathcal{F}_n] = M_n$ a.s. But this 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 that is bounded from below has a 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 $X_n \xrightarrow{a.s.} X_\infty$.*

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 Y admits an a.s. limit and so does X . \square

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 $X_n \rightarrow X_\infty$ both almost surely and in \mathcal{L}^1 . Moreover, the extended supermartingale property holds, i.e.*

$$\mathbb{E}[X_\infty | \mathcal{F}_n] \leq X_n. \quad (10.1)$$

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 $m \rightarrow \infty$ by \mathcal{L}^1 convergence. The result for supermartingales follows. When X is martingale, one applies the same reasoning to $|\mathbb{E}\mathbf{1}_G X_\infty - \mathbb{E}\mathbf{1}_G X_n|$. \square

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_G for any version of $\mathbb{E}[\xi | \mathcal{G}]$. Then the collection $\mathcal{C} = \{X_G : \mathcal{G} \in \mathbb{G}\}$ 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

$$\sup\{\mathbb{E}|X_{\mathcal{G}}|\mathbf{1}_{\{|X_{\mathcal{G}}|>k}\} : \mathcal{G} \in \mathbb{G}\} < \varepsilon. \quad (10.2)$$

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 8.7, we have

$$|X_{\mathcal{G}}| \leq \mathbb{E}[|\xi| | \mathcal{G}] \quad (10.3)$$

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 \square

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 $\nu_\infty(G) := \mathbb{E}\mathbf{1}_GX_\infty$ define finite measures on \mathcal{F}_∞ (Exercise 4.9) with $\nu_\xi(\Omega) = \nu_\infty(\Omega) = \mathbb{E}\xi$. Moreover, they coincide on the algebra $\cup_{n=1}^\infty \mathcal{F}_n$, since for $G \in \mathcal{F}_n$ one has $\nu_\xi(G) = \mathbb{E}\mathbf{1}_GX_n$ by the definition of X and $\nu_\infty(G) = \mathbb{E}\mathbf{1}_GX_n$ by Theorem 10.8. We conclude that ν_ξ and ν_∞ are the same on \mathcal{F}_∞ . 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}_∞ -measurable, they must be equal. \square

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 $X_n \rightarrow X_{-\infty}$ both a.s. and in \mathcal{L}^1 as $n \rightarrow -\infty$. 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 $\mathbb{E}U_{-N}(a, b) \leq \mathbb{E}(X_{-1}-a)^-/(b-a)$. Hence also $U_{-\infty}(a, b) := \lim_{N \rightarrow \infty} U_{-N}(a, b)$ 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. \square

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 $\|X_n - X_\infty\|_p \rightarrow 0$, equivalently $\mathbb{E}|X_n - X_\infty|^p \rightarrow 0$. It turns out useful to formulate most results in terms of submartingales. Here is the first one.

Proposition 10.12 *Let X be a nonnegative submartingale and write $X_n^* = \max\{X_1, \dots, X_n\}$. For $\lambda > 0$ it holds that*

$$\lambda \mathbb{P}(X_n^* \geq \lambda) \leq \mathbb{E} \mathbf{1}_{\{X_n^* \geq \lambda\}} X_n. \quad (10.4)$$

It then follows that $\lambda \mathbb{P}(X_n^ \geq \lambda) \leq \mathbb{E} X_n$.*

Proof Let $T = \inf\{n \geq 0 : X_n \geq \lambda\}$. Then T is a stopping time and $\{X_n^* \geq \lambda\} = \{T \leq n\}$. It holds that $E_k := \{T = k\} \subset \{X_k \geq \lambda\}$ and by the submartingale property of X we have $\mathbb{E} \mathbf{1}_{E_k} X_k \leq \mathbf{1}_{E_k} X_n$, since $E_k \in \mathcal{F}_k$. Then we get $\lambda \mathbb{P}(E_k) = \lambda \mathbb{E} \mathbf{1}_{E_k} \leq \mathbb{E} \mathbf{1}_{E_k} X_k \leq \mathbb{E} \mathbf{1}_{E_k} X_n$. Equation (10.4) follows by summing over k . \square

Remark 10.13 Let ξ_1, ξ_2, \dots be a sequence of independent random variables with expectation zero and finite second moment. Let $M_n = \sum_{k=1}^n \xi_k$ and $M_n^* = \max\{|M_1|, \dots, |M_n|\}$. 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*

$$\lambda\mathbb{P}(Y \geq \lambda) \leq \mathbb{E} \mathbf{1}_{\{Y \geq \lambda\}} X \quad (10.5)$$

holds for all $\lambda > 0$. If $X \in \mathcal{L}^p(\Omega, \mathcal{F}, \mathbb{P})$ for $p > 1$, 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 Since (10.5) holds for all $\lambda > 0$ we can integrate both sides after multiplication with $p\lambda^{p-2}$. On the left we get the integral $\int_0^\infty p\lambda^{p-1}\mathbb{P}(Y \geq \lambda) d\lambda = \mathbb{E} Y^p$, in view of Exercise 5.12. On the right we compute using Fubini's theorem the integral

$$\begin{aligned} \int_0^\infty p\lambda^{p-2}\mathbb{E} \mathbf{1}_{\{Y \geq \lambda\}} X d\lambda &= \mathbb{E} \int_0^\infty p\lambda^{p-2} \mathbf{1}_{\{Y \geq \lambda\}} X d\lambda \\ &= \mathbb{E} \left(X \int_0^Y p\lambda^{p-2} d\lambda \right) \\ &= \frac{p}{p-1} \mathbb{E} X Y^{p-1}. \end{aligned}$$

We have established $\|Y\|_p^p = \mathbb{E} Y^p \leq q \mathbb{E} X Y^{p-1}$. It follows from Hölder's inequality (Theorem 4.33) that $\mathbb{E} X Y^{p-1} \leq (\mathbb{E} X^p)^{1/p} (\mathbb{E} Y^p)^{1/q} = \|X\|_p (\|Y\|_p)^{p/q}$ 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 $\|Y\|_p < \infty$. Notice that the event $\{Y_n \geq \lambda\}$ is empty for $n < \lambda$ and equal to $\{Y \geq \lambda\}$ otherwise. It follows that (10.5) 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. \square

Here is the result we were aiming at.

Theorem 10.15 *Let X be a nonnegative submartingale that is bounded in \mathcal{L}^p for $p > 1$. Let $X^* = \sup_n X_n$.*

(i) It holds that $X^ \in \mathcal{L}^p(\Omega, \mathcal{F}_\infty, \mathbb{P})$ and $\|X^*\|_p \leq q \sup_n \|X_n\|_p$, where $q = \frac{p}{p-1}$.*

(ii) The a.s. limit $X_\infty = \lim_{n \rightarrow \infty} X_n$ exists and moreover, $X_n \xrightarrow{\mathcal{L}^p} X_\infty$ and $\|X_\infty\|_p = \sup_n \|X_n\|_p$.

Proof (i) Put $X_n^* = \max\{X_1, \dots, X_n\}$. Combining Proposition 10.12 and Lemma 10.14, we get the inequality $\|X_n^*\|_p \leq q\|X_n\|_p$. If $K = \sup_n \|X_n\|_p < \infty$, then we obtain by monotone convergence $\|X^*\|_p \leq qK$.

(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 $\|X_n - X_\infty\|_p \leq 2\|X^*\|_p < \infty$. By the Dominated Convergence Theorem we get $X_n \xrightarrow{\mathcal{L}^p} X_\infty$. Finally, $\|X_\infty\|_p \leq \|X_\infty - X_n\|_p + \sup_n \|X_n\|_p$ yields the last assertion. \square

Corollary 10.16 *Let M be a martingale that is bounded in \mathcal{L}^p for $p > 1$. The a.s. limit $M_\infty = \lim_{n \rightarrow \infty} M_n$ exists and moreover, $M_n \xrightarrow{\mathcal{L}^p} M_\infty$.*

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. \square

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 $\lim_{n \rightarrow \infty} w_{nk} = 0$ for every $k \geq 1$ and $\lim_{n \rightarrow \infty} \sum_{k=1}^n w_{nk} = 1$. Let (x_n) be sequence of real numbers with $\lim_{n \rightarrow \infty} x_n = x \in \mathbb{R}$. Put $\bar{x}_n = \sum_{k=1}^n w_{kn} x_k$. Then $\lim_{n \rightarrow \infty} \bar{x}_n = x$.*

(ii) *Let $(w_n)_{n \geq 0}$ be an increasing sequence of nonnegative real numbers with $w_0 = 0$ and $w_n \rightarrow \infty$. 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 $\frac{1}{w_n} \sum_{k=1}^n x_k \rightarrow 0$.*

Proof (i) Let $\varepsilon > 0$ and choose m such that $|x_n - x| < \varepsilon$ for $n > m$. Then we have for $n > m$

$$\begin{aligned} |\bar{x}_n - x| &\leq \sum_{k=1}^n w_{kn} |x_k - x| + \left| \sum_{k=1}^n w_{kn} - 1 \right| |x| \\ &\leq \sum_{k=1}^m w_{kn} |x_k - x| + \sum_{k=m+1}^n w_{kn} |x_k - x| + \left| \sum_{k=1}^n w_{kn} - 1 \right| |x| \\ &\leq \sum_{k=1}^m w_{kn} |x_k - x| + \varepsilon \sum_{k=m+1}^n w_{kn} + \left| \sum_{k=1}^n w_{kn} - 1 \right| |x|. \end{aligned}$$

It follows from the assumptions that $\limsup_{n \rightarrow \infty} |\bar{x}_n - x| \leq \varepsilon$. Since $\varepsilon > 0$ is arbitrary the result follows.

(ii) Given $n \geq 1$, 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 $y_n \rightarrow 0$. We compute

$$\begin{aligned} \frac{1}{w_n} \sum_{k=1}^n x_k &= -\frac{1}{w_n} \sum_{k=1}^n w_k (y_k - y_{k-1}) \\ &= \frac{1}{w_n} \left(-w_n y_n + \sum_{k=1}^n y_k (w_k - w_{k-1}) \right) \\ &= -y_n + \sum_{k=1}^n w_{kn} y_k. \end{aligned}$$

Application of part (i) yields the result. \square

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 $1 \leq k \leq n$ 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 $\frac{1}{\langle M \rangle_n} M_n \rightarrow 0$ a.s. on the set $\{\langle M \rangle_n \rightarrow \infty\}$.*

Proof Let X be the martingale transform $X = \frac{1}{1+\langle M \rangle} \cdot M$. Then

$$\begin{aligned} \mathbb{E} [\Delta X_n^2 | \mathcal{F}_{n-1}] &= \frac{1}{(1 + \langle M \rangle)^2} \mathbb{E} [\Delta M_n^2 | \mathcal{F}_{n-1}] \\ &= \frac{1}{(1 + \langle M \rangle)^2} \Delta \langle M \rangle_n \\ &\leq \frac{1}{1 + \langle M \rangle_{n-1}} - \frac{1}{1 + \langle M \rangle_n}. \end{aligned}$$

It follows that $\Delta \langle X \rangle_n \leq \frac{1}{1 + \langle M \rangle_{n-1}} - \frac{1}{1 + \langle M \rangle_n}$ and hence that $\langle X \rangle_n \leq 1$ 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. \square

Corollary 10.20 *Let X_1, X_2, \dots be an independent sequence with $\mathbb{E} X_k = 0$ and $\sigma_k^2 := \text{Var} X_k < \infty$ for all k . Let $\alpha_n = \sum_{k=1}^n \sigma_k^2$. If $\alpha_n \rightarrow \infty$, then $\frac{1}{\alpha_n} \sum_{k=1}^n X_k \rightarrow 0$ 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. \square

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 $\frac{1}{n} \sum_{k=1}^n X_k \rightarrow 0$, 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, \dots be a sequence of random variables. Define $\mathcal{T}_n = \sigma(X_{n+1}, X_{n+2}, \dots)$ and $\mathcal{T} = \bigcap_{n=0}^{\infty} \mathcal{T}_n$, \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 *iid* 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 iid sequence. Then $\mathbb{P}(F) \in \{0, 1\}$.*

Proof Put $\mathcal{F}_n = \sigma(X_1, \dots, X_n)$. 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. \square

Example 10.22 Here is a first application. Let X_1, X_2, \dots 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 $\bar{X} = \limsup \frac{1}{n} (X_{n+1}, \dots, X_{n+m})$, 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, \dots be an iid sequence in $\mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$. Let $\mu = \mathbb{E} X_1$ and $M_{-n} = \frac{1}{n} \sum_{k=1}^n X_k$. Then $M_{-n} \rightarrow \mu$ a.s. and in \mathcal{L}^1 as $n \rightarrow \infty$.*

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 to be a clever one. Let $S_n = \sum_{k=1}^n X_k$ and put $\mathcal{F}_{-n} = \sigma(S_n, S_{n+1}, \dots)$, $n \geq 1$. In Example 8.8 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 $M_{-n} \rightarrow M_{-\infty}$ 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 . \square

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 $k = 0, \dots, 2^n - 1$ and \mathcal{F}_n be the σ -algebra by the I_k^n for $k = 0, \dots, 2^n - 1$. 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 $X_n \rightarrow X_\infty$ 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 $X_n \rightarrow X_\infty$ 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 $Y \in \mathcal{L}^1$, (\mathcal{F}_n) and 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 $X_n \rightarrow X_\infty$ a.s. Show that for $Y \in \mathcal{L}^2$, we have $X_n \xrightarrow{\mathcal{L}^2} X_\infty$. Find a condition such that $X_\infty = Y$. Give also an example in which $P(X_\infty = Y) = 0$.

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) $\lim_{n \rightarrow -\infty} \mathbb{E}X_n < \infty$.

(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 (Ω, \mathcal{F}) such that $\mathbb{Q} \ll \mathbb{P}$ with $d\mathbb{Q}/d\mathbb{P} = M_\infty$. 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

$$\frac{d\mathbb{Q}_n}{d\mathbb{P}_n} = M_n,$$

where $M_n = \mathbb{E}_{\mathbb{P}}[M_\infty|\mathcal{F}_n]$.

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 and k 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 $d\mathbb{Q}/d\mathbb{P}$.

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

$$\mathbb{E}[X_n | \mathcal{F}_n] \xrightarrow{\text{a.s.}} \mathbb{E}[X | \mathcal{F}_\infty].$$

(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

$$\mathbb{P}(\sup_{k \leq n} M_k \geq x) \leq \exp\left(-\frac{x^2}{2 \sum_{i=1}^n c_i^2}\right).$$

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, \dots 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 w.r.t. an obvious filtration (which one?). 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

$$\prod_{k=1}^{\infty} r_k > 0. \tag{10.6}$$

Show that N is bounded in \mathcal{L}^2 and that consequently M is uniformly integrable.

- (c) Show that (10.6) implies that $\mathbb{E} M_\infty = 1$.
- (d) Show that $\mathbb{E} M_\infty = 1$ implies that (10.6) holds. *Hint:* Reason by contradiction.

10.13 Let X_1, X_2, \dots be real valued functions defined on some Ω and take $\mathcal{F} = \sigma(X_1, X_2, \dots)$. Assume there exist probability measures \mathbb{P} and \mathbb{Q} on (Ω, \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 $\mathbb{P} \perp \mathbb{Q}$. This phenomenon is known as *Kakutani's dichotomy*, or *Kakutani's alternatives*.

10.14 Consider the setup of Exercise 10.13. Assume that

$$\prod_{k=1}^n \mathbb{E}_{\mathbb{P}} \sqrt{\frac{g(X_k)}{f(X_k)}} \rightarrow 0.$$

Suppose one observes X_1, \dots, 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*: $\mathbb{Q}(M_n \leq c_n) \rightarrow 0$. (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 $N_n \rightarrow N_\infty$ a.s. Show (use Kronecker's lemma) that $\frac{M_n}{(1+A_n)^p}$ has an a.s. finite limit.

11 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 so far. First we sketch a functional analytic background of weak convergence. After that we go back to our probabilistic environment.

Consider a normed vector space $(X, \|\cdot\|)$, often a Banach space, and let X^* be the vector space of all continuous linear functionals $T : X \rightarrow \mathbb{R}$, also called the (strong) dual space of X . The operator norm of $T \in X^*$ is defined as

$$\|T\| = \sup\left\{\frac{|Tx|}{\|x\|} : \|x\| \neq 0\right\}.$$

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, $\|x_n - x\| \rightarrow 0$. If $T \in X^*$, we then have

$$|Tx_n - Tx| \leq \|T\| \|x_n - x\| \rightarrow 0. \quad (11.1)$$

If a sequence (x_n) satisfies (11.1) 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 $\|T_n - T\| \rightarrow 0$. Of course, we then also have for all $y^* \in Y^*$ that

$$|y^*T_n - y^*T| \rightarrow 0. \quad (11.2)$$

Paralleling the above, we say that T_n converges weakly to T if (11.2) 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 (11.2) for $y^* = x(\cdot)$ and we have

$$|T_n x - Tx| \rightarrow 0, \forall x \in X. \quad (11.3)$$

If (11.3) 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 4.35, 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 $\mu : f \rightarrow \mu(f)$ 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 $\mu_n(f) \rightarrow \mu(f)$, 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. You can view $C_b(\mathbb{R})$ as a space of *test functions*. Other choices for 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 $\pm\infty$, $C_0(\mathbb{R})$. To distinguish between the convergence concepts related to the different spaces of test functions, 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 11.15.

Below we will adhere to the custom followed in probability theory.

11.1 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 11.1 Let μ, μ_1, μ_2, \dots 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 $\mu_n(f) \rightarrow \mu(f)$ for all $f \in C_b(\mathbb{R})$. If X, X_1, X_2, \dots are (real) random variables (possibly defined on different probability spaces) with distributions μ, μ_1, μ_2, \dots 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 11.1. Let (x_n) be a convergent sequence, suppose with $\lim x_n = 0$. Then for every $f \in C_b(\mathbb{R})$ one has $f(x_n) \rightarrow f(0)$. 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 $\mu_n(B) \rightarrow \mu(B)$ 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 $\mu_n(B) = 1$ and $\mu(B) = 1$. 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 $\mu_n(B) \rightarrow \mu(B)$ 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 $F_n(x) \rightarrow F(x)$, for all $x \in \mathbb{R}$ *except for* $x = 0$.

Suppose that a random variable X has distribution μ . Recall from Proposition 4.24 that $\mathbb{E}f(X) = \mathbb{E}f \circ X = \int f d\mu$. It follows that $X_n \xrightarrow{w} X$ iff $\mathbb{E}f(X_n) \rightarrow \mathbb{E}f(X)$ 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 given 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 11.2 *Let the random variables X, X_1, X_2, \dots be defined on a single probability space. If $X_n \xrightarrow{a.s.} X$ or if $X_n \xrightarrow{\mathbb{P}} X$, then $X_n \xrightarrow{w} X$. If $X_n \xrightarrow{w} X$ and $g : \mathbb{R} \rightarrow \mathbb{R}$ 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 $X_n \xrightarrow{a.s.} X$. If $f \in C_b(\mathbb{R})$, then also $f(X_n) \xrightarrow{a.s.} f(X)$. Since f is bounded, we can apply the Dominated Convergence Theorem to get the assertion. The remainder of the proof is left as Exercise 11.1. \square

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 11.3 *Let μ, μ_1, μ_2 be a sequence of probability measures on $(\mathbb{R}, \mathcal{B})$ and let F, F_1, F_2, \dots 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*

$x \in \mathbb{R}$. Then $\lim F_n(x) = F(x)$ for all $x \in C_F$, the set of points where F is continuous.

Proof By definition, $F(x) = \mu((-\infty, x]) = \int \mathbf{1}_{(-\infty, x]} d\mu$, 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. \square

The last proposition has a converse statement, which can be directly proved, see Exercise 11.5, but we prefer to let it follow from Skorohod's representation theorem, Theorem 11.4, as a trivial consequence. Note that this theorem is a subtle kind of converse to Proposition 11.2.

Theorem 11.4 *Let F, F_1, F_2, \dots be distribution functions satisfying $F_n(x) \rightarrow F(x)$ 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, \dots defined on it such that X has distribution function F , the X_n have distribution functions F_n and $X_n \xrightarrow{a.s.} X$.*

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 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. \square

Corollary 11.5 Let μ, μ_1, μ_2, \dots be probability measures on $(\mathbb{R}, \mathcal{B})$ with distribution functions F, F_1, F_2, \dots . Equivalence holds between (i) $F_n(x) \rightarrow F(x)$ for all $x \in C_F$ and (ii) $\mu_n \xrightarrow{w} \mu$.

Proof In view of Proposition 11.3 it is sufficient to establish (i) \Rightarrow (ii). This implication follows by combining Proposition 11.2 and Theorem 11.4. \square

Here is a result that gives an appealing sufficient condition for weak convergence, when the random variables involved admit densities.

Theorem 11.6 Consider real random variables X, X_1, X_2, \dots having densities f, f_1, f_2, \dots w.r.t. Lebesgue measure λ . Suppose that $f_n \rightarrow f$ λ -a.e. Then $X_n \xrightarrow{w} X$.

Proof Apply Scheffé's lemma, Lemma 4.19, to conclude that $f_n \rightarrow f$ in $\mathcal{L}^1(\mathbb{R}, \mathcal{B}, \lambda)$. Let $g \in C_b(\mathbb{R})$. Since g is bounded, we also have $f_n g \rightarrow fg$ in $\mathcal{L}^1(\mathbb{R}, \mathcal{B}, \lambda)$ and hence $X_n \xrightarrow{w} X$. \square

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}^∞ . But if extra properties are imposed, there can still be an affirmative answer. Something like that happens in the next proposition that is known as Helly's selection principle. 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 $\lim_{x \rightarrow \infty} F(x) = 1$ and $\lim_{x \rightarrow -\infty} F(x) = 0$ fails to hold.

Proposition 11.7 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 $F_{n_k}(x) \rightarrow F(x)$, 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 $\mathbb{Q} = \{q_1, q_2, \dots\}$ 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 (n_k^2) , we have a limit, $\ell(q_2)$ say. Note that along the thinned subsequence, we still have the limit $\lim_{k \rightarrow \infty} F_{n_k^2}(q_1) = \ell(q_1)$. Continue like this to construct a *nested* sequence of subsequence (n_k^j) for which we have that $\lim_{k \rightarrow \infty} F_{n_k^j}(q_i) = \ell(q_i)$ holds for every $i \leq j$. Put $n_k = n_k^k$, then (n_k) is a subsequence of (n_k^i) for every $i \leq k$. It follows that $\lim_{k \rightarrow \infty} F_{n_k}(q_i) = \ell(q_i)$ for all q_i . In this way we have constructed a function $\ell : \mathbb{Q} \rightarrow [0, 1]$ 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 $F(q) \geq \ell(q)$ 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 $\lim_{x \rightarrow \infty} F(x) < 1$, 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 $F_{n_k}(x) \rightarrow F(x)$ 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 $F(y) \leq \ell(r)$, hence $F(x) < \ell(r) + \varepsilon$. So we have the inequalities

$$\ell(q) - \varepsilon < F(x) < \ell(r) + \varepsilon.$$

Then $\limsup F_{n_k}(x) \leq \lim F_{n_k}(q) = \ell(q) < F(x) + \varepsilon$ and $\liminf F_{n_k}(x) \geq \liminf F_{n_k}(r) = \ell(r) > F(x) - \varepsilon$. The result follows since ε is arbitrary. \square

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 $\lim_{n \rightarrow \infty} F_n(x) = 0$. Hence any limit function F in Proposition 11.7 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, $\mu(\mathbb{R}) \leq 1$. 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 $n \rightarrow \infty$, 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 measure 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 $\lim_{M \rightarrow \infty} \mu([-M, M]) = 1$. If F is the distribution function of μ , we equivalently have $\lim_{M \rightarrow \infty} (F(M) - F(-M)) = 1$. Note that $F(M) - F(-M) = \mu((-M, M])$. The next condition, *tightness*, gives a uniform version of this.

Definition 11.8 A sequence of probability measures (μ_n) on $(\mathbb{R}, \mathcal{B})$ is called tight, if $\lim_{M \rightarrow \infty} \inf_n \mu_n([-M, M]) = 1$.

Remark 11.9 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 11.10 (i) Let (μ_n) be 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 distribution function F and a subsequence (F_{n_k}) such that $F_{n_k}(x) \rightarrow F(x)$, for all $x \in C_F$.

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 $F_n(\pm M) \rightarrow F(\pm M)$. 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 Proposition 11.7 and choose N such that $|F_{n_k}(\pm M) - F(\pm M)| < \varepsilon$. 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. \square

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 sequence 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 basically nothing is known of the *joint* distribution of (X_n, Y_n) . Under extra conditions something can be said though.

Proposition 11.11 *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, we also have weak convergence of $(g(X_n, Y_n))$ for a continuous $g : \mathbb{R}^2 \rightarrow \mathbb{R}$. In particular there holds weak convergence for $X_n + Y_n$.*

Proof Exercise 11.2 □

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 $\partial E = \text{Cl } E \setminus \text{Int } E$.

Theorem 11.12 *Let μ, μ_1, μ_2, \dots be probability measures on some metric space (S, d) . The following statements are equivalent.*

(i) $\mu_n \xrightarrow{w} \mu$.

(ii) $\limsup_{n \rightarrow \infty} \mu_n(F) \leq \mu(F)$ for all closed sets F .

(iii) $\liminf_{n \rightarrow \infty} \mu_n(G) \geq \mu(G)$ for all open sets G .

(iv) $\lim_{n \rightarrow \infty} \mu_n(E) = \mu(E)$ for all sets E with $\mu(\partial E) = 0$.

Proof We start with (i) \Rightarrow (ii), the proof of which is similar to the one of Proposition 11.3. 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) \Leftrightarrow (iii) is almost trivial. Knowing this, the implication (iii) \Rightarrow (iv) is easy. The proof of (iv) \Rightarrow (i) roughly follows a pattern needed in Exercise 11.5. 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 $0 = x_0 < \dots < x_m = B$ be a finite set of points not in D such that $\max\{x_k - x_{k-1} : k = 1, \dots, m\} < \varepsilon$. Let $I_k = (x_{k-1}, x_k]$. The continuity of g implies that $\partial g^{-1}[I_k] \subset \{x_{k-1}, x_k\}$. Hence $\mu(\partial g^{-1}[I_k]) = 0$. Let $\tilde{g} = \sum_{k=1}^m x_k \mathbf{1}_{I_k}$. Then $|\mu_n(\tilde{g}) - \mu(\tilde{g})| \leq \sum_{k=1}^m x_k |\mu_n(I_k) - \mu(I_k)|$, which tends to zero as $n \rightarrow \infty$. 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 that

$$\limsup_{n \rightarrow \infty} |\mu_n(g) - \mu(g)| \leq 2\varepsilon.$$

This finishes the proof. \square

11.2 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, \dots be random variables and recall that $X_n \xrightarrow{w} X$ can be cast as

$$\mathbb{E} f(X_n) \rightarrow \mathbb{E} f(X), \forall f \in C_b(\mathbb{R}). \quad (11.4)$$

As a matter of fact one can show that weak convergence takes place, if (11.4) holds for all bounded *uniformly continuous* functions (Exercise 11.4). In the present section we take this as our characterization of weak convergence.

Our goal is to prove Theorem 11.16 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 11.13 *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*

$$|\mathbb{E} f(X) - \mathbb{E} f(X + Y)| \leq \varepsilon + 2\|f\| \mathbb{P}(|Y| \geq \delta). \quad (11.5)$$

Proof Let $\varepsilon > 0$ be given and choose $\delta > 0$ such that $|f(x) - f(y)| < \varepsilon$ whenever $|x - y| < \delta$. Then

$$\begin{aligned} |\mathbb{E} f(X) - \mathbb{E} f(X + Y)| &\leq \mathbb{E} (\mathbf{1}_{\{|Y| < \delta\}} |f(X) - f(X + Y)|) \\ &\quad + \mathbb{E} (\mathbf{1}_{\{|Y| \geq \delta\}} (|f(X)| + |f(X + Y)|)) \\ &\leq \varepsilon + 2\|f\| \mathbb{P}(|Y| \geq \delta). \end{aligned}$$

\square

Lemma 11.14 *Let Y, X, X_1, X_2, \dots 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 (11.5) it follows that

$$|\mathbb{E} f(X) - \mathbb{E} f(X + \sigma Y)| \leq \varepsilon + 2\|f\| \mathbb{P}(|Y| \geq \frac{\delta}{\sigma})$$

and

$$|\mathbb{E} f(X_n) - \mathbb{E} f(X_n + \sigma Y)| \leq \varepsilon + 2\|f\| \mathbb{P}(|Y| \geq \frac{\delta}{\sigma}).$$

Now we consider

$$\begin{aligned}
|\mathbb{E} f(X_n) - \mathbb{E} f(X)| &\leq |\mathbb{E} f(X_n) - \mathbb{E} f(X_n + \sigma Y)| \\
&\quad + |\mathbb{E} f(X_n + \sigma Y) - \mathbb{E} f(X + \sigma Y)| \\
&\quad + |\mathbb{E} f(X) - \mathbb{E} f(X + \sigma Y)| \\
&\leq 2\varepsilon + 4\|f\| \mathbb{P}(|Y| \geq \frac{\delta}{\sigma}) \\
&\quad + |\mathbb{E} f(X_n + \sigma Y) - \mathbb{E} f(X + \sigma Y)|.
\end{aligned}$$

By assumption, the last term tends to zero for $n \rightarrow \infty$. Letting then $\sigma \downarrow 0$, we obtain $\limsup_n |\mathbb{E} f(X_n) - \mathbb{E} f(X)| \leq 2\varepsilon$, which finishes the proof, since ε is arbitrary. \square

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

$$\mathbb{E}[f(Z)|X = x] = \int_{-\infty}^{\infty} f(z) \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{1}{2\sigma^2}(z-x)^2) dz =: f_\sigma(x).$$

Hence

$$\mathbb{E} f(X + \sigma Y) = \mathbb{E} f_\sigma(X). \quad (11.6)$$

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 11.3) that f has derivatives of all orders given by

$$f_\sigma^{(k)}(x) = \int_{-\infty}^{\infty} f(z) p_\sigma^{(k)}(z-x) dz. \quad (11.7)$$

Hence f_σ is a smooth function. Write \mathcal{C}^∞ 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 11.15 *Let X, X_1, X_2, \dots be random variables. The weak convergence $X_n \xrightarrow{w} X$ takes place iff $\mathbb{E} f(X_n) \rightarrow \mathbb{E} f(X)$, for all $f \in \mathcal{C}^\infty$.*

Proof Suppose that $\mathbb{E} f(X_n) \rightarrow \mathbb{E} f(X)$, for all $f \in \mathcal{C}^\infty$, then it holds in particular for any f_σ . In view of (11.6), this means that $X_n + \sigma Y \xrightarrow{w} X + \sigma Y$ for all $\sigma > 0$. Now Lemma 11.14 applies. \square

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 (11.11). Let $f \in \mathcal{C}^\infty$ and put

$$R(x, y) = f(x + y) - f(x) - yf'(x) - \frac{1}{2}y^2f''(x).$$

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

$$|\mathbb{E} f(X + Y) - \mathbb{E} f(X + W)| \leq \mathbb{E} |R(X, Y)| + \mathbb{E} |R(X, W)|. \quad (11.8)$$

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

$$R(x, y) = f(x + y) - f(x) - yf'(x) - \frac{1}{2}y^2f''(x) = \frac{1}{2}y^2(f''(\theta_2(x, y)) - f''(x)),$$

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 $\mathbb{E} |R(X, Y)|$:

$$\mathbb{E} |R(X, Y)| \leq C \mathbb{E} |Y|^3 \quad (11.9)$$

and

$$\mathbb{E} |R(X, Y)| \leq C(\varepsilon \mathbb{E} Y^2 + \mathbb{E} Y^2 \mathbf{1}_{\{|Y| > \varepsilon\}}). \quad (11.10)$$

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 η_1, \dots, η_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 $Z = X_1 + \eta_1$. Repetitive use of the triangle inequality and application of (11.8) gives

$$\begin{aligned} |\mathbb{E} f(S) - \mathbb{E} f(Z)| &\leq \sum_{j=1}^n |\mathbb{E} f(X_j + \xi_j) - \mathbb{E} f(X_j + \eta_j)| \\ &\leq \sum_{j=1}^n \mathbb{E} |R(X_j, \xi_j)| + \mathbb{E} |R(X_j, \eta_j)|. \end{aligned} \quad (11.11)$$

Theorem 11.16 (Central Limit Theorem) *Let for each $n \in \mathbb{N}$ be given a sequence $\xi_{n1}, \dots, \xi_{nk_n}$ of independent random variables with $\mathbb{E} \xi_{nj} = 0$ and $\sum_{j=1}^{k_n} \text{Var} \xi_{nj} = 1$. 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: $L_n(\varepsilon) \rightarrow 0$ as $n \rightarrow \infty$ for every $\varepsilon > 0$. Then $S_n := \sum_{j=1}^{k_n} \xi_{nj} \xrightarrow{w} Z$, where Z has a $N(0,1)$ distribution.

Proof Let $S_n = \sum_{j=1}^{k_n} \xi_{nj}$ and let η_{nj} ($j = 1, \dots, k_n$, $n \in \mathbb{N}$) 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} ($j = 1, \dots, k_n$) 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}^∞ . Recall Theorem 11.15. Take such $f \in \mathcal{C}^\infty$ and apply (11.11) to get

$$\begin{aligned} |\mathbb{E} f(S_n) - \mathbb{E} f(Z)| &= |\mathbb{E} f(S_n) - \mathbb{E} f(Z_n)| \\ &\leq \sum_{j=1}^{k_n} \mathbb{E} |R(X_{nj}, \xi_{nj})| + \mathbb{E} |R(X_{nj}, \eta_{nj})|, \end{aligned} \quad (11.12)$$

with an obvious meaning of the X_{nj} . For the first error terms in (11.12) we use the estimate of (11.10) which yields $\mathbb{E} \sum_{j=1}^{k_n} \mathbb{E} |R(X_{nj}, \xi_{nj})| \leq C(\varepsilon + L_n(\varepsilon))$. In view of the Lindeberg condition, this term can be made arbitrarily small. We now focus on the second error term in (11.12). Let $\sigma_{nj}^2 = \mathbb{E} \xi_{nj}^2 = \mathbb{E} \eta_{nj}^2$ and use (11.9) to obtain

$$\sum_{j=1}^{k_n} \mathbb{E} |R(X_{nj}, \eta_{nj})| \leq C \sum_{j=1}^{k_n} \mathbb{E} |\eta_{nj}|^3 = C \sum_{j=1}^{k_n} \sigma_{nj}^3 \mathbb{E} |N(0,1)|^3.$$

To finish the proof, we first observe that

$$\max_j \sigma_{nj}^2 = \max_j \mathbb{E} \xi_{nj}^2 = \max_j \mathbb{E} \xi_{nj}^2 (\mathbf{1}_{\{|\xi_{nj}| \leq \varepsilon\}} + \mathbf{1}_{\{|\xi_{nj}| > \varepsilon\}}) \leq \varepsilon^2 + L_n(\varepsilon).$$

Hence (use $\sum_{j=1}^{k_n} \sigma_{nj}^2 = 1$)

$$\sum_{j=1}^{k_n} \sigma_{nj}^3 \leq \max_j \sigma_{nj} \sum_{j=1}^{k_n} \sigma_{nj}^2 \leq (\varepsilon^2 + L_n(\varepsilon))^{1/2}.$$

And, again, this term can be made arbitrarily small, because of the Lindeberg condition. \square

11.3 Exercises

11.1 Prove the remaining assertions of Proposition 11.2.

11.2 Prove Proposition 11.11.

11.3 Show, using the Dominated Convergence Theorem, that (11.7) holds. Show also that all the derivatives are bounded functions.

11.4 Show that $X_n \xrightarrow{w} X$ iff $\mathbb{E}f(X_n) \rightarrow \mathbb{E}f(X)$ for all bounded uniformly continuous functions f . *Hint:* for one implication the proof of Proposition 11.3 is instructive.

11.5 Show the implication $F_n(x) \rightarrow F(x)$ for all $x \in C_F \Rightarrow \mu_n \xrightarrow{w} \mu$ of Corollary 11.5 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.*

11.6 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 $\mathbb{E}X_n \rightarrow \mathbb{E}X$.

11.7 Show the following variation on Fatou's lemma: if $X_n \xrightarrow{w} X$, then $\mathbb{E}|X| \leq \liminf_{n \rightarrow \infty} \mathbb{E}|X_n|$.

11.8 Show that the weak limit of a sequence of probability measures is unique.

11.9 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?

11.10 For each n we have a sequence $\xi_{n1}, \dots, \xi_{nk_n}$ of independent random variables with $\mathbb{E}\xi_{nj} = 0$ and $\sum_{j=1}^{k_n} \text{Var} \xi_{nj} = 1$. If $\sum_{j=1}^{k_n} \mathbb{E}|\xi_{nj}|^{2+\delta} \rightarrow 0$ as $n \rightarrow \infty$ for some $\delta > 0$, then $\sum_{j=1}^{k_n} \xi_{nj} \xrightarrow{w} N(0, 1)$. Show that this follows from the Lindeberg Central Limit Theorem.

11.11 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 $0 < \text{Var } X_j = \sigma^2 < \infty$. Show that this follows from the Lindeberg Central Limit Theorem.

12 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.

12.1 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 12.1 Let μ be a probability measure on $(\mathbb{R}, \mathcal{B})$. Its characteristic function $\phi : \mathbb{R} \rightarrow \mathbb{C}$ is defined by

$$\phi(u) = \int e^{iux} \mu(dx). \quad (12.1)$$

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 $f = g + ih$ into its real part g and imaginary part h , we define $\int f d\mu := \int g d\mu + i \int h d\mu$. For integrals of complex valued functions, previously shown theorem are, mutatis mutandis, true. For instance, one has $|\int f d\mu| \leq \int |f| d\mu$, where $|\cdot|$ denotes the norm of a complex number.

If X is a random variable with distribution μ , then it follows from Proposition 4.24 applied to $h(x) = \exp(iux)$, that ϕ_μ can alternatively be expressed by $\phi(u) = \mathbb{E} \exp(iuX)$. 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

$$\phi_X(u) = \int e^{iux} f(x) dx. \quad (12.2)$$

Analysts define for $f \in \mathcal{L}^1(\mathbb{R}, \mathcal{B}, \lambda)$ (λ Lebesgue measure) the Fourier transform \hat{f} by

$$\hat{f}(u) = \int e^{-iux} f(x) dx,$$

sometimes also by dividing this expression by $\sqrt{2\pi}$. What we thus see, is the equality $\phi_X(u) = \hat{f}(-u)$.

Proposition 12.2 *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) $\phi_{aX+b}(u) = \phi_X(au)e^{iub}$.
- (iv) ϕ is real valued and symmetric around zero, if X and $-X$ have the same distribution.
- (v) If X and Y are independent random variables, 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 $\phi^{(k)}(0) = i^k \mathbb{E}X^k$.

Proof Properties (i), (iii) and (iv) are trivial. Consider (ii). Fixing $u \in \mathbb{R}$, we consider $\phi(u+t) - \phi(u)$ for $t = t_n \rightarrow 0$. We have $|\phi(u+t_n) - \phi(u)| = |\mathbb{E} \exp(iuX)(\exp(it_n X) - 1)| \leq \mathbb{E} |\exp(it_n X) - 1|$. The random variables $Y_n = |\exp(it_n X) - 1|$ converge to zero 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 4.30. Finally, property (vi) for $k = 1$ follows by an application of the Dominated Convergence Theorem and the inequality $|e^{ix} - 1| \leq |x|$, for $x \in \mathbb{R}$. The other cases can be treated similarly. \square

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. The following theorem is similar to Fourier inversion. Note that the integration interval in (12.3) is symmetric around zero. This is essential.

Theorem 12.3 *Let F be a distribution function and ϕ its characteristic function. Then, for all $a < b$*

$$\lim_{T \rightarrow \infty} \frac{1}{2\pi} \int_{-T}^T \frac{e^{-iua} - e^{-iub}}{iu} \phi(u) du = \tilde{F}(b) - \tilde{F}(a). \quad (12.3)$$

Proof Let $a < b$. We compute, using Fubini's theorem which we will justify

below,

$$\Phi_T = \frac{1}{2\pi} \int_{-T}^T \frac{e^{-iua} - e^{-iub}}{iu} \phi(u) du \quad (12.4)$$

$$\begin{aligned} &= \frac{1}{2\pi} \int_{-T}^T \frac{e^{-iua} - e^{-iub}}{iu} \int e^{iux} \mu(dx) du \\ &= \frac{1}{2\pi} \int \int_{-T}^T \frac{e^{-iua} - e^{-iub}}{iu} e^{iux} du \mu(dx) \\ &= \frac{1}{2\pi} \int \int_{-T}^T \frac{e^{i(x-a)u} - e^{i(x-b)u}}{iu} du \mu(dx) \quad (12.5) \\ &=: \int E_T(x) \mu(dx) \end{aligned}$$

Application of Fubini's theorem is justified as follows. First, the integrand in (12.5) is bounded by $b - a$, because $|e^{ix} - e^{iy}| \leq |x - y|$ 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

$$\begin{aligned} E_T(x) &= \frac{1}{2\pi} \int_{-T}^T \frac{\sin(x-a)u - \sin(x-b)u}{u} du \\ &= \frac{1}{2\pi} \int_{-T}^T \frac{\sin(x-a)u}{u} du - \frac{1}{2\pi} \int_{-T}^T \frac{\sin(x-b)u}{u} du \\ &= \frac{1}{2\pi} \int_{-T(x-a)}^{T(x-a)} \frac{\sin v}{v} dv - \frac{1}{2\pi} \int_{-T(x-b)}^{T(x-b)} \frac{\sin v}{v} dv. \end{aligned}$$

The function g given by $g(s, t) = \int_s^t \frac{\sin y}{y} dy$ is continuous in (s, t) . Hence it is bounded on any compact subset of \mathbb{R}^2 . Moreover, $g(s, t) \rightarrow \pi$ as $s \rightarrow -\infty$ and $t \rightarrow \infty$. 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 (12.5), since μ is a finite measure. The second ingredient is $E(x) := \lim_{T \rightarrow \infty} E_T(x)$. From Exercise 5.10 we deduce that

$$\int_0^\infty \frac{\sin \alpha x}{x} dx = \operatorname{sgn}(\alpha) \frac{\pi}{2}.$$

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

$$\Phi_T \rightarrow \mu(a, b) + \frac{1}{2} \mu(\{a, b\}) = \tilde{F}(b) - \tilde{F}(a).$$

□

Corollary 12.4 *If μ and ν are two probability measures on $(\mathbb{R}, \mathcal{B})$ whose characteristic functions are the same, then they coincide.*

Proof Exercise 12.1. □

The content of Corollary 12.4 explains why characteristic functions are called characteristic.

Theorem 12.5 *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

$$f(x) = \frac{1}{2\pi} \int e^{-iux} \phi(u) \, du. \quad (12.6)$$

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 (12.3) is equal to the (Lebesgue) integral $\int \frac{e^{-iua} - e^{-iub}}{iu} \phi(u) \, du$, again because of dominated convergence. We use Fubini's theorem to compute for $a < b$

$$\begin{aligned} \int_a^b f(x) \, dx &= \frac{1}{2\pi} \int_a^b \int e^{-iux} \phi(u) \, du \, dx \\ &= \frac{1}{2\pi} \int \phi(u) \int_a^b e^{-iux} \, dx \, du \\ &= \frac{1}{2\pi} \int \phi(u) \frac{e^{-iua} - e^{-iub}}{iu} \, du \\ &= F(b) - F(a), \end{aligned}$$

for every a and b , because of Theorem 12.3 and because $\int_a^b f(x) \, dx$ is continuous in a and b . It also follows that f must be nonnegative and so it is a density. □

Remark 12.6 Note the duality between the expressions (12.2) and (12.6). Apart from the presence of the minus sign in the integral and the factor 2π in the denominator in (12.6), 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 (12.1). We have with $u, x \in \mathbb{R}^k$, $\langle \cdot, \cdot \rangle$ the standard inner product,

$$\phi(u) = \int e^{i\langle u, x \rangle} \mu(dx).$$

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 12.2. As a consequence we have the following characterization of independent random variables.

Proposition 12.7 *Let $X = (X_1, \dots, X_k)$ be a k -dimensional random vector. Then X_1, \dots, X_k are independent random variables iff $\phi_X(u) = \prod_{i=1}^k \phi_{X_i}(u_i)$, $\forall u = (u_1, \dots, u_k) \in \mathbb{R}^k$.*

Proof If the X_i are independent, the statement about the characteristic functions is proved in the same way as Proposition 12.2 (v). If the characteristic function ϕ_X factorizes as stated, the result follows by the uniqueness property of characteristic functions. \square

12.2 Characteristic functions and weak convergence

The first result says that weak convergence of probability measures implies pointwise convergence of their characteristic functions.

Proposition 12.8 *Let μ, μ_1, μ_2, \dots be probability measures on $(\mathbb{R}, \mathcal{B})$ and let $\phi, \phi_1, \phi_2, \dots$ be their characteristic functions. If $\mu_n \xrightarrow{w} \mu$, then $\phi_n(u) \rightarrow \phi(u)$ for every $u \in \mathbb{R}$.*

Proof Consider for fixed u the function $f(x) = e^{iux}$. It is obviously bounded and continuous and we obtain straight from Definition 11.1 that $\mu_n(f) \rightarrow \mu(f)$. But $\mu_n(f) = \phi_n(u)$. \square

Proposition 12.9 *Let μ_1, μ_2 be probability measures on $(\mathbb{R}, \mathcal{B})$. Let ϕ_1, ϕ_2, \dots be the corresponding characteristic functions. Assume that the sequence (μ_n) is tight and that for all $u \in \mathbb{R}$ the limit $\phi(u) := \lim_{n \rightarrow \infty} \phi_n(u)$ 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 11.10 to deduce that there exists a weakly converging subsequence (μ_{n_k}) with a probability measure as limit. Call this limit μ . From Proposition 12.8 we know that $\phi_{n_k}(u) \rightarrow \phi_\mu(u)$ 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

$$|\mu_{n'_k}(f) - \mu(f)| > \varepsilon. \quad (12.7)$$

Using Proposition 11.10, 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 $\mu_{n''_k}(f) \rightarrow \mu(f)$, which contradicts (12.7). \square

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 $\int_{-a}^a (1 - \phi(u)) du \in \mathbb{R}$ for every $a > 0$.

Lemma 12.10 *Let a random variable X have distribution μ and characteristic function ϕ . Then for every $K > 0$*

$$P(|X| > 2K) \leq K \int_{-1/K}^{1/K} (1 - \phi(u)) du. \quad (12.8)$$

Proof It follows from Fubini's theorem and $\int_{-a}^a e^{iux} du = 2 \frac{\sin ax}{x}$ that

$$\begin{aligned} K \int_{-1/K}^{1/K} (1 - \phi(u)) du &= K \int_{-1/K}^{1/K} \int (1 - e^{iux}) \mu(dx) du \\ &= \int K \int_{-1/K}^{1/K} (1 - e^{iux}) du \mu(dx) \\ &= 2 \int (1 - \frac{\sin x/K}{x/K}) \mu(dx) \\ &\geq 2 \int_{|x/K| > 2} (1 - \frac{\sin x/K}{x/K}) \mu(dx) \\ &\geq \mu([-2K, 2K]^c). \end{aligned}$$

since $\frac{\sin x}{x} \leq \frac{1}{2}$ for $x > 2$. □

The following theorem is known as *Lévy's continuity theorem*.

Theorem 12.11 *Let μ_1, μ_2 be probability measures on $(\mathbb{R}, \mathcal{B})$ and ϕ_1, ϕ_2, \dots be the corresponding characteristic functions. Assume that for all $u \in \mathbb{R}$ the limit $\phi(u) := \lim_{n \rightarrow \infty} \phi_n(u)$ 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 12.10. Let $\varepsilon > 0$. Since ϕ is continuous at zero, the same holds for $\bar{\phi}$, 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 $0 \leq \int_{-\delta}^{\delta} (2 - \phi(u) - \phi(-u)) du < 2\delta\varepsilon$.

By the convergence of the characteristic functions (which are bounded), the Dominated Convergence Theorem implies that for big enough n , $n \geq N$ say,

$$\int_{-\delta}^{\delta} (2 - \phi_n(u) - \phi_n(-u)) du < \int_{-\delta}^{\delta} (2 - \phi(u) - \phi(-u)) du + 2\delta\varepsilon.$$

Hence, by taking $n \geq N$ we have

$$\int_{-\delta}^{\delta} (2 - \phi_n(u) - \phi_n(-u)) du < 4\delta\varepsilon.$$

Since $\int_{-\delta}^{\delta} (\phi_n(u) + \phi_n(-u)) du = 2 \int_{-\delta}^{\delta} \phi_n(u) du$, it follows from Lemma 12.10 that for $n \geq N$ and $K = 1/\delta$

$$\begin{aligned} P(|X_n| > 2K) &\leq \frac{1}{\delta} \int_{-\delta}^{\delta} (1 - \phi_n(u)) du \\ &\leq \frac{1}{2\delta} \int_{-\delta}^{\delta} (2 - \phi_n(u) - \phi_n(-u)) du \\ &< 2\varepsilon. \end{aligned}$$

We conclude that $(\mu_n)_{n \geq N}$ is tight. Apply now Proposition 12.8. \square

Corollary 12.12 *Let μ, μ_1, μ_2 be probability measures on $(\mathbb{R}, \mathcal{B})$ and $\phi, \phi_1, \phi_2, \dots$ be their corresponding characteristic functions. Then $\mu_n \xrightarrow{w} \mu$ if and only if $\phi_n(u) \rightarrow \phi(u)$ for all $u \in \mathbb{R}$.*

Proof If $\phi_n(u) \rightarrow \phi(u)$ for all $u \in \mathbb{R}$, then we can apply Theorem 12.11. 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 12.8. \square

12.3 The Central Limit Theorem revisited

The proof of the Theorem 11.16 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 12.13 *Let X be a random variable with $\mathbb{E} X^2 < \infty$ and with characteristic function ϕ . Then*

$$\begin{aligned} |\phi(u) - 1| &\leq \mathbb{E} \min\{2, |uX|\}, \\ |\phi(u) - 1 - iu\mathbb{E} X| &\leq \mathbb{E} \min\{2|u||X|, \frac{1}{2}u^2 X^2\} \end{aligned}$$

and

$$|\phi(u) - 1 - iu\mathbb{E} X + \frac{1}{2}u^2\mathbb{E} X^2| \leq \mathbb{E} \min\{u^2 X^2, \frac{1}{6}|u|^3 |X|^3\}.$$

Proof Let $x \in \mathbb{R}$. Then $|e^{ix} - 1| \leq 2$ and $|e^{ix} - 1| = |\int_0^x ie^{iy} dy| \leq |x|$. Hence $|e^{ix} - 1| \leq \min\{2, |x|\}$. Since

$$e^{ix} - 1 - ix = \int_0^x (e^{iy} - 1) dy,$$

and

$$e^{ix} - 1 - ix + \frac{1}{2}x^2 = - \int_0^x \int_0^y (e^{it} - 1) dt dy,$$

we arrive at $|e^{ix} - 1 - ix| \leq \min\{2|x|, \frac{1}{2}x^2\}$ and $|e^{ix} - 1 - ix + \frac{1}{2}x^2| \leq \min\{x^2, |x|^3/6\}$. Replacing x with uX and taking expectations yields the assertions. \square

We are now ready to give the announced alternative proof of theorem 11.16.

Proof of Theorem 11.16 Let $\phi_{n_j}(u) = \mathbb{E} \exp(iu\xi_{n_j})$ and $\phi_n(u) = \mathbb{E} \exp(iuS_n)$. Because of independence we have

$$\phi_n(u) = \prod_{j=1}^{k_n} \phi_{n_j}(u).$$

First we show that

$$\sum_{j=1}^{k_n} (\phi_{n_j}(u) - 1) \rightarrow -\frac{1}{2}u^2. \quad (12.9)$$

We write

$$\sum_{j=1}^{k_n} (\phi_{n_j}(u) - 1) = \sum_{j=1}^{k_n} (\phi_{n_j}(u) - 1 + \frac{1}{2}u^2 \mathbb{E} \xi_{n_j}^2) - \sum_{j=1}^{k_n} \frac{1}{2}u^2 \mathbb{E} \xi_{n_j}^2.$$

The last term gives the desired limit, so it suffices to show that the first term converges to zero. By virtue of lemma 12.13, we can bound its absolute value by

$$\sum_{j=1}^{k_n} \mathbb{E} \min\{u^2 \xi_{n_j}^2, \frac{1}{6}|u|^3 |\xi_{n_j}|^3\}. \quad (12.10)$$

But

$$\mathbb{E} \min\{u^2 \xi_{n_j}^2, \frac{1}{6}|u|^3 |\xi_{n_j}|^3\} \leq \frac{1}{6}|u|^3 \varepsilon \mathbb{E} \xi_{n_j}^2 \mathbf{1}_{\{|\xi_{n_j}| \leq \varepsilon\}} + u^2 \mathbb{E} \xi_{n_j}^2 \mathbf{1}_{\{|\xi_{n_j}| > \varepsilon\}}.$$

Hence we get that the expression in (12.10) is majorized by

$$\frac{1}{6}|u|^3 \varepsilon \sum_{j=1}^{k_n} \mathbb{E} \xi_{n_j}^2 + u^2 L_n(\varepsilon) = \frac{1}{6}|u|^3 \varepsilon + u^2 L_n(\varepsilon),$$

which tends to $\frac{1}{6}|u|^3 \varepsilon$. Since ε is arbitrary, we have proved (12.9). It then also follows that

$$\exp\left(\sum_{j=1}^{k_n} (\phi_{n_j}(u) - 1)\right) \rightarrow \exp\left(-\frac{1}{2}u^2\right). \quad (12.11)$$

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 (12.11), we are finished as soon as we have shown that

$$\prod_{j=1}^{k_n} \phi_{n_j}(u) - \exp\left(\sum_{j=1}^{k_n} (\phi_{n_j}(u) - 1)\right) \rightarrow 0. \quad (12.12)$$

The displayed difference is in absolute value less than

$$\sum_{j=1}^{k_n} |\phi_{nj}(u) - \exp(\phi_{nj}(u) - 1)|, \quad (12.13)$$

because of the following elementary result: if a_i and b_i are complex numbers with norm less than or equal to one, then

$$\left| \prod_{i=1}^n a_i - \prod_{i=1}^n b_i \right| \leq \sum_{i=1}^n |a_i - b_i|.$$

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 12.3).

Let $M_n(u) = \max_j |\phi_{nj}(u) - 1|$. Now we use the inequality $|e^z - 1 - z| \leq |z|^2 e^{|z|}$ (which easily follows from a Taylor expansion) with $z = \phi_{nj}(u) - 1$ to bound (12.13) 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 12.13, 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

$$\max_j \mathbb{E} \xi_{nj}^2 \rightarrow 0 \quad (12.14)$$

and then by Lemma 12.13 and Jensen's inequality

$$M_n(u) = \max_j |\phi_{nj}(u) - 1| \leq \max_j |u| \mathbb{E} |\xi_{nj}| \leq |u| (\max_j \mathbb{E} \xi_{nj}^2)^{1/2} \rightarrow 0.$$

This proves (12.12) and hence it completes the proof of the theorem. \square

Remark 12.14 The Lindeberg condition in the theorem is almost necessary. One can show that if (12.14) holds and if the weak convergence as in the theorem takes place, then also the Lindeberg condition is satisfied.

12.4 Exercises

12.1 Prove Corollary 12.4

12.2 Let μ and ν be probability measures on $(\mathbb{R}, \mathcal{B})$ with corresponding characteristic functions ϕ and ψ .

(a) Show that $\int_{\mathbb{R}} \exp(-iuy) \phi(y) \nu(dy) = \int_{\mathbb{R}} \psi(x - u) \mu(dx)$.

- (b) Assume that ν is the law of a random variable Y with a $N(0, \frac{1}{\sigma^2})$ distribution, then $\psi(u) = \exp(-\frac{1}{2}u^2/\sigma^2)$. Let f_{σ^2} be the density of the $N(0, \sigma^2)$ distribution. Show that

$$\frac{1}{2\pi} \int_{\mathbb{R}} \exp(-iuy)\phi(y) \exp(-\frac{1}{2}\sigma^2 y^2) dy = \int_{\mathbb{R}} f_{\sigma^2}(u-x) \mu(dx),$$

and show that the right hand side gives the *density* of $X + Y$ (see Exercise 5.6).

- (c) Write $Y = \sigma Z$, with Z having the standard normal distribution. It follows that ϕ determines the distribution of $X + \sigma Z$ for all $\sigma \in \mathbb{R}$. Show that ϕ uniquely determines μ . (This gives an alternative proof of the assertion of Corollary 12.4).

12.3 Let X_1, X_2, \dots 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 $\psi = \exp(\lambda\phi - \lambda)$.

12.4 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) $\phi_{N(\mu, \sigma^2)}(u) = \exp(iu\mu - \frac{1}{2}\sigma^2 u^2)$.
- (c) If X has an exponential distribution with parameter λ , then $\phi_X(u) = \lambda/(\lambda - iu)$.
- (d) If X has a Cauchy distribution, then $\phi_X(u) = \exp(-|u|)$. Show also that $\frac{1}{n} \sum_{k=1}^n X_k$ has a Cauchy distribution, if the X_k are *iid* and Cauchy distributed.

12.5 Let X_n have a Bin($n, \lambda/n$) distribution (for $n > \lambda$). Show that $X_n \xrightarrow{w} X$, where X has a Poisson(λ) distribution.

12.6 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 : $\phi(u) = \mathbb{E} \exp(iuX)$.

- (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 $p(z) = \frac{1}{2\pi\sigma} \int \phi(-y/\sigma) \exp(iyz/\sigma - \frac{1}{2}y^2) dy$.

12.7 Let X, X_1, X_2, \dots 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 $\phi_n \rightarrow \phi$ pointwise, then $p_n \rightarrow p$ pointwise. Invoke Exercise 12.6 and the dominated convergence theorem to show this.

- (b) Let $f \in C_b(\mathbb{R})$ 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 $|\mathbb{E}f(X_n + \sigma Y) - \mathbb{E}f(X + \sigma Y)| \rightarrow 0$ if $\phi_n \rightarrow \phi$ pointwise.
- (d) Prove Corollary 12.12: $X_n \xrightarrow{w} X$ iff $\phi_n \rightarrow \phi$ pointwise.

12.8 Let X_1, X_2, \dots, 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 $X_{(1)} = \min\{X_1, \dots, X_n\}$, $X_{(2)} = \min\{X_1, \dots, X_n\} \setminus \{X_{(1)}\}$, etc. The resulting $X_{(1)}, X_{(2)}, \dots, X_{(n)}$ is called the ordered sample. The *sample median* M_n of X_1, \dots, X_n is by definition $X_{(k)}$. Show that with $U_{nj} = \mathbf{1}_{\{X_j \leq m + n^{-1/2}x\}}$ we have

$$\mathbb{P}(n^{1/2}(M_n - m) \leq x) = \mathbb{P}\left(\sum_j U_{nj} \geq k\right).$$

- (c) Let $p_n = \mathbb{P}U_{nj}$, $b_n = (np_n(1-p_n))^{1/2}$, $\xi_{nj} = (U_{nj} - p_n)/b_n$, $Z_n = \sum_{j=1}^n \xi_{nj}$, $t_n = (k - np_n)/b_n$. Rewrite the probabilities in part (b) as $\mathbb{P}(Z_n \geq t_n)$ and show that $t_n \rightarrow t := -2xf(m)$.
- (d) Show that $\mathbb{P}(Z_n \geq t) \rightarrow 1 - \Phi(t)$, where Φ is the standard normal distribution.
- (e) Show that $\mathbb{P}(Z_n \geq t_n) \rightarrow \Phi(2f(m)x)$ and conclude that the *Central Limit Theorem for the sample median* holds:

$$2f(m)n^{1/2}(M_n - m) \xrightarrow{w} N(0, 1).$$

12.9 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 $\Gamma(t) = \int_0^\infty y^{t-1}e^{-y} dy$ 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 $\phi_t(u) = \mathbb{E}e^{iuX_t}$ of X_t is given by

$$\phi_t(u) = e^{-iu\sqrt{t}} \frac{1}{\left(1 - \frac{iu}{\sqrt{t}}\right)^t}$$

and conclude that $\phi_t(u) \rightarrow e^{-\frac{1}{2}u^2}$ as $t \rightarrow \infty$.

(c) Show that

$$\frac{t^{t-\frac{1}{2}} e^{-t}}{\Gamma(t)} = \frac{1}{2\pi} \int_{-\infty}^{\infty} \phi_t(u) \, du.$$

(d) Prove Stirling's formula

$$\lim_{t \rightarrow \infty} \frac{\Gamma(t)}{\sqrt{2\pi} e^{-t} t^{t-\frac{1}{2}}} = 1.$$

12.10 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 $\mathbb{E}[e^{iuX} | \mathcal{G}]$. 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 ($\mathbb{E} |\hat{\phi}(u+h) - \hat{\phi}(u)| \rightarrow 0$ for $h \rightarrow 0$).
- (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 $\phi : \mathbb{R} \rightarrow \mathcal{C}$ such that $\phi(u)$ is a version of $\mathbb{E}[e^{iuX} | \mathcal{G}]$ 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 Brownian motion

This chapter is devoted to showing the existence of Brownian motion as a well defined mathematical well defined object. 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.

13.1 The space $C[0, \infty)$

In this section we summarize some facts concerning the space $C[0, \infty)$ of real valued continuous functions defined on $[0, \infty)$. For $x_1, x_2 \in C[0, \infty)$ we define

$$\rho(x_1, x_2) = \sum_{n \geq 1} 2^{-n} (\max\{|x_1(t) - x_2(t)| : 0 \leq t \leq n\} \wedge 1). \quad (13.1)$$

Then ρ defines a metric on $C[0, \infty)$ (which we use throughout these notes) and we have

Proposition 13.1 *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

$$m^T(x, \delta) = \max\{|x(t) - x(s)| : s, t \in [0, T], |s - t| \leq \delta\}. \quad (13.2)$$

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 13.2 *A set A in $C[0, \infty)$ is relatively compact (has compact closure) iff (i) $\sup\{|x(0)| : x \in A\} < \infty$ and (ii) for all $T > 0$ $\lim_{\delta \downarrow 0} \sup\{m^T(x, \delta) : x \in A\} = 0$.*

Under requirement (ii) in this proposition, the functions in A are uniformly equicontinuous.

Cylinder sets of $C[0, \infty)$ have the typical form $\{x : (x(t_1), \dots, x(t_k)) \in A\}$, where $A \in \mathcal{B}(\mathbb{R}^k)$ for some $k \geq 1$ and $t_1, \dots, t_k \in [0, \infty)$. A finite dimensional projection on $(C[0, \infty), \rho)$ is by definition of the following type: $\pi_{t_1, \dots, t_k}(x) = (x(t_1), \dots, x(t_k))$, 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 (Ω, \mathcal{F}) be a measurable space. A map $X : \Omega \rightarrow C[0, \infty)$ is called a

random element of $C[0, \infty)$ if it is \mathcal{F}/\mathcal{B} -measurable. It follows that $\pi_{t_1, \dots, t_k} \circ X$ is random vector in \mathbb{R}^k , for any finite dimensional projection π_{t_1, \dots, t_k} , and it is usually denoted by $(X_{t_1}, \dots, X_{t_k})$. 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 (Ω, \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 $(X_{t_1}, \dots, X_{t_k})$ on \mathbb{R}^k ($k \geq 1, t_i \in [0, \infty)$).

13.2 Weak convergence on metric spaces

Let (S, ρ) be a metric space and P, P^1, P^2, \dots be probability measures on the Borel σ -algebra $\mathcal{B}(S)$. Like in the real case we say that P^n converges weakly to P (notation $P^n \xrightarrow{w} P$) iff for all $f \in C_b(S)$ one has $\lim P^n f = P f$. If X, X^1, X^2, \dots are random variables defined on probability spaces (Ω, \mathcal{F}, 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 P^n of X^n converge weakly to the law P of X , equivalently, iff $P^n f(X^n) \rightarrow P f(X)$ for all $f \in C_b(S)$.

We need a generalization of tightness as given in Definition 11.8 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 $\inf\{P(K) : P \in \Pi\} > 1 - \varepsilon$. 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 (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 13.3 *A family Π of probability measures on a complete separable metric space is tight iff it is relatively compact.*

We will also need the following perturbation result.

Proposition 13.4 *Let X^1, X^2, \dots and Y^1, Y^2, \dots be random sequences in a metric space (S, ρ) and defined on a single probability space. If $X^n \xrightarrow{w} X$ and $\rho(Y^n, X^n) \xrightarrow{P} 0$, 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 13.5 *Let P^1, P^2, \dots be a sequence of probability measures on the space $(C[0, \infty), \mathcal{B})$. This sequence is tight iff*

$$\lim_{\lambda \uparrow \infty} \sup\{P^n(x : |x(0)| > \lambda) : n \geq 1\} = 0 \quad (13.3)$$

and

$$\limsup_{\delta \downarrow 0} \{P^n(x : m^T(x, \delta) > \varepsilon) : n \geq 1\} = 0, \forall T, \varepsilon > 0. \quad (13.4)$$

Proof If the sequence is tight, the result is a straightforward application of Theorem 13.2. For every $\varepsilon > 0$ we can find a compact K such that $\inf_n P^n(K) > 1 - \varepsilon$. 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 (13.3) and (13.4) and let $\varepsilon, T > 0$, T integer, be given. Choose λ_T such that $\sup_n P^n(x : |x(0)| > \lambda_T) \leq \varepsilon 2^{-T-1}$. For each $k \geq 1$ we can also find δ_k such that $\sup_n P^n(x : m^T(x, \delta_k) > 1/k) \leq \varepsilon 2^{-T-k-1}$. 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 13.2 we obtain that K has compact closure and it is thus compact itself. Finally we compute $P^n(K^c) \leq \sum_{T \geq 1} \mathbb{P}(A_{T,0}^c) + \sum_{k \geq 1} \mathbb{P}(A_{T,k}^c) \leq \varepsilon$. \square

We have seen that any finite dimensional projection is continuous. Hence, if X, X_1, X_2, \dots are random elements of $(C[0, \infty), \mathcal{B})$ and if we assume that $X_n \xrightarrow{w} X$, then also $(X_{t_1}^n, \dots, X_{t_k}^n)$ considered as random elements in \mathbb{R}^k converge in distribution to $(X_{t_1}, \dots, X_{t_k})$. This is then true for any finite set of t_i 's and we say that all finite dimensional distributions converge weakly. The converse does not hold in general, unless one assumes tightness.

Theorem 13.6 *Let X_1, X_2, \dots be random elements of $C[0, \infty)$. Assume that their collection $\{P^1, P^2, \dots\}$ of distributions is tight and that all finite dimensional distributions of the X_n converge weakly. Then there exists a probability measure P on $(C[0, \infty), \mathcal{B})$ such that $P^n \xrightarrow{w} P$.*

Proof Every subsequence of (P^n) is tight as well and thus has a convergent subsequence. Different subsequences have to converge to the same limit, call it P , since the finite dimensional distributions corresponding to these sequences converge. Hence, if (P^n) has a limit, it must be P . Suppose therefore that the P^n don't converge. Then there is bounded and continuous f and an $\varepsilon > 0$ such that $|P^{n_k} f - P f| > \varepsilon$ along a subsequence (P^{n_k}) . No further subsequence of this can have P as a limit which contradicts what we just showed. \square

13.3 An invariance principle

Throughout this section we work with a real valued *iid* sequence ξ_1, ξ_2, \dots 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 $t \geq 0$

$$X_t^n = \frac{1}{\sigma\sqrt{n}}(S_{[nt]} + (nt - [nt])\xi_{[nt]+1}). \quad (13.5)$$

The processes X^n have continuous paths and can be considered as random elements of $C[0, \infty)$. Notice that the increments $X_t^n - X_s^n$ of each X^n over

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 have $\text{Var}(X_t^n - X_s^n) = t - s$, the central limit theorem should be helpful to understand the limit behavior.

Theorem 13.7 *Let $0 = t_0 < t_1 < \dots < t_k$. Then the k -vector of increments $X_{t_j}^n - X_{t_{j-1}}^n$ with $j = 1, \dots, k$ converges in distribution to a random vector with independent elements N_j , where each N_j has a $N(0, t_j - t_{j-1})$ distribution.*

Proof Since the term in (13.5) with the $\xi_{[nt]}$ tends to zero in probability, we can ignore it as a consequence of proposition 13.4. But then the conclusion follows from the ordinary Central Limit Theorem. \square

Denote by P^n the law of X^n . We have the following important result, whose proof is deferred to the next section.

Theorem 13.8 *The sequence of probability measures P^n is tight.*

Combined with Theorems 13.6 and 13.7 one obtains

Theorem 13.9 *There exists a probability measure P^* on $(C[0, \infty), \mathcal{B})$ such that $P^n \xrightarrow{w} P^*$.*

Any process defined on some probability space that has continuous paths, that starts in zero and that has independent increments over non-overlapping intervals (s, t) with a $N(0, t - s)$ distribution is called a Wiener process, also called a Brownian motion. Let W denotes the *coordinate mapping process* on $\Omega = C[0, \infty)$, i.e. W is defined by $W_t(\omega) = \omega(t)$ for all $t \geq 0$. Under the measure P^* 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 P^* is called Wiener measure. Notice that P^* is also the law of W .

We can rephrase Theorem 13.9 as

Theorem 13.10 *The processes X^n of this section converge in distribution to a Wiener process W .*

Both Theorems 13.9 and 13.10 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.

13.4 The proof of Theorem 13.8

Consider the process S_n of section 13.3. To prove Theorem 13.8 we use the following results.

Lemma 13.11 *Let $\gamma > 0$, $n \geq 1$, $N \geq n$ and $\eta \geq \sigma\sqrt{2(n-1)}$. The following inequalities are valid.*

$$\mathbb{P}(\max_{j \leq n} |S_j| > \gamma) \leq 2\mathbb{P}(|S_n| > \gamma - \eta) \quad (13.6)$$

$$\mathbb{P}(\max_{\substack{1 \leq j \leq n \\ 0 \leq k \leq N}} |S_{j+k} - S_k| > \gamma) \leq \left(\frac{N}{n} + 2\right)\mathbb{P}(\max_{j \leq n} |S_j| > \gamma/3). \quad (13.7)$$

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

$$\mathbb{P}(\tau \leq n, |S_n| > \gamma - \eta) + \mathbb{P}(\tau \leq n, |S_n| \leq \gamma - \eta) \quad (13.8)$$

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 Chebyshev'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 (13.8), 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 (13.6) follows.

To prove (13.7) we argue as follows. Let $m = \lfloor \frac{N}{n} \rfloor$ and consider the 'intervals' $\{pn, \dots, (p+1)n-1\}$, for $p = 0, \dots, m$. 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. \square

With this lemma we prove 13.8 as follows. According to Theorem 13.5 it is sufficient to show that

$$\limsup_{\delta \downarrow 0} \mathbb{P}(\max_{n \geq 1} \sup_{\substack{|s-t| \leq \delta \\ 0 \leq t, s \leq T}} |X_t^n - X_s^n| > \varepsilon) = 0 \text{ for all } T, \varepsilon > 0. \quad (13.9)$$

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 (13.9) is less than

$$\mathbb{P}\left(\max_{\substack{|s-t| \leq [n\delta]+1 \\ 0 \leq t, s \leq [nT]+1}} |Y_t - Y_s| > \varepsilon\sigma\sqrt{n}\right).$$

But, since Y is piecewise linear between the integer values of its arguments, the max is attained at integer numbers. Hence we consider

$$\mathbb{P}\left(\max_{\substack{0 \leq j \leq [n\delta]+1 \\ 0 \leq k \leq [nT]+1}} |S_{j+k} - S_k| > \varepsilon\sigma\sqrt{n}\right).$$

Now we apply inequality (13.7) and bound this probability by

$$\left(\frac{[nT] + 1}{[n\delta] + 1} + 2\right)\mathbb{P}\left(\max_{j \leq [n\delta] + 1} |S_j| > \varepsilon\sigma\sqrt{n}/3\right). \quad (13.10)$$

In view of (13.6) (take $\eta = \sigma\sqrt{2[n\delta]}$) the probability in (13.10) 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: $\frac{1}{\sigma\sqrt{[n\delta]}}S_{[n\delta] + 1} \xrightarrow{w} Z$, where Z has a $N(0, 1)$ distribution. So for $n \rightarrow \infty$ 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 limsup in (13.10) for $n \rightarrow \infty$ is less than $\frac{T\delta + 2\delta^2}{(\varepsilon/3 - \sqrt{2\delta})^4} \mathbb{E}Z^4$, from which we obtain (13.9). \square

.

13.5 Exercises

13.1 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 randomness here). Show that all finite dimensional distributions of the X^n converge, but X^n does not converge in distribution.

13.2 Show that ρ as in (1.1) defines a metric.

13.3 Suppose that the ξ_i of section 4 of the lecture notes are *iid* normally distributed random variables. Use Doob’s inequality to obtain $\mathbb{P}(\max_{j \leq n} |S_j| > \gamma) \leq 3\gamma^{-4}n^2$.

13.4 Show that a finite dimensional projection on $C[0, \infty)$ (with the metric ρ) is continuous.

13.5 Consider $C[0, \infty)$ with the Borel σ -algebra \mathcal{B} induced by ρ and some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. If $X : (\Omega, \mathcal{F}) \rightarrow (C[0, \infty), \mathcal{B})$ 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 5.7.

13.6 Prove Proposition 13.4.