Measure Theoretic Probability (condensed lecture notes)

Tinbergen Institute

P.J.C. Spreij

this version: September 21, 2021

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.

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

The present set of lecture notes, composed for the course on *Measure theory and asymptotic statistics* at the Tinbergen Institute is a short version of the extended set Measure theoretic probability.

Amsterdam, October 2018
Peter Spreij

Contents

1 σ-algebras and measures

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

1.1 σ-algebras

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

(i) $S \in \Sigma_0$,

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

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

Notice that always $\emptyset$ belongs to an algebra, since $\emptyset = S^c$. Of course property (iii) extends to finite unions by induction. Moreover, in an algebra we also have $E, F \in \Sigma_0 \Rightarrow E \cap F \in \Sigma_0$, since $E \cap F = (E^c \cup F^c)^c$. Furthermore $E \setminus F = E \cap F^c \in \Sigma_0$. It follows that an algebra is closed under finitely many of the usual set operations.

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

Alternatively, a collection Σ is a σ-algebra if (i) and (ii) of Definition 1.1 are valid together with $\bigcup_{n=1}^\infty E_n \in \Sigma$ as soon as $E_n \in \Sigma$ $(n = 1, 2 \ldots)$. It follows that an algebra is closed under countably many of the usual set operations.

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. The σ-algebra $\mathcal{B}(\mathbb{R})$ is also generated by all open intervals, or all closed intervals. An alternative generating collection is given next.

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

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

For the proof of the reverse inclusion we proceed in three steps. First we observe that $(-\infty, x) = \cup_n(-\infty, x - \frac{1}{n}] \in \sigma(\mathcal{I})$. Knowing this, we conclude that $(a, b) = (-\infty, b) \setminus (-\infty, a] \in \sigma(\mathcal{I})$. Let then G be an arbitrary open set. Since G is open, for every $x \in G$ there exists a rational $\varepsilon_x > 0$ such that $(x - 2\varepsilon_x, x + 2\varepsilon_x) \subset G$. Consider $(x - \varepsilon_x, x + \varepsilon_x)$ and choose a rational q_x in this interval, note that $|x - q_x| \leq \varepsilon_x$. It follows that $x \in (q_x - \varepsilon_x, q_x + \varepsilon_x) \subset (x - 2\varepsilon_x, x + 2\varepsilon_x) \subset G$. Hence $G \subset \cup_{x \in G}(q_x - \varepsilon_x, q_x + \varepsilon_x) \subset G$, and so $G = \cup_{x \in G}(q_x - \varepsilon_x, q_x + \varepsilon_x)$. But the union here is in fact a countable union, since there are only countably many q_x and ε_x. (Note that the arguments above can be used for any metric space with a countable dense subset to get that an open G is a countable union of open balls.) It follows that $G \in \sigma(\mathcal{I})$, hence $\mathcal{O} \subset \sigma(\mathcal{I})$, and therefore (recall $\mathcal{B}$ is the smallest σ-algebra containing $\mathcal{O}$) $\mathcal{B} \subset \sigma(\mathcal{I})$. □

An obvious question to ask is whether every subset of $\mathbb{R}$ belongs to $\mathcal{B} = \mathcal{B}(\mathbb{R})$. The answer is no. It is a fact, albeit not easy to prove, that the cardinality of $\mathcal{B}(\mathbb{R})$ is the same as the cardinality of $\mathbb{R}$, from which the negative answer follows.

1.2 Measures

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

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

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

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

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

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

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

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

The proof of this theorem can be found in the literature, or in the extended version of these notes. 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.

Here are the first elementary properties of a measure.

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

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

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

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

If μ is finite, we also have

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

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

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

Measures have certain continuity properties.

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

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

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

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

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

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

Proof Exercise 1.2. □

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

1.3 Null sets

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

1.4 π- and d-systems

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

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

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

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

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

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

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

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

Proof Let Σ be a π-system and a d-system. We check the defining conditions of a σ-algebra. (i) Since Σ is a d-system, $S \in \Sigma$. (ii) Complements of sets in Σ are in Σ as well, again because Σ is a d-system. (iii) If $E, F \in \Sigma$, then

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

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

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

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

Step 1. Put $\mathcal{D}_1 = \{B \in d(\mathcal{I}) : B \cap C \in d(\mathcal{I}), \forall C \in \mathcal{I}\}$. We claim that $\mathcal{D}_1$ is a d-system. Given that this holds and because, obviously, $\mathcal{I} \subset \mathcal{D}_1$, also $d(\mathcal{I}) \subset \mathcal{D}_1$. Since $\mathcal{D}_1$ is defined as a subset of $d(\mathcal{I})$, we conclude that these two collections are the same. We now show that the claim holds. Evidently $S \in \mathcal{D}_1$. Let $B_1, B_2 \in \mathcal{D}_1$ with $B_1 \subset B_2$ and $C \in \mathcal{I}$. Write $(B_2 \setminus B_1) \cap C$ as $(B_2 \cap C) \setminus (B_1 \cap C)$. The last two intersections belong to $d(\mathcal{I})$ by definition of $\mathcal{D}_1$ and so does their difference, since $d(\mathcal{I})$ is a d-system. For $B_n \uparrow B$, $B_n \in \mathcal{D}_1$ and $C \in \mathcal{I}$ we have $(B_n \cap C) \in d(\mathcal{I})$ which then converges to $B \cap C \in d(\mathcal{I})$. So $B \in \mathcal{D}_1$.

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

Sometimes another version of Lemma 1.13 is useful.

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

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

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

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

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

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

1.5 Probability language

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

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

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

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

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

Note that the sets $F_m = \cap_{n \geq m} E_n$ form an increasing sequence and the sets $D_m = \cup_{n \geq m} E_n$ form a decreasing sequence. Clearly, $\mathcal{F}$ is closed under taking limsup and liminf. The terminology is explained by (i) of Exercise 1.4. In probabilistic terms, $\limsup E_n$ is described as the event that the E_n occur *infinitely often*, abbreviated by E_n i.o. Likewise, $\liminf E_n$ is the event that the E_n occur *eventually*. The former interpretation follows by observing that $\omega \in \limsup E_n$ iff for all m, there exists $n \geq m$ such that $\omega \in E_n$. In other words, a particular outcome ω belongs to $\limsup E_n$ iff it belongs to some (infinite) subsequence of (E_n).

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

1.6 Exercises

1.1 Prove the following statements.

(a) The intersection of an arbitrary family of d-systems is again a d-system.

(b) The intersection of an arbitrary family of σ-algebras is again a σ-algebra.

(c) If $\mathcal{C}_1$ and $\mathcal{C}_2$ are collections of subsets of Ω with $\mathcal{C}_1 \subset \mathcal{C}_2$, then $d(\mathcal{C}_1) \subset d(\mathcal{C}_2)$.

1.2 Prove Corollary 1.8.

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

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

(a) Show that $\mathbf{1}_{\liminf E_n} = \liminf \mathbf{1}_{E_n}$.

(b) Show that $\mu(\liminf E_n) \leq \liminf \mu(E_n)$. (Use Proposition 1.7.)

(c) Show also that $\mu(\limsup E_n) \geq \limsup \mu(E_n)$, provided that μ is finite.

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

(a) Show that Σ^* is a σ-algebra and that $\Sigma^* = \Sigma \vee \mathcal{N} (= \sigma(\mathcal{N} \cup \Sigma))$.

(b) Show that μ^* restricted to Σ coincides with μ and that $\mu^*(E)$ doesn't depend on the specific choice of F in its definition.

(c) Show that the collection of (μ^*, Σ^*)-null sets is $\mathcal{N}$.

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

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

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

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

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

1.11 Let $S = \{1, 2, 3, 4\}$, $\mathcal{C} = \{\{1, 2\}, \{3\}\}$, $\mathcal{F} = \sigma(\mathcal{C})$. A measure μ is fixed by $\mu(\{1, 2\}) = 0$, $\mu(\{3\}) = 1$, $\mu(\{4\}) = 1$.

(a) Give all sets in $\mathcal{F}$ (there are 8 of them) together with their measure.

(b) Show that $\{1\}$ and $\{1, 2\}$ are null sets. Do they belong to $\mathcal{F}$?

(c) Let $\mathcal{F}'$ be the power set of S (all subsets of S). Put $\mu'(\{1\}) = 0$ and $\mu'(E) = \mu(E)$ for all $E \in \mathcal{F}$. Show that μ' is the unique extension of μ to $\mathcal{F}'$.

(d) What are the μ'-null sets? Are they elements of $\mathcal{F}'$?

1.12 Consider $\mathbb{R}$ together with the Borel sets and the Lebesgue measure λ.

(a) Show that all finite and countable sets have measure zero. What is $\lambda(\mathbb{Q})$?

(b) Is $\lambda(\mathbb{R}) = \sum_{x \in \mathbb{R}} \lambda(\{x\})$ (whatever the sum could mean)?

2 Measurable functions and random variables

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

2.1 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 $\mathbb{R}$. Finally, recall some notation. For a mapping $h : S \to \mathbb{R}$ and $B \subset \mathbb{R}$, the set $\{s \in S : h(s) \in B\}$ is denoted $h^{-1}[B]$.

Definition 2.1 A mapping $h : S \to \mathbb{R}$ is called *measurable* if $h^{-1}[B] \in \Sigma$ for all $B \in \mathcal{B}$.

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

Remark 2.2 Consider $E \subset S$. Recall that the indicator function of E is defined by $\mathbf{1}_E(s) = 1$ if $s \in E$ and $\mathbf{1}_E(s) = 0$ if $s \notin E$. Check that $\mathbf{1}_E$ is a measurable function iff E is a measurable set.

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

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

Proposition 2.3 *Let (S, Σ) be a measurable space and $h : S \to \mathbb{R}$.*

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

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

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

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

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

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

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

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

Remark 2.4 There are many variations on the assertions of Proposition 2.3 possible. For instance in (ii) we could also use $\{h < c\}$, or $\{h > c\}$. Furthermore, (ii) is true for $h : S \to [-\infty, \infty]$ as well. We proved (iv) by a simple composition argument, which also applies to a more general situation. Let (S_i, Σ_i) be measurable spaces ($i = 1, 2, 3$), $h : S_1 \to S_2$ is Σ_1/Σ_2-measurable and $f : S_2 \to S_3$ is Σ_2/Σ_3-measurable. Then $f \circ h$ is Σ_1/Σ_3-measurable.

The set of measurable functions will also be denoted by Σ. This notation is of course a bit ambiguous, but it turns out, that no confusion can arise. Remark 2.2, in a way justifies this notation. The remark can, with the present convention, be rephrased as $\mathbf{1}_E \in \Sigma$ iff $E \in \Sigma$. Later on we often need the set of nonnegative measurable functions, denoted Σ^+.

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

Proposition 2.5 *We have the following properties.*

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

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

Proof (i) If $h \in \Sigma$ and $\lambda \in \mathbb{R}$, then λh is also measurable (use (ii) of the previous proposition for $\lambda \neq 0$). To show that the sum of two measurable functions is measurable, we first note that $\{(x_1, x_2) \in \mathbb{R}^2 : x_1 + x_2 > c\} = \cup_{q \in \mathbb{Q}} \{(x_1, x_2) \in \mathbb{R}^2 : x_1 > q, x_2 > c - q\}$ (draw a picture!). But then we also have $\{h_1 + h_2 > c\} = \cup_{q \in \mathbb{Q}}(\{h_1 > q\} \cap \{h_2 > c - q\})$, a countable union. To show that products of measurable functions are measurable is left as Exercise 2.1.

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

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

Theorem 2.6 *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, then $f \in \mathcal{H}$.

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

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

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

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

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

2.2 Random variables

We return to the setting of Section 1.5 and so we consider a set (of outcomes) Ω and $\mathcal{F}$ a σ-algebra (of events) defined on it. In this setting Definition 2.1 takes the following form.

Definition 2.7 A function $X : \Omega \to \mathbb{R}$ is called a *random variable* if it is ($\mathcal{F}$-)measurable.

Following the tradition, we denote random variables by X (or other capital letters), rather than by h, as in the previous sections. By definition, random variables are nothing else but measurable functions with respect to a given σ-algebra $\mathcal{F}$. Given $X : \Omega \to \mathbb{R}$, let $\sigma(X) = \{X^{-1}[B] : B \in \mathcal{B}\}$. Then $\sigma(X)$ is a σ-algebra, and X is a random variable in the sense of Definition 2.7 iff $\sigma(X) \subset \mathcal{F}$. It follows that $\sigma(X)$ is the smallest σ-algebra on Ω such that X is a random variable. See also Exercise 2.2.

If we have a collection of mappings $X := \{X_i : \Omega \to \mathbb{R} | i \in I\}$, then we denote by $\sigma(X)$ the smallest σ-algebra on Ω such that all the X_i become measurable. See Exercise 2.3.

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

$$\mu(B) := \mathbb{P}(X \in B) = \mathbb{P}(X^{-1}[B]). \tag{2.1}$$

It is straightforward to check that μ is a probability measure on $\mathcal{B}$. Commonly used alternative notations for μ are $\mathbb{P}^X$, or $\mathcal{L}_X$, $\mathcal{L}^X$. This probability measure is referred to as the *distribution* of X or the *law* of X. Along with the distribution of X, we introduce its *distribution function*, usually denoted by F (or F_X, or F^X). By definition it is the function $F : \mathbb{R} \to [0,1]$, given by $F(x) = \mu((-\infty, x]) = \mathbb{P}(X \leq x)$.

Proposition 2.8 *The distribution function of a random variable is right continuous, non-decreasing and satisfies* $\lim_{x\to\infty} F(x) = 1$ *and* $\lim_{x\to-\infty} F(x) = 0$. *The set of points where F is discontinuous is at most countable.*

Proof Exercise 2.4. □

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

Proposition 2.9 *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 2.8 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 2.10, we justify this terminology. We will see that for any distribution function F, it is possible to construct a random variable on some $(\Omega, \mathcal{F}, \mathbb{P})$, whose distribution function equals F. This theorem is founded on the existence of the Lebesgue measure λ on the Borel sets $\mathcal{B}[0,1]$ of $[0,1]$, see Theorem 1.5.

We now give a probabilistic translation of this theorem. Consider $(\Omega, \mathcal{F}, \mathbb{P}) = ([0,1], \mathcal{B}[0,1], \lambda)$. Let $U : \Omega \to [0,1]$ be the identity map. The distribution of U on $[0,1]$ is trivially the Lebesgue measure again, in particular the distribution function F^U of U satisfies $F^U(x) = x$ for $x \in [0,1]$ and so $\mathbb{P}(a < U \leq b) = F^U(b) - F^U(a) = b - a$ for $a, b \in [0,1]$ with $a \leq b$. Hence, to the distribution function F^U corresponds a probability measure on $([0,1], \mathcal{B}[0,1])$ and there exists a random variable U on this space, such that U has F^U as its distribution function. The random variable U is said to have the standard uniform distribution.

The proof of Theorem 2.10 (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 2.6), 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 2.10 *Let F be a distribution function on $\mathbb{R}$. Then there exists a probability space and a random variable $X : \Omega \to \mathbb{R}$ such that F is the distribution function of X.*

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

$$X^-(\omega) \leq c \Leftrightarrow F(c) \geq \omega. \tag{2.2}$$

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 (2.2) 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 (2.2). In order to find the distribution function of X^-, we compute $\mathbb{P}(X^- \leq c) = \mathbb{P}([0, F(c)]) = \lambda([0, F(c)]) = F(c)$. □

2.3 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 2.11 We have the following different definitions of independence, in decreasing order of generality.

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

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

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

The above definition also applies to finite sequences. For instance, a finite sequence of σ-algebras $\mathcal{F}_1, \ldots, \mathcal{F}_n$ is called independent if the infinite sequence $\mathcal{F}_1, \mathcal{F}_2, \ldots$ is independent in the sense of part (ii) of the above definition, where $\mathcal{F}_m = \{\emptyset, \Omega\}$ for $m > n$. It follows that two σ-algebras $\mathcal{F}_1$ and $\mathcal{F}_2$ are independent, if $\mathbb{P}(E_1 \cap E_2) = \mathbb{P}(E_1)\mathbb{P}(E_2)$ for all $E_1 \in \mathcal{F}_1$ and $E_2 \in \mathcal{F}_2$. It also follows that two events E and F are independent iff $\sigma(E)$ and $\sigma(F)$ are independent σ-algebras, this is Exercise 2.12.

To check independence of two σ-algebras, Theorem 1.15 is again helpful. It tells you that for two σ-algebras to be independent, it is sufficient to check the product rule for generating π-systems.

Proposition 2.12 *Let $\mathcal{I}$ and $\mathcal{J}$ be π-systems and suppose that for all $I \in \mathcal{I}$ and $J \in \mathcal{J}$ the product rule $\mathbb{P}(I \cap J) = \mathbb{P}(I)\mathbb{P}(J)$ holds. Then the σ-algebras $\sigma(\mathcal{I})$ and $\sigma(\mathcal{J})$ are independent.*

Proof Put $\mathcal{G} = \sigma(\mathcal{I})$ and $\mathcal{H} = \sigma(\mathcal{J})$. We define for each $I \in \mathcal{I}$ the *finite measures* μ_I and ν_I on $\mathcal{H}$ by $\mu_I(H) = \mathbb{P}(H \cap I)$ and $\nu_I(H) = \mathbb{P}(H)\mathbb{P}(I)$ $(H \in \mathcal{H})$. Notice that μ_I and ν_I coincide on $\mathcal{J}$ by assumption and that $\mu_I(\Omega) = \mathbb{P}(I) = \nu_I(\Omega)$. Theorem 1.15 yields that $\mu_I(H) = \nu_I(H)$ for all $H \in \mathcal{H}$.

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

We next present an important consequence concerning two random variables X_1 and X_2 for which we need some notation. Put $F_X : \mathbb{R}^2 \to [0, 1]$, defined by $F_X(x_1, x_2) = \mathbb{P}(\{X_1 \leq x_1\} \cap \{X_2 \leq x_2\})$. The function F_X is called the joint distribution function of X_1 and X_2.

Corollary 2.13 *Let X_1, X_2 be random variables defined on some $(\Omega, \mathcal{F}, \mathbb{P})$. Then X_1 and X_2 are independent iff $\mathbb{P}(\{X_1 \leq x_1\} \cap \{X_2 \leq x_2\}) = \mathbb{P}(X_1 \leq x_1)\mathbb{P}(X_2 \leq x_2)$ for all $x_1, x_2 \in \mathbb{R}$. In terms of the distribution functions this can be written as $F_X(x_1, x_2) = F_{X_1}(x_1)F_{X_2}(x_2)$ for all $x_1, x_2 \in \mathbb{R}$.*

Proof Combine Proposition 2.12 and Exercise 2.2. □

2.4 Exercises

2.1 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.

2.2 Let X be a random variable. Show that $\Pi(X) := \{\{X \leq x\} : x \in \mathbb{R}\}$ is a π-system and that it generates $\sigma(X)$. Formulate a similar statement for a two-dimensional random vector X.

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

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

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

2.4 Prove Proposition 2.8.

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

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

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

2.8 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.

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

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

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

2.10 Consider an independent sequence $X_1, X_2, \ldots$. Let $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$ and $\mathcal{T}_n = \sigma(X_{n+1}, X_{n+2}, \ldots)$, $n \geq 1$. Let $\mathcal{I}$ be the collection of events of the type $\{X_1 \in B_1, \ldots, X_n \in B_n\}$, with the B_i Borel sets in $\mathbb{R}$. Show that $\mathcal{I}$ is a π-system that generates $\mathcal{F}_n$. Find a π-system that generates $\mathcal{T}_n$ and show that $\mathcal{F}_n$ and $\mathcal{T}_n$ are independent. (Use Proposition 2.12.)

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

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

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

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

2.12 Show that two events E and F are independent iff $\sigma(E)$ and $\sigma(F)$ are independent σ-algebras

3 Integration

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

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

3.1 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 $E \in \Sigma$. This should be seen as the product of height times width, the classical formula, with height equal to a and width equal to the measure of E, $\mu(E)$. We extend this definition to the class of *simple functions*.

Definition 3.1 A function $f : S \to [0,\infty)$ is called a nonnegative simple function, if it has a representation as a finite sum

$$f = \sum_{i=1}^n a_i \mathbf{1}_{A_i}, \tag{3.1}$$

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

The representation in (3.1) is inherently non-unique. Take $n=2$, $a_1 = a_2 = a$, A_1 and A_2 disjoint sets and think of a representation of f with $n=1$. Note that a simple function is measurable. Since we remember that Riemann integrals are linear operators and knowing the definition of integral for an indicator function, we now present the definition of the integral of $f \in \mathfrak{S}^+$.

Definition 3.2 Let $f \in \mathfrak{S}^+$. The (*Lebesgue*) *integral* of f with respect to the measure μ is defined as

$$\int f \, \mathrm{d}\mu := \sum_{i=1}^{n} a_i \mu(A_i), \tag{3.2}$$

when f has representation (3.1).

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

For products ab for $a, b \in [0, \infty]$, we use the convention $ab = 0$, when $a = 0$.

It should be clear that this definition of integral is, at first sight, troublesome. As the representation of a simple function is not unique, 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 3.3 *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 (3.1) and define $\phi : S \to \{0, 1\}^n$ by $\phi(s) = (\mathbf{1}_{A_1}(s), \ldots, \mathbf{1}_{A_n}(s))$. Let $\{0, 1\}^n = \{u_1, \ldots, u_m\}$ where $m = 2^n$ and put $U_k = \phi^{-1}(u_k)$. Then the collection $\{U_1, \ldots, U_m\}$ is a measurable partition of S (the sets U_k are measurable). We will also need the sets $S_i = \{k : U_k \subset A_i\}$ and $T_k = \{i : U_k \subset A_i\}$. Note that these sets are dual in the sense that $k \in S_i$ iff $i \in T_k$.

Below we will use the fact $A_i = \cup_{k \in S_i} U_k$, when we rewrite (3.1). We obtain by interchanging the summation order

$$\begin{aligned} f = \sum_i a_i \mathbf{1}_{A_i} &= \sum_i a_i \Big(\sum_{k \in S_i} \mathbf{1}_{U_k} \Big) \\ &= \sum_k \Big(\sum_{i \in T_k} a_i \Big) \mathbf{1}_{U_k}. \end{aligned} \tag{3.3}$$

Now apply the definition of $\mu(f)$ by using the representation of f given by (3.3). This gives $\mu(f) = \sum_k (\sum_{i \in T_k} a_i) \mu(U_k)$. Interchanging the summation order, we see that this is equal to $\sum_i a_i (\sum_{k \in S_i} \mu(U_k)) = \sum_i a_i \mu(A_i)$, which coincides with (3.2). We conclude that if f is given by (3.1), we can also represent f in a similar fashion by using a partition, and that both representations give the same value for the integral.

Step 2: Suppose that we have two representations of a simple function f, one is as in (3.1) with the collection of A_i a measurable partition of S. The other one is

$$f = \sum_{j=1}^{m} b_j \mathbf{1}_{B_j}, \tag{3.4}$$

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 (3.2) by using the representation (3.1) of f, but $\sum_j b_j \mu(B_j)$ if we use (3.4). 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 (3.1) and (3.4). According to step 1, we can replace each of them with a representation in terms of a measurable partition, without changing the value of the integral. According to step 2, each of the representations in terms of the partitions also gives the same value of the integral. This proves the proposition.

□

Corollary 3.4 *Let $f \in \mathfrak{S}^+$ and suppose that f assumes the* different *values $0 \leq a_1, \ldots, a_n < \infty$. Then $\mu(f) = \sum_{i=1}^n a_i \mu(\{f = a_i\})$. If f is indentically zero, then $\mu(f) = 0$.*

Proof We have the representation $f = \sum_{i=1}^n a_i \mathbf{1}_{\{f=a_i\}}$. The expression for $\mu(f)$ follows from Definition 3.2, which is unambiguous by Proposition 3.3. The result for the zero function follows by the representation $f = 0 \times \mathbf{1}_S$ and the convention $ab = 0$ for $a = 0$ and $b \in [0, \infty]$.

□

Example 3.5 Here is an instructive example. Let $(S, \Sigma, \mu) = (\mathbb{N}, 2^{\mathbb{N}}, \tau)$, with counting measure τ. A function f on $\mathbb{N}$ can be identified with a sequence (f_i). Then f can be represented in a somewhat cumbersome way (but it just means that the value $f(k)$ of the function f at the variable k equals the number f_k) by

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

For now, we assume $f_i = 0$ for $i > n$ and $f_i \geq 0$ for $i \leq n$; obviously, f is a simple function. Since $\tau(\{i\}) = 1$, we get $\tau(f) = \sum_{i=1}^n f_i$, nothing else but the finite sum of the f_i. In this case, integration is just summation. Of course, a different representation would yield the same answer. A generalization occurs when the set of values $\{f_i : i \in \mathbb{N}\}$ is finite. Then f is still a simple function if the f_i are nonnegative, as in Corollary 3.4. But note that now it may happen that $\tau(f) = \infty$ (think of all f_i equal to one).

Example 3.6 Let $(S, \Sigma, \mu) = ([0,1], \mathcal{B}([0,1]), \lambda)$ and f the indicator of the rational numbers in $[0,1]$, $f = \mathbf{1}_{\mathbb{Q}\cap[0,1]}$. We know that $\lambda(\mathbb{Q}\cap[0,1]) = 0$ and it follows that $\lambda(f) = 0$. This f is a nice example of a function that is not Riemann integrable, whereas its Lebesgue integral trivially exists and has a very sensible value.

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

Proposition 3.7 *Let $f, g \in \mathfrak{S}^+$ and $c \in [0, \infty)$.*

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

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

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

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

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

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

3.2 A general definition of integral

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

Definition 3.8 Let f be a nonnegative measurable function. The integral of f is defined as $\mu(f) := \sup\{\mu(h) : h \leq f, h \in \mathfrak{S}^+\}$, where $\mu(h)$ is as in Definition 3.2.

Notice that for functions $f \in \mathfrak{S}^+$, Definition 3.8 yields for $\mu(f)$ the same as Definition 3.2 in the previous section. Thus there is no ambiguity in notation by using the same symbol μ. We immediately have some extensions of results in the previous section.

Proposition 3.9 *Let $f, g \in \Sigma^+$. If $f = 0$ a.e., then $\mu(f) = 0$. If $f \leq g$ a.e., then $\mu(f) \leq \mu(g)$, and if $f = g$ a.e., then $\mu(f) = \mu(g)$.*

Proof Let $f \in \Sigma^+$, $f = 0$ a.e. Take $h \in \mathfrak{S}^+$ with $h \leq f$. From this inequality we obtain $\{h > 0\} \subset \{f > 0\}$, and hence $\mu(\{h > 0\}) \leq \mu(\{f > 0\})$, but the latter measure is zero and hence $h = 0$ a.e. By Corollary 3.4 and Proposition 3.7(ii), $\mu(h) = 0$. Therefore, $\mu(f)$, being the supremum of those $\mu(h)$, is also zero.

Let $f, g \in \Sigma^+$ and $N = \{f > g\}$. Take $h \in \mathfrak{S}^+$ with $h \leq f$. Then also $h\mathbf{1}_N, h\mathbf{1}_{N^c} \in \mathfrak{S}^+$ and by Proposition 3.7(iii) and the fact that $h\mathbf{1}_N = 0$ a.e., we then have $\mu(h) = \mu(h\mathbf{1}_N) + \mu(h\mathbf{1}_{N^c}) = \mu(h\mathbf{1}_{N^c})$. Moreover,

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

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

Example 3.10 We extend the situation of Example 3.5, by allowing infinitely many f_i to be positive. The result will be $\tau(f) = \sum_{i=1}^\infty f_i$, classically defined as $\lim_{n\to\infty} \sum_{i=1}^n f_i$. Check that this is in agreement with Definition 3.8. See also Exercise 3.1.

The following will frequently be used.

Lemma 3.11 *Let $f \in \Sigma^+$ and suppose that $\mu(f) = 0$. Then $f = 0$ a.e.*

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

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

Theorem 3.12 *Let (f_n) be a sequence in Σ^+, such that $f_{n+1} \geq f_n$ a.e. for each n. Let $f = \limsup f_n$. Then $\mu(f_n) \uparrow \mu(f) \leq \infty$.*

Proof We first consider the case where $f_{n+1}(s) \geq f_n(s)$ for all $s \in S$, so (f_n) is increasing *everywhere*. Then $f(s) = \lim f_n(s)$ for all $s \in S$, possibly with value infinity. It follows from Proposition 3.9, that $\mu(f_n)$ is an increasing sequence, bounded by $\mu(f)$. Hence we have $\ell := \lim \mu(f_n) \leq \mu(f)$.

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

$$\ell \geq \mu(f_n) \geq \mu(f_n \mathbf{1}_{E_n}) \geq c\mu(h\mathbf{1}_{E_n}). \tag{3.5}$$

Suppose that h has representation (3.1). Then $\mu(h\mathbf{1}_{E_n}) = \sum_i a_i \mu(A_i \cap E_n)$. This is a finite sum of nonnegative numbers and hence the limit of it for $n \to \infty$ can be taken inside the sum and thus equals $\mu(h)$, since $E_n \uparrow S$ and the continuity of the measure (Proposition 1.7). From (3.5) we then conclude $\ell \geq c\mu(h)$, for all $c \in (0,1)$, and thus $\ell \geq \mu(h)$. Since this holds for all our h, we get $\ell \geq \mu(f)$ by taking the supremum over h. This proves the first case.

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

Example 3.13 Here is a nice application of Theorem 3.12. Let $f \in \Sigma^+$ and, for each $n \in \mathbb{N}$, put $E_{n,i} = \{i2^{-n} \leq f < (i+1)2^{-n}\}$ $(i \in I_n := \{0, \ldots, n2^n - 1\})$, similar to the sets in the proof of Theorem 2.6. Put also $E_n = \{f \geq n\}$. Note that the sets $E_{n,i}$ and E_n are in Σ. Define

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

These f_n form an increasing sequence in Σ^+, even in $\mathfrak{S}^+$, with limit f. Theorem 3.12 yields $\mu(f_n) \uparrow \mu(f)$. We have exhibited a sequence of simple functions with limit f, that can be used to approximate $\mu(f)$.

Proposition 3.14 *Let $f, g \in \Sigma^+$ and $\alpha, \beta > 0$. Then $\mu(\alpha f + \beta g) = \alpha\mu(f) + \beta\mu(g) \leq \infty$.*

Proof Exercise 3.2. □

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

Lemma 3.15 *Let (f_n) be an arbitrary sequence in Σ^+. Then $\liminf \mu(f_n) \geq \mu(\liminf f_n)$. If there exists a function $h \in \Sigma^+$ such that $f_n \leq h$ a.e., and $\mu(h) < \infty$, then $\limsup \mu(f_n) \leq \mu(\limsup f_n)$.*

Proof Put $g_n = \inf_{m \geq n} f_m$. We have for all $m \geq n$ the inequality $g_n \leq f_m$. Then also $\mu(g_n) \leq \mu(f_m)$ for $m \geq n$, and even $\mu(g_n) \leq \inf_{m \geq n} \mu(f_m)$. We want to take limits on both side of this inequality. On the right hand side we get $\liminf \mu(f_n)$. The sequence (g_n) is increasing, with limit $g = \liminf f_n$, and by Theorem 3.12, $\mu(g_n) \uparrow \mu(\liminf f_n)$ on the left hand side. This proves the first assertion. The second assertion follows by considering $\bar{f}_n = h - f_n \geq 0$. Check where it is used that $\mu(h) < \infty$. □

Remark 3.16 Let (E_n) be a sequence of sets in Σ, and let $f_n = \mathbf{1}_{E_n}$ and $h = 1$. The statements of Exercise 1.4 follow from Lemma 3.15.

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

Definition 3.17 Let $f \in \Sigma$ and assume that $\mu(f^+) < \infty$ or $\mu(f^-) < \infty$. Then we define $\mu(f) := \mu(f^+) - \mu(f^-)$. If both $\mu(f^+) < \infty$ and $\mu(f^-) < \infty$, we say that f is *integrable*. The collection of all integrable functions is denoted by $\mathcal{L}^1(S, \Sigma, \mu)$. Note that $f \in \mathcal{L}^1(S, \Sigma, \mu)$ implies that $|f| < \infty$ μ-a.e.

Proposition 3.18 *The following natural properties hold.*

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

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

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

Proof Exercise 3.3. □

Example 3.19 Let $(S, \Sigma, \mu) = ([0, 1], \mathcal{B}([0, 1]), \lambda)$, where λ is Lebesgue measure. Assume that $f \in C[0, 1]$. Exercise 3.5 yields that $f \in \mathcal{L}^1([0, 1], \mathcal{B}([0, 1]), \lambda)$ and that $\lambda(f)$ is equal to the Riemann integral $\int_0^1 f(x)\,\mathrm{d}x$. This implication fails to hold if we replace $[0, 1]$ with an unbounded interval, see Exercise 3.6.

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. More knowledge is required for a precise statement and its proof.

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

Theorem 3.20 *Let $(f_n) \subset \Sigma$ and $f \in \Sigma$. Assume that $f_n(s) \to f(s)$ for all s outside a set of measure zero. Assume also that there exists a function $g \in \Sigma^+$ such that $\sup_n |f_n| \leq g$ a.e. and that $\mu(g) < \infty$. Then $\mu(|f_n - f|) \to 0$ (often denoted $f_n \xrightarrow{\mathcal{L}^1} f$), and hence $\mu(f_n) \to \mu(f)$.*

Proof The second assertion easily follows from the first one, which we prove now for the case that $f_n \to f$ *everywhere*. One has the inequality $|f| \leq g$, whence $|f_n - f| \leq 2g$. The second assertion of Fatou's lemma immediately yields $\limsup \mu(|f_n - f|) \leq 0$, which is what we wanted. The *almost everywhere* version is left as Exercise 3.4. □

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.

3.3 Integrals over subsets

This section is in a sense a prelude to the theorem of Radon-Nikodym, Theorem 5.4. Let $f \in \Sigma^+$ and $E \in \Sigma$. Then we may define

$$\int_E f \, \mathrm{d}\mu := \mu(\mathbf{1}_E f). \tag{3.6}$$

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

Proposition 3.21 *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 3.7. □

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

$$\nu(E) = \mu(\mathbf{1}_E f) \, (= \int_E f \, \mathrm{d}\mu). \tag{3.7}$$

One verifies (Exercise 3.8) 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 (3.7). More generally we have

Proposition 3.22 *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 3.9. □

For the measure ν above, Proposition 3.22 states that $\int h \, \mathrm{d}\nu = \int hf \, \mathrm{d}\mu$, valid for all $h \in \mathcal{L}^1(S, \Sigma, \nu)$. The notation $f = \frac{\mathrm{d}\nu}{\mathrm{d}\mu}$ is often used and looks like a derivative. We will return to this in Chapter 5, where we discuss Radon-Nikodym derivatives. The equality $\int h \, \mathrm{d}\nu = \int hf \, \mathrm{d}\mu$ now takes the appealing form

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

Example 3.23 Let $(S, \Sigma, \mu) = (\mathbb{R}, \mathcal{B}, \lambda)$, $f \geq 0$, Borel measurable, $\nu(E) = \int \mathbf{1}_E f \, \mathrm{d}\lambda$ and $hf \in \mathcal{L}^1(\mathbb{R}, \mathcal{B}, \lambda)$. Then

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

where the equality is valid under conditions as for instance in Example 3.19.

Remark 3.24 If f is continuous, see Example 3.19, then $x \mapsto F(x) = \int_{[0,x]} f \, \mathrm{d}\lambda$ defines a differentiable function on $(0, 1)$, with $F'(x) = f(x)$. This follows from the theory of Riemann integrals. We adopt the conventional notation $F(x) = \int_0^x f(u) \, \mathrm{d}u$. This case can be generalized as follows. If $f \in \mathcal{L}^1(\mathbb{R}, \mathcal{B}, \lambda)$, then (using a similar notational convention) $x \mapsto F(x) = \int_{-\infty}^x f(u) \, \mathrm{d}u$ is well defined for all $x \in \mathbb{R}$. Moreover, F is at (Lebesgue) almost all points x of $\mathbb{R}$ differentiable with derivative $F'(x) = f(x)$. The proof of this result, the *fundamental theorem of calculus for the Lebesgue integral*, is not given here.

3.4 Expectation and integral

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

$$\mathbb{P}(X) = \int_\Omega X \, \mathrm{d}\mathbb{P}, \tag{3.8}$$

provided that the integral is well defined, which is certainly the case if X is nonnegative or if $\mathbb{P}(|X|) < \infty$. Other often used notations for the integral in (3.8) are $\mathbb{P}X$ and $\mathbb{E}\,X$. The latter is the favorite one among probabilists and one speaks of the Expectation of X. Note also that $\mathbb{E}\,X$ is always defined when $X \geq 0$ *almost surely*. The latter concept meaning almost everywhere w.r.t. the probability measure $\mathbb{P}$. We abbreviate almost surely by a.s.

Example 3.25 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, a function on $\Omega = \mathbb{N}$, by $X(n) = x_n$. In a spirit similar to what we have seen in Examples 3.5 and 3.10, we get $\mathbb{E}\,X = \sum_{n=1}^\infty x_n p_n$. Let us switch to a different approach. Let $\xi_1, \xi_2, \ldots$ be the different elements of the set $\{x_1, x_2, \dots\}$ 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 $h : \mathbb{R} \to \mathbb{R}$ is Borel measurable, then $Y := h \circ X$ (we also write $Y = h(X)$) is a random variable as well. There are two recipes to compute $\mathbb{E}\, Y$. One is of course the direct application of the definition of expectation to Y. But we also have

Proposition 3.26 *Let X be a random variable, and $h : \mathbb{R} \to \mathbb{R}$ Borel measurable. Let $\mathbb{P}^X$ be the distribution of X. Then $h \circ X \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ iff $h \in \mathcal{L}^1(\mathbb{R}, \mathcal{B}, \mathbb{P}^X)$, in which case*

$$\mathbb{E}\, h(X) = \int_{\mathbb{R}} h \, \mathrm{d}\mathbb{P}^X. \tag{3.9}$$

Proof Exercise 3.10. □

It follows from this proposition that one can also compute $\mathbb{E}\, Y$ as $\mathbb{E}\, Y = \int_{\mathbb{R}} y \, \mathbb{P}^Y(\,\mathrm{d}y)$, and of course $\mathbb{E}\, X = \int_{\mathbb{R}} x \, \mathbb{P}^X(\,\mathrm{d}x)$.

Example 3.27 Suppose there exists $f \geq 0$, Borel-measurable such that for all $B \in \mathcal{B}$ one has $\mathbb{P}^X(B) = \lambda(\mathbf{1}_B f)$, in which case it is said that X has a *density* f. Then, provided that the expectation is well defined, Example 3.23 yields

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

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

We conclude that the definition of expectation as a Lebesgue integral w.r.t. a probability measure as in (3.8) yields the familiar formulas, sums for discrete random variables and Riemann integrals for random variables having an ordinary density function, as special cases. So, we see that the Lebesgue integral serves as a unifying concept for expectation. At least as important is that we can use the powerful convergence theorems (obtained for integrals) of Section 3.2 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 3.20) with the function g equal to a constant. Here is a simple example of the application of the Monotone Convergence Theorem.

Example 3.28 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 3.11.

The next two propositions have proven to be very useful in proofs of results in Probability Theory.

Proposition 3.29 *Let X be a real valued random variable and $g : \mathbb{R} \to [0, \infty]$ an increasing function. Then $\mathbb{E}\, g(X) \geq g(c)\mathbb{P}(X \geq c)$.*

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

The inequality in Proposition 3.29 is known as Markov's inequality. An example is obtained by taking $g(x) = x^+$ and by replacing X with $|X|$. One gets $\mathbb{E}\,|X| \geq c\mathbb{P}(|X| \geq c)$. For the special case where $g(x) = (x^+)^2$, it is known as Chebychev's inequality. This name is especially used, if we apply it with $|X - \mathbb{E}\,X|$ instead of X. For $c \geq 0$ we then obtain $\operatorname{Var} X \geq c^2\mathbb{P}(|X - \mathbb{E}\,X| \geq c)$.

We now turn to a result that is known as *Jensen's inequality*, Proposition 3.30 below. Recall that a function $g : G \to \mathbb{R}$ is convex, if G is a convex set and if for all $x, y \in G$ and $\alpha \in [0, 1]$ one has

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

We consider only the case where G is an interval.

Let us first give a property of a convex function. For all x in the interior of G there is a number $d(x)$ such that for all $z \in G$ it holds that

$$g(z) - g(x) \geq d(x)(z - x). \tag{3.10}$$

Verify this property graphically. The $d(x)$ are also called subgradients of g. The following proposition (Jensen's inequality) is now easy to prove, and you check where we use in the proof the fact that $\mathbb{P}$ is a *probability* measure.

Proposition 3.30 *Let $g : G \to \mathbb{R}$ be convex and X a random variable with $\mathbb{P}(X \in G) = 1$. Assume that $\mathbb{E}\,|X| < \infty$ and $\mathbb{E}\,|g(X)| < \infty$. Then*

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

Proof We exclude the trivial case $\mathbb{P}(X = x_0) = 1$ for some $x_0 \in G$. Since $\mathbb{P}(X \in G) = 1$, we have $\mathbb{E}\,X \in \operatorname{Int} G$ (Exercise 3.15) and (3.10) with $x = \mathbb{E}\,X$ and z replaced with X holds a.s. So, in view of (3.10),

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

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

3.5 $\mathcal{L}^p$-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 3.31 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 explain this until the end of this section. It is however obvious that $\mathcal{L}^p := \mathcal{L}^p(\Omega, \mathcal{F}, \mathbb{P})$ is a vector space, since $|X+Y|^p \leq (|X|+|Y|)^p \leq 2^p(|X|^p + |Y|^p)$.

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

Proposition 3.32 *Let $X, Y \in \mathcal{L}^2$. Then $\mathbb{E}\,|XY| \leq ||X||_2\,||Y||_2$.*

Proof If $\mathbb{E}\,Y^2 = 0$, then $Y = 0$ a.s. (Lemma 3.11), so also $XY = 0$ a.s. and there is nothing to prove. Assume then that $\mathbb{E}\,Y^2 > 0$ and let $c = \mathbb{E}\,|XY|/\mathbb{E}\,Y^2$. One trivially has $\mathbb{E}\,(|X| - c|Y|)^2 \geq 0$. But, by the choice of c, the left hand side equals $\mathbb{E}\,X^2 - \frac{(\mathbb{E}\,|XY|)^2}{\mathbb{E}\,Y^2}$. □

Proposition 3.32 tells us that $X, Y \in \mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$ is sufficient to guarantee that the product XY is integrable. For independent X and Y weaker integrability assumptions suffice and the product rule for probabilities of intersections extends to a product rule for expectations.

Proposition 3.33 *Let $X, Y \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ be independent random variables. Then $XY \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ and $\mathbb{E}\,(XY) = \mathbb{E}\,X \cdot \mathbb{E}\,Y$.*

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

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

Proposition 3.34 *Let $1 \leq p \leq r$ and $X \in \mathcal{L}^r(\Omega, \mathcal{F}, \mathbb{P})$, then $X \in \mathcal{L}^p(\Omega, \mathcal{F}, \mathbb{P})$ and $||X||_p \leq ||X||_r$.*

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

3.6 $\mathcal{L}^p$-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 3.31.

Definition 3.35 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 3.36 *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 3.11 and we have a trivial inequality. Let then $0 < ||f||_p < \infty$. We now define a probability measure $\mathbb{P}$ on Σ by

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

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

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

and

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

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

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

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

Remark 3.37 For $p = 2$ Theorem 3.36 yields the Cauchy-Schwarz inequality $||fg||_1 \leq ||f||_2 ||g||_2$ for square integrable functions. Compare to Proposition 3.32.

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

Theorem 3.38 *Let $f, g \in \mathcal{L}^p(S, \Sigma, \mu)$ and $p \in [1, \infty]$. Then $||f+g||_p \leq ||f||_p + ||g||_p$.*

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

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

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

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

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

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

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

Theorem 3.39 *Let $p \in [1, \infty]$. The space $\mathcal{L}^p(\Omega, \mathcal{F}, \mathbb{P})$ is* complete *in the following sense. Let (X_n) be a Cauchy-sequence in $\mathcal{L}^p$: $||X_n - X_m||_p \to 0$ for $n, m \to \infty$. Then there exists* a limit *$X \in \mathcal{L}^p$ such that $||X_n - X||_p \to 0$. The limit is unique in the sense that any other limit X' satisfies $||X - X'||_p = 0$.*

Proof Omitted. □

Remark 3.40 Notice that it follows from Theorem 3.39, and the discussion preceding it, that $L^p(\Omega, \mathcal{F}, \mathbb{P})$ is a truly complete normed space, called a *Banach* space. The same is true for $L^p(S, \Sigma, \mu)$ ($p \in [1, \infty]$), for which you need Exercise 3.12. For the special case $p = 2$ we endow $L^2(S, \Sigma, \mu)$ with the inner product $\langle f, g \rangle := \int fg \, d\mu$, and it is then called a *Hilbert* space. Likewise $L^2(\Omega, \mathcal{F}, \mathbb{P})$ is a Hilbert space with inner product $\langle X, Y \rangle := \mathbb{E}\, XY$.

3.7 Exercises

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

3.2 Prove Proposition 3.14.

3.3 Prove Proposition 3.18 (assume Proposition 3.14). Show also that $|\mu(f)| \leq \mu(|f|)$, if $f \in \mathcal{L}^1(S, \Sigma, \mu)$.

3.4 Prove the 'almost everywhere' version of Theorem 3.20 by using the 'everywhere' version.

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

3.6 Let $f : [0, \infty) \to \mathbb{R}$ be given by $f(x) = \frac{\sin x}{x}$ for $x > 0$ and $f(0) = 1$. Show that $I := \int_0^\infty f(x)\,\mathrm{d}x$ exists as an improper Riemann integral (i.e. the limit $\lim_{T\to\infty} \int_0^T f(x)\,\mathrm{d}x$ exists and is finite), but that $f \notin \mathcal{L}^1([0, \infty), \mathcal{B}([0, \infty)), \lambda)$. In Exercise 4.9 you compute that $I = \frac{\pi}{2}$.

3.7 Prove Proposition 3.21 by means of the standard machinery.

3.8 Verify that ν defined in (3.7) is a measure.

3.9 Prove Proposition 3.22. *Hint*: The standard machine works.

3.10 Prove Proposition 3.26. *Hint*: Use the standard machinery for h.

3.11 Give the details for Example 3.28.

3.12 Give the proof of Theorem 3.39 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 $\mathcal{L}^\infty(S, \Sigma, \mu)$.

3.13 *This exercise concerns a more general version of Theorem 3.20.* Let $(f_n) \subset \Sigma$ and $f = \limsup f_n$ and assume that $f_n(s) \to f(s)$ for all s outside a set of measure zero. Assume also there exist functions $g, g_n \in \Sigma^+$ such that $|f_n| \leq g_n$ a.e. with $g_n(s) \to g(s)$ for all s outside a set of measure zero and that $\mu(g_n) \to \mu(g) < \infty$. Show that $\mu(|f_n - f|) \to 0$.

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

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

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

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

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

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

4 Product measures

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

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

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

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

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

4.1 Product of two measure spaces

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

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

Next to the projections, we now consider *embeddings*. For fixed $s_1 \in S_1$ we define $e_{s_1} : S_2 \to S$ by $e_{s_1}(s_2) = (s_1, s_2)$. Similarly we define $e^{s_2}(s_1) = (s_1, s_2)$. One easily checks that the embeddings e_{s_1} are Σ_2/Σ-measurable and that the e^{s_2} are Σ_1/Σ-measurable (Exercise 4.1). As a consequence we have the following proposition.

Proposition 4.1 *Let $f : S \to \mathbb{R}$ be Σ-measurable. Then the* marginal *mappings $s_1 \mapsto f(s_1, s_2)$ and $s_2 \mapsto f(s_1, s_2)$ are Σ_1-, respectively Σ_2-measurable, for any $s_2 \in S_2$, respectively $s_1 \in S_1$.*

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

Remark 4.2 The converse statement of Proposition 4.1 is in general not true, but a counterexample is beyond the scope of these lecture notes; see the full version for details. There are functions $f : S \to \mathbb{R}$ that are not measurable w.r.t. the product σ-algebra Σ, although the mappings $s_1 \mapsto f(s_1, s_2)$ and $s_2 \mapsto f(s_1, s_2)$ are Σ_1-, respectively Σ_2-measurable. Counterexamples are not obvious, see below for a specific one. Fortunately, there are also conditions that are sufficient to have measurability of f w.r.t. Σ, when measurability of the marginal functions is given. See Exercise 4.8.

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

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

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

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

Lemma 4.3 *Let f be a bounded Σ-measurable function. Then the mappings $I_i^f : S_i \to \mathbb{R}$ are Σ_i-measurable $(i = 1, 2)$. Moreover we have the identity*

$$\mu_1(I_1^f) = \mu_2(I_2^f), \tag{4.1}$$

or, in a more appealing notation,

$$\int_{S_1} \Big(\int_{S_2} f(s_1, s_2)\mu_2(\mathrm{d}s_2) \Big) \mu_1(\mathrm{d}s_1) = \int_{S_2} \Big(\int_{S_1} f(s_1, s_2)\mu_1(\mathrm{d}s_1) \Big) \mu_2(\mathrm{d}s_2). \tag{4.2}$$

Proof We use the Monotone Class Theorem, Theorem 2.6, and so we have to find a good vector space $\mathcal{H}$. The obvious candidate is the collection of all bounded Σ-measurable functions f that satisfy the assertions of the lemma.

First we notice that $\mathcal{H}$ is indeed a vector space, since sums of measurable functions are measurable and by linearity of the integral. Obviously, the constant functions belong to $\mathcal{H}$. Then we have to show that if $f_n \in \mathcal{H}$, $f_n \geq 0$ and $f_n \uparrow f$, where f is bounded, then also $f \in \mathcal{H}$. Of course here the Monotone Convergence Theorem comes into play. First we notice that measurability of the I_i^f follows from measurability of the $I_i^{f_n}$ for all n. Theorem 3.12 yields that the sequences $I_i^{f_n}(s_i)$ are increasing and converging to $I_i^f(s_i)$. Another application of this theorem yields that $\mu_1(I_1^{f_n})$ converges to $\mu_1(I_1^f)$ and that $\mu_2(I_2^{f_n})$ converges to $\mu_2(I_2^f)$. Since $\mu_1(I_1^{f_n}) = \mu_2(I_2^{f_n})$ for all n, we conclude that $\mu_1(I_1^f) = \mu_2(I_2^f)$, whence $f \in \mathcal{H}$.

Next we check that $\mathcal{H}$ contains the indicators of sets in $\mathcal{R}$. A quick computation shows that for $f = \mathbf{1}_{E_1 \times E_2}$ one has $I_1^f = \mathbf{1}_{E_1}\mu_2(E_2)$, which is Σ_1-measurable, $I_2^f = \mathbf{1}_{E_2}\mu_1(E_1)$, and $\mu_1(I_1^f) = \mu_2(I_2^f) = \mu_1(E_1)\mu_2(E_2)$. Hence $f \in \mathcal{H}$. By Theorem 2.6 we conclude that $\mathcal{H}$ coincides with the space of all bounded Σ-measurable functions. □

It follows from Lemma 4.3 that for all $E \in \Sigma$, the indicator function $\mathbf{1}_E$ satisfies the assertions of the lemma. This shows that the following definition is meaningful.

Definition 4.4 We define $\mu : \Sigma \to [0, \infty)$ by $\mu(E) = \mu_2(I_2^{\mathbf{1}_E})$ for $E \in \Sigma$.

In Theorem 4.5 below (known as Fubini's theorem) 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 4.5 *The mapping μ of Definition 4.4 has the following properties.*

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

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

$$\mu(f) = \mu_2(I_2^f) = \mu_1(I_1^f) \leq \infty. \tag{4.3}$$

(iii) If $f \in \mathcal{L}^1(S, \Sigma, \mu)$, then Equation (4.3) 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 (4.3) 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 4.2. □

Theorem 4.5 has been proved under the standing assumption that the initial measures μ_1 and μ_2 are finite. The results extend to the case where both these measures are σ-finite. The approach is as follows. Write $S_1 = \cup_{i=1}^\infty S_1^i$ with the $S_1^i \in \Sigma_1$ and $\mu_1(S_1^i) < \infty$. Without loss of generality, we can take the S_1^i disjoint. Take a similar partition (S_2^j) of S_2. Then $S = \cup_{i,j} S_{ij}$, where the $S_{ij} := S_1^i \times S_2^j$, form a countable disjoint union as well. Let $\Sigma_{ij} = \{E \cap S_{ij} : E \in \Sigma\}$. On each measurable space (S_{ij}, Σ_{ij}) the above results apply and one has e.g. identity of the involved integrals by splitting the integration over the sets S_{ij} and adding up the results.

We note that if one goes beyond σ-finite measures (often a good thing to do if one wants to have counterexamples), the assertion may no longer be true. Let $S_1 = S_2 = [0, 1]$ and $\Sigma_1 = \Sigma_2 = \mathcal{B}[0, 1]$. Take μ_1 equal to Lebesgue measure and μ_2 the counting measure, the latter is not σ-finite. It is a nice exercise to show that $\Delta := \{(x, y) \in S : x = y\} \in \Sigma$. Let $f = \mathbf{1}_\Delta$. Obviously $I_1^f(s_1) \equiv 1$ and $I_2^f(s_2) \equiv 0$ and the two iterated integrals in (4.3) are 1 and 0. So, more or less everything above concerning product measures 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. This results in product spaces of the form $(S_1 \times \ldots \times S_n, \Sigma_1 \times \ldots \times \Sigma_n, \mu_1 \times \ldots \times \mu_n)$ under conditions similar to those of Theorem 4.5. The product σ-algebra is again defined as the smallest σ-algebra that makes all projections measurable. Existence of product measures is proved in just the same way as before, using an induction argument. Note that there will be many possibilities to extend (4.1) and (4.2), since there are $n!$ different integration orders. We leave the details to the reader.

4.2 Application in Probability theory

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

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

Section 1.1), and, continuing our discussion of the previous section, the product σ-algebra $\mathcal{B}(\mathbb{R}) \times \mathcal{B}(\mathbb{R})$.

Proposition 4.6 *It holds that* $\mathcal{B}(\mathbb{R}^2) = \mathcal{B}(\mathbb{R}) \times \mathcal{B}(\mathbb{R})$.

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

Remark 4.7 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 4.8 *Let* $X_1, X_2 : \Omega \to \mathbb{R}$ *be given. The vector mapping* $X = (X_1, X_2) : \Omega \to \mathbb{R}^2$ *is a random vector iff the* X_i *are random variables.*

Proof Exercise 4.3. □

Remark 4.9 Let X_1, X_2 be two random variables. We already knew that $X_1 + X_2$ is a random variable too. This also follows from the present results. Let $f : \mathbb{R}^2 \to \mathbb{R}$ be a continuous function. Then it is also $\mathcal{B}(\mathbb{R}^2)$-measurable, and by Corollary 4.8 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 2.2 the distribution, or the law, of a random variable. Next to random variables, one can also look at random *vectors*. These are mappings $X : \Omega \to \mathbb{R}^n$ of which the components, written as X_i for $i = 1, \ldots, n$, are random variables. Also random vectors have distributions, now probability measures on the Borel sets of $\mathbb{R}^n$. So, for the Borel sets B in $\mathbb{R}^n$, $B \in \mathcal{B}(\mathbb{R}^n)$, we define $\mathbb{P}^X(B) = \mathbb{P}(X \in B)$ and as in the one-dimensional case, $\mathbb{P}^X$ is a probability measure on $\mathcal{B}(\mathbb{R}^n)$, the *joint* distribution of X, also called (joint) law of X. The *marginal* distribution of e.g. X^1 (for the other components something similar applies) is given by $\mathbb{P}^{X_1}(E) = \mathbb{P}^X(E \times \mathbb{R}^{n-1})$, with $E \in \mathcal{B}(\mathbb{R})$.

Random vector have distribution functions as well, also called *joint* distribution functions. The latter are functions $F : \mathbb{R}^n \to [0, 1]$, defined by $F(x_1, \ldots, x_n) = \mathbb{P}(\{X_1 \leq x_1\} \cap \cdots \{X_n \leq x_n\})$ for all $(x_1, \ldots, x_n) \in \mathbb{R}^n$. We also use the notation F_X for such a distribution function. Note that one has (in obvious notation) $F_{X_1}(x_1) = \mathbb{P}(\{X_1 \leq x_1\} \cap \mathbb{R}^{n-1})$, which gives the (*marginal*) distribution function of X_1. For the other components one similarly has the (marginal) distribution functions F_{X_i}.

Let us specialize to $n = 2$. Then we have the relations $\mathbb{P}^X(B_1 \times \mathbb{R}) = \mathbb{P}(X_1 \in B_1) = \mathbb{P}^{X_1}(B_1)$ for the marginal distribution, or marginal law, of X_1. The joint distribution function $F = F_X : \mathbb{R}^2 \to [0, 1]$ can be written as

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

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

An important case happens if there exists a nonnegative $\mathcal{B}(\mathbb{R}^2)$-measurable function f such that $\mathbb{P}^X(E) = \int_E f \, \mathrm{d}(\lambda \times \lambda)$, for all $E \in \mathcal{B}(\mathbb{R}^2)$. In that case, f is called the (joint) density of X. The obvious marginal density f_{X_1} of X_1 is defined by $f_{X_1}(x_1) = \int f(x_1, x_2) \, \lambda(\,\mathrm{d}x_2)$. One similarly defines the marginal density of X_2. Check these are indeed densities in the sense of Example 3.27.

Independence (of random variables) had to do with multiplication of probabilities (see Definition 2.11), so it should in a natural way be connected to product measures.

Proposition 4.10 *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 4.4. □

The results of the present section (Proposition 4.6, Corollary 4.8, Proposition 4.10) have obvious extensions to higher dimensional situations. We leave the formulation to the reader.

Remark 4.11 Suppose one is given a random variable X, defined on a given $(\Omega, \mathcal{F}, \mathbb{P})$. Sometimes one needs an additional random variable Y having a specified distribution. It may happen that the given probability space is not rich enough to have such a random variable well defined. Suppose $\Omega = \{0, 1\}$ and $X(\omega) = \omega$, having a Bernoulli distribution for $\mathbb{P}$ defined on the power set of Ω with $\mathbb{P}(\{1\}) = p$. Clearly, it is impossible to define on this Ω a random variable having more than two different outcomes. Extending the probability space to a suitable product space offers a way out, see Exercise 4.13, from which it even follows that X and Y are independent.

4.3 Exercises

4.1 Show that the embeddings e_{s_1} are Σ_2/Σ-measurable and that the e^{s_2} are Σ_1/Σ-measurable. Also prove Proposition 4.1.

4.2 Prove part (iii) of Fubini's theorem (Theorem 4.5) for $f \in \mathcal{L}^1(S, \Sigma, \mu)$ (you already know it for $f \in \Sigma^+$). Explain why $s_1 \mapsto f(s_1, s_2)$ is in $\mathcal{L}^1(S_1, \Sigma_1, \mu_1)$ for all s_2 outside a set N of μ_2-measure zero and that I_2^f is well defined on N^c.

4.3 Prove Corollary 4.8.

4.4 Prove Proposition 4.10.

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

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

Define

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

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

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

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

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

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

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

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

$$\mathbb{E} \sum_{k=1}^{\infty} Z_k = \sum_{k=1}^{\infty} \mathbb{E} \, Z_k. \tag{4.4}$$

(a) Show that this follows from Fubini's theorem (as an alternative to the arguments of Exercise 3.11). If $\sum_{k=1}^{\infty} \mathbb{E} \, Z_k < \infty$, what is $\mathbb{P}(\sum_{k=1}^{\infty} Z_k = \infty)$?

(b) Formulate a result similar to (4.4) for random variables Z_k that may assume negative values as well.

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

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

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

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

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

4.9 Show that for $t > 0$

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

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

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

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

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

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

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

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

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

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

4.11 Let F be the distribution function of a nonnegative random variable X and $\alpha > 0$. Show (use Exercise 4.10 for instance, or write $\mathbb{E}\, X^\alpha = \mathbb{E}\, f(X)$, with $f(x) = \int_0^x \alpha y^{\alpha-1} \, \mathrm{d}y$) that

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

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

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

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

4.13 Let X be a random variable, defined on some $(\Omega, \mathcal{F}, \mathbb{P})$. Let Y be random variable defined on another probability space $(\Omega', \mathcal{F}', \mathbb{P}')$. Consider the product space, with the product σ-algebra and the product probability measure. Redefine X and Y on the product space by $X(\omega, \omega') = X(\omega)$ and $Y(\omega, \omega') = Y(\omega')$. Show that the redefined X and Y are independent and that their marginal distributions are the same as they were originally.

5 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 3.22, known as the Radon-Nikodym theorem, Theorem 5.4.

5.1 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 3.3.

Definition 5.1 Let μ and ν be measures 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 is provided by the measures ν and μ of (3.7), $\nu \ll \mu$. See also Proposition 5.3 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 5.1.

Proposition 5.2 *Let μ, ν_a and ν_s be measures on (S, Σ). Assume that $\nu_a \ll \mu$ and $\nu_s \perp \mu$. Put*

$$\nu = \nu_a + \nu_s. \tag{5.1}$$

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 Omitted. □

The content of Proposition 5.2 is that the decomposition (5.1) of ν, if it exists, is unique. We will see in Section 5.2 that, given a σ-finite measure μ, such a decomposition exists for any σ-finite measure ν and it is called the *Lebesgue decomposition* of ν w.r.t. μ. We extend the definition of the measure ν as given in (3.7) to the real and complex case.

Proposition 5.3 *Let μ be a measure on (S, Σ) and h a nonnegative measurable function on S. Then the map $\nu : \Sigma \to [0, \infty]$ defined by*

$$\nu(E) = \mu(\mathbf{1}_E h) \tag{5.2}$$

is a measure on (S, Σ) that is absolutely continuous w.r.t. μ.

Proof See Exercise 3.8. □

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 (5.2). We will use in that case the notation

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

5.2 The Radon-Nikodym theorem

As an appetizer for the Radon-Nikodym theorem (Theorem 5.4) we consider a special case. Let S be a finite or countable set and $\Sigma = 2^S$. Let μ be a σ-finite measure on (S, Σ) and ν another, finite, measure such that $\nu \ll \mu$. Define $h(x) = \frac{\nu(\{x\})}{\mu(\{x\})}$ if $\mu(\{x\}) > 0$ and zero otherwise. It is easy to verify that $h \in \mathcal{L}^1(S, \Sigma, \mu)$ and

$$\nu(E) = \mu(\mathbf{1}_E h), \forall E \subset S. \tag{5.3}$$

Observe that we have obtained an expression like (5.2), but now starting from the assumption $\nu \ll \mu$. The principal theorem on absolute continuity (and singularity) is the following.

Theorem 5.4 *Let μ be a σ-finite measure and let ν be a finite measure. Then there exists a unique decomposition $\nu = \nu_a + \nu_s$ and a nonnegative function $h \in \mathcal{L}^1(S, \Sigma, \mu)$ such that $\nu_a(E) = \mu(\mathbf{1}_E h)$ for all $E \in \Sigma$ (so $\nu_a \ll \mu$) and $\nu_s \perp \mu$. Moreover, h is unique in the sense that any other h' with this property is such that $\mu(\{h \neq h'\}) = 0$. The function h is called the Radon-Nikodym derivative of ν_a w.r.t. μ and is often written as*

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

Proof Omitted. □

Remark 5.5 If ν is a 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 we may still take $h \geq 0$.

Remark 5.6 The function h of Theorem 5.4, the Radon-Nikodym derivative of ν_a w.r.t. μ, is also called the *density* of ν_a w.r.t. μ. If λ is Lebesgue measure on $(\mathbb{R}, \mathcal{B})$ and ν is the law of a random variable X that is absolutely continuous w.r.t. λ, we have that $F(x) := \nu((-\infty, x]) = \int_{(-\infty,x]} f \,\mathrm{d}\lambda$, where $f = \frac{\mathrm{d}\nu}{\mathrm{d}\lambda}$. Traditionally, the function f was called the density of X, and we see that calling a Radon-Nikodym derivative a density is in agreement with this tradition, but also extends it.

Theorem 5.4 is often used for probability measures $\mathbb{Q}$ and $\mathbb{P}$ with $\mathbb{Q} \ll \mathbb{P}$. Write $Z = \frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}$ and note that $\mathbb{E}\, Z = 1$. It is immediate from Proposition 3.22 that for $X \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{Q})$ one has

$$\mathbb{E}_{\mathbb{Q}}\, X = \mathbb{E}\,[XZ], \tag{5.4}$$

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

5.3 Decomposition of a distribution function

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

Proposition 5.7 *Let F be a distribution function, $F : \mathbb{R} \to \mathbb{R}$. Then there exists a purely discontinuous right-continuous nondecreasing function F_d, with $\lim_{x\to-\infty} F_d(x) = 0$, a nonnegative Borel-measurable function f and a nondecreasing continuous function F_s with $\lim_{x\to-\infty} F_s(x) = 0$ such that the decomposition*

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

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

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

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

One verifies that F_d has the asserted properties, the set of discontinuities of F_d is also D and that $F_c := F - F_d$ is continuous. Up to the normalization constant $1 - F_d(\infty)$, F_c is a distribution function if $F_d(\infty) < 1$ and equal to zero if $F_d(\infty) = 1$. Hence there exists a subprobability measure μ_c on $\mathcal{B}$ such that $\mu_c((-\infty, x]) = F_c(x)$, Theorem 2.10. According to the Radon-Nikodym theorem, we can split $\mu_c = \mu_{ac} + \mu_s$, where μ_{ac} is absolutely continuous w.r.t. Lebesgue measure λ. Hence, there exists a λ-a.e. unique function f in $\mathcal{L}^1_+(\mathbb{R}, \mathcal{B}, \lambda)$ such that $\mu_{ac}(B) = \int_B f\,\mathrm{d}\lambda$. □

We have already encountered two examples, where the above decomposition consists of a single term only. If a random variable X has a discrete distribution, there are x_k, $k = 1, 2, \ldots$ with $\sum_{k\geq1} \mathbb{P}(X = x_k) = 1$, $F = F_d$, and if the distribution function admits a density f, then $F = F_{ac}$. Another extreme case occurs when F is the distribution function of Exercise 5.2, then $F = F_s$. Think of an example of a random variable for which all three terms in the decomposition of Proposition 5.7 are nontrivial.

5.4 Exercises

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

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

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

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

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

(a) Find the distribution of X.

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

$$Y = 3 \sum_{k=1}^{\infty} 4^{-k} X_{2k},$$

where the factor 3 only appears for esthetic reasons. Show that the distribution function $F : [0,1] \to \mathbb{R}$ of Y is constant on $(\frac{1}{4}, \frac{3}{4})$, that $F(1-x) = 1 - F(x)$ and that it satisfies $F(x) = 2F(x/4)$ for $x < \frac{1}{4}$.

(c) Make a sketch of F and show that F is continuous, but not absolutely continuous w.r.t. Lebesgue measure. (Hence there is no Borel measurable function f such that $F(x) = \int_{[0,x]} f(u) \, \mathrm{d}u$, $x \in [0,1]$).

5.3 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.

5.4 Let μ and ν be σ-finite measures and ϕ an arbitrary measure on a measurable space (S, Σ). Assume that $\phi \ll \nu$ and $\nu \ll \mu$. Show that $\phi \ll \mu$ and that

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

5.5 Let ν and μ be σ-finite measures on (S, Σ) with $\nu \ll \mu$ and let $h = \frac{\mathrm{d}\nu}{\mathrm{d}\mu}$, the standing assumptions in this exercise. Show that $\nu(\{h = 0\}) = 0$. Show that $\mu(\{h = 0\}) = 0$ iff $\mu \ll \nu$. What is $\frac{\mathrm{d}\mu}{\mathrm{d}\nu}$ if this happens?

5.6 Let μ and ν be σ-finite measures and ϕ a finite measure on (S, Σ). Assume that $\phi \ll \mu$ and $\nu \ll \mu$ with Radon-Nikodym derivatives h and k respectively. Let $\phi = \phi_a + \phi_s$ be the Lebesgue decomposition of ϕ w.r.t. μ. Show that (ν-a.e.)

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

5.7 Consider the measurable space $(\Omega, \mathcal{F})$ and a measurable map $X : \Omega \to \mathbb{R}^n$ ($\mathbb{R}^n$ is endowed with the usual Borel σ-algebra $\mathcal{B}^n$). Consider two probability measure $\mathbb{P}$ and $\mathbb{Q}$ on $(\Omega, \mathcal{F})$ and let $\mathbb{P}^X$ and $\mathbb{Q}^X$ be the corresponding distributions (laws) on $(\mathbb{R}^n, \mathcal{B}^n)$. Assume that $\mathbb{P}^X$ and $\mathbb{Q}^X$ are both absolutely continuous w.r.t. some σ-finite measure (e.g. Lebesgue measure), with corresponding Radon-Nikodym derivatives (in this context often called densities) f and g respectively, so $f, g : \mathbb{R}^n \to [0, \infty)$. Assume that $g > 0$. Show that for $\mathcal{F} = \sigma(X)$ it holds that $\mathbb{P} \ll \mathbb{Q}$ and that (look at Exercise 5.6) the Radon-Nikodym derivative here can be taken as the *likelihood ratio*

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

5.8 Show that the two displayed formulas in Exercise 4.10 are valid for functions F and G that are of bounded variation over some interval $(a, b]$. The integrals should be taken in the Lebesgue-Stieltjes sense.

5.9 Let a random variable X have distribution function F with the decomposition as in Proposition 5.7.

(a) Suppose that $F_s = 0$. Assume that $\mathbb{E}\, X$ is well defined. How would one compute this expectation practically? See also the introductory paragraph of Chapter 3 for an example where this occurs, and compute for that case $\mathbb{E}\, X$ explicitly. Verify the answer by exploiting the independence of Y and Z.

(b) As an example of the other extreme case, suppose that F is the distribution of Y as in Exercise 5.2, so $F = F_s$. What is $\mathbb{E}\, Y$ here?

6 Conditional expectation

6.1 A simple, finite case

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

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

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

$$\mathbb{E}\,\hat{X}\mathbf{1}_E = \mathbb{E}\,X\mathbf{1}_E, \forall E \in \sigma(Y). \tag{6.1}$$

The just described two properties of the conditional expectation will lie at the heart of a more general concept, *conditional expectation* of a random variable given a σ-algebra, see Section 6.2.

The random variable $\hat{X}$ is a.s. the only $\sigma(Y)$-measurable random variable that satisfies (6.1). 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.

6.2 Conditional expectation for $X \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $\mathcal{G}$ a sub-σ-algebra of $\mathcal{F}$. Assume that $X \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$. Inspired by the results of the previous section we adopt the following definition.

Definition 6.1 A random variable $\hat{X}$ is called a version of the conditional expectation $\mathbb{E}\,[X|\mathcal{G}]$, if it is $\mathcal{G}$-measurable and if

$$\mathbb{E}\, \hat{X}\mathbf{1}_G = \mathbb{E}\, X\mathbf{1}_G, \forall G \in \mathcal{G}. \tag{6.2}$$

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 6.2 *If $X \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$, then a version of the conditional expectation $\mathbb{E}\,[X|\mathcal{G}]$ exists and moreover, any two versions are a.s. equal.*

Proof For any $G \in \mathcal{G}$ we define $\nu^+(G) := \mathbb{E}\, X^+\mathbf{1}_G$ and $\nu^-(G) := \mathbb{E}\, X^-\mathbf{1}_G$ We have seen that ν^+ and ν^- are finite measures on the measurable space $(\Omega, \mathcal{G})$. Moreover, $\nu^+ \ll \mathbb{P}$ and $\nu^- \ll \mathbb{P}$ on this space. According to the Radon-Nikodym Theorem 5.4 there exist nonnegative $\mathcal{G}$-measurable functions ξ^+ and ξ^- such that $\nu^+(G) = \mathbb{E}\, \xi^+\mathbf{1}_G$ and $\nu^-(G) = \mathbb{E}\, \xi^-\mathbf{1}_G$. These functions are a.s. unique. Then $\hat{X} = \xi^+ - \xi^-$ is a version of $\mathbb{E}\,[X|\mathcal{G}]$. □

Remark 6.3 The ξ^+ and ξ^- in the above proof are in general not equal to the positive and negative parts $\hat{X}^+$ and $\hat{X}^-$ of $\hat{X}$. Think of a simple example.

Remark 6.4 It is common to call a given version of $\mathbb{E}\,[X|\mathcal{G}]$ *the* conditional expectation of X given $\mathcal{G}$, but one should take care with this custom. In fact one should consider $\mathbb{E}\,[X|\mathcal{G}]$ as an equivalence class of random variables, where equivalence $Y_1 \sim Y_2$ for $\mathcal{G}$-measurable functions means that $\mathbb{P}(Y_1 = Y_2) = 1$. As such one can consider $\mathbb{E}\,[X|\mathcal{G}]$ as an element of $L^1(\Omega, \mathcal{G}, \mathbb{P})$. Later on we will often identify a version $\hat{X}$ of $\mathbb{E}\,[X|\mathcal{G}]$ with $\mathbb{E}\,[X|\mathcal{G}]$.

Remark 6.5 One can also define versions of conditional expectations for random variables X with $\mathbb{P}(X \in [0, \infty]) = 1$ without requiring that $\mathbb{E}\, X < \infty$. Again this follows from the Radon-Nikodym theorem. The definition of conditional expectation can also be extended to e.g. the case where $X = X^+ - X^-$, where $\mathbb{E}\, X^- < \infty$, but $\mathbb{E}\, X^+ = \infty$.

Let us present the most relevant properties of conditional expectation. As before, we let $X \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$, $\mathcal{G}$ a sub-σ-algebra of $\mathcal{F}$ and $\hat{X}$ is a version of $\mathbb{E}\,[X|\mathcal{G}]$. Other random variables below that are versions of a conditional expectation given $\mathcal{G}$ are similarly denoted with a 'hat'.

Proposition 6.6 *The following elementary properties hold.*

(i) If $X \geq 0$ a.s., then $\hat{X} \geq 0$ a.s. If $X \geq Y$ a.s., then $\hat{X} \geq \hat{Y}$ a.s.

(ii) $\mathbb{E}\,\hat{X} = \mathbb{E}\,X$.

(iii) If $a, b \in \mathbb{R}$ and if $\hat{X}$ and $\hat{Y}$ are versions of $\mathbb{E}\,[X|\mathcal{G}]$ and $\mathbb{E}\,[Y|\mathcal{G}]$, then $a\hat{X} + b\hat{Y}$ is a version of $\mathbb{E}\,[aX + bY|\mathcal{G}]$.

(iv) If X is $\mathcal{G}$-measurable, then X is a version of $\mathbb{E}\,[X|\mathcal{G}]$.

Proof (i) Let $G = \{\hat{X} < 0\}$. Then we have from (6.2) 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 (6.2).

(iii) Just verify that $\mathbb{E}\,\mathbf{1}_G(a\hat{X} + b\hat{Y}) = \mathbb{E}\,\mathbf{1}_G(aX + bY)$, for all $G \in \mathcal{G}$.

(iv) Obvious. □

We have taken some care in formulating the assertions of the previous theorem concerning versions. Bearing this in mind and being a bit less precise at the same time, one often phrases e.g. (iii) as $\mathbb{E}\,[aX + bY|\mathcal{G}] = a\mathbb{E}\,[X|\mathcal{G}] + b\mathbb{E}\,[Y|\mathcal{G}]$. Some convergence properties are listed in the following theorem.

Theorem 6.7 *The following convergence properties for conditional expectation given a fixed sub-σ-algebra hold.*

(i) If (X_n) is an a.s. increasing sequence of nonnegative random variables, then the same holds for versions $(\hat{X}_n)$. If moreover $X_n \uparrow X$ a.s., then $\hat{X}_n \uparrow \hat{X}$ a.s. (monotone convergence for conditional expectations)

(ii) If (X_n) is a sequence of a.s. nonnegative random variables, and $(\hat{X}_n)$ are corresponding versions of the conditional expectations, then $\liminf_{n\to\infty} \hat{X}_n \geq \hat{\ell}$ a.s., where $\hat{\ell}$ is a version of the conditional expectation of $\ell := \liminf_{n\to\infty} X_n$. (Fatou's lemma for conditional expectations)

(iii) If (X_n) is a sequence of random variables such that for some X one has $X_n \to X$ a.s. and if there is a random variable Y such that $\mathbb{E}\,Y < \infty$ and $|X_n| \leq Y$ a.s. for all n. Then $\hat{X}_n \to \hat{X}$ a.s. (dominated convergence for conditional expectations)

Proof (i) From the previous theorem we know that the $\hat{X}_n$ form a.s. an increasing sequence. Let $\hat{X} := \limsup \hat{X}_n$, then $\hat{X}$ is $\mathcal{G}$-measurable and $\hat{X}_n \uparrow \hat{X}$ a.s. We verify that this $\hat{X}$ is a version of $\mathbb{E}\,[X|\mathcal{G}]$. But this follows by application of the Monotone Convergence Theorem to both sides of $\mathbb{E}\,\mathbf{1}_G X_n = \mathbb{E}\,\mathbf{1}_G\hat{X}_n$ for all $G \in \mathcal{G}$.

(ii) and (iii) These properties follow by mimicking the proofs of the ordinary versions of Fatou's Lemma and the Dominated Convergence Theorem, Exercises 6.4 and 6.5. □

Theorem 6.8 *Additional properties of conditional expectations are as follows.*

(i) If $\mathcal{H}$ is a sub-σ-algebra of $\mathcal{G}$, then any version of $\mathbb{E}\,[\hat{X}|\mathcal{H}]$ is also a version of $\mathbb{E}\,[X|\mathcal{H}]$ and vice versa (tower property).

(ii) *If Z is $\mathcal{G}$-measurable such that $ZX \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$, then $Z\hat{X}$ is a version of $\mathbb{E}[ZX|\mathcal{G}]$. We write $Z\mathbb{E}[X|\mathcal{G}] = \mathbb{E}[ZX|\mathcal{G}]$.*

(iii) *Let $\hat{X}$ be a version of $\mathbb{E}[X|\mathcal{G}]$. If $\mathcal{H}$ is independent of $\sigma(X) \vee \mathcal{G}$, then $\hat{X}$ is a version of $\mathbb{E}[X|\mathcal{G} \vee \mathcal{H}]$. In particular, $\mathbb{E}\,X$ is a version of $\mathbb{E}[X|\mathcal{H}]$ if $\sigma(X)$ and $\mathcal{H}$ are independent.*

(iv) *Let X be a $\mathcal{G}$-measurable random variable and let the random variable Y be independent of $\mathcal{G}$. Assume that $h \in \mathcal{B}(\mathbb{R}^2)$ is such that $h(X, Y) \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$. Put $\hat{h}(x) = \mathbb{E}[h(x, Y)]$. Then $\hat{h}$ is a Borel function and $\hat{h}(X)$ is a version of $\mathbb{E}[h(X, Y)|\mathcal{G}]$.*

(v) *If $c : \mathbb{R} \to \mathbb{R}$ is a convex function and $\mathbb{E}\,|c(X)| < \infty$, then $c(\hat{X}) \leq C$, a.s., where C is any version of $\mathbb{E}[c(X)|\mathcal{G}]$. We often write $c(\mathbb{E}[X|\mathcal{G}]) \leq \mathbb{E}[c(X)|\mathcal{G}]$ (Jensen's inequality for conditional expectations).*

(vi) *$||\hat{X}||_p \leq ||X||_p$, for every $p \geq 1$.*

Proof (i) Let $\tilde{X}$ be a version of $\mathbb{E}[\hat{X}|\mathcal{H}]$. By definition, we have $\mathbb{E}\,\mathbf{1}_H\tilde{X} = \mathbb{E}\,\mathbf{1}_H\hat{X}$, for all $H \in \mathcal{H}$. But since $\mathcal{H} \subset \mathcal{G}$, it also holds that $\mathbb{E}\,\mathbf{1}_H\hat{X} = \mathbb{E}\,\mathbf{1}_H X$, by (6.2). Hence $\tilde{X}$ is a version of $\mathbb{E}[X|\mathcal{H}]$.

(ii) We give the proof for bounded Z. Certainly ZX is integrable in this case and its conditional expectation exists. Without loss of generality we may then even assume that $Z \geq 0$ a.s. (Add a constant c to Z to have $Z + c \geq 0$, if this is not the case and the result will follow from the case of nonnegative Z). Assume first that also X is nonnegative. If $Z = \mathbf{1}_G$ for some $G \in \mathcal{G}$, then the result directly follows from the definition. By linearity the assertion holds for nonnegative simple Z. For arbitrary $Z \geq 0$, we choose simple Z_n such that $Z_n \uparrow Z$. Since we know (in the sense of versions) $Z_n\hat{X} = \mathbb{E}[Z_nX|\mathcal{G}]$, we apply Theorem 6.7 (i)) to settle the case for $X \geq 0$. If X is arbitrary, linearity yields the assertion by applying the previous results for X^+ and X^-.

(iii) It is sufficient to show this for nonnegative X. Let $G \in \mathcal{G}$ and $H \in \mathcal{H}$. By the independence assumption, we have $\mathbb{E}\,\mathbf{1}_G\mathbf{1}_H X = \mathbb{E}\,\mathbf{1}_G X\,\mathbb{P}(H)$ and $\mathbb{E}\,\mathbf{1}_G\mathbf{1}_H\hat{X} = \mathbb{E}\,\mathbf{1}_G\hat{X}\,\mathbb{P}(H)$. It follows that $\mathbb{E}\,\mathbf{1}_G\mathbf{1}_H X = \mathbb{E}\,\mathbf{1}_G\mathbf{1}_H\hat{X}$, since $\hat{X}$ is version of $\mathbb{E}[X|\mathcal{G}]$. Recall from Exercise 1.6 that the collection $\mathcal{C} := \{G \cap H : G \in \mathcal{G}, H \in \mathcal{H}\}$ is a π-system that generates $\mathcal{G} \vee \mathcal{H}$. Observe that $E \mapsto \mathbb{E}\,\mathbf{1}_E X$ and $E \mapsto \mathbb{E}\,\mathbf{1}_E\hat{X}$ both define measures on $\mathcal{G} \vee \mathcal{H}$ and that these measures have been seen to coincide on $\mathcal{C}$. It follows from Theorem 1.15 that these measures are the same. The second statement follows by taking $\mathcal{G} = \{\emptyset, \Omega\}$.

(iv) We use the Monotone Class Theorem, Theorem 2.6 and for simplicity of notation we take X and Y real valued. Let V be the collection of all bounded measurable functions for which the statement holds true. Using (iii), one easily checks that $h = \mathbf{1}_{B\times C} \in V$, where B, C are Borel sets in $\mathbb{R}$. The sets $B \times C$ form a π-system that generates $\mathcal{B}(\mathbb{R}^2)$. The collection V is obviously a vector space and the constant functions belong to it. Let (h_n) be an increasing sequence of nonnegative functions in V that converge to some bounded function h. If $\hat{h}_n(x) = \mathbb{E}\,h_n(x, Y)$ and $\hat{h}(x) = \mathbb{E}\,h(x, Y)$, then we also have $\hat{h}_n(x) \uparrow \hat{h}(x)$ for all x by the Monotone Convergence Theorem. We will see that $\hat{h}(X)$ is a version of $\mathbb{E}[h(X, Y)|\mathcal{G}]$. Let $G \in \mathcal{G}$. For all n it holds that $\mathbb{E}\,\mathbf{1}_G\hat{h}_n(X) = \mathbb{E}\,\mathbf{1}_G h_n(X, Y)$.

Invoking the Monotone Convergence Theorem again results in $\mathbb{E}\,\mathbf{1}_G\hat{h}(X) = \mathbb{E}\,\mathbf{1}_G h(X,Y)$. Since all $\hat{h}_n(X)$ are $\mathcal{G}$-measurable, the same holds for $\hat{h}(X)$ and we conclude that $h \in V$. The remainder of the proof is Exercise 6.11.

(v) Since c is convex, there are sequences (a_n) and (b_n) in $\mathbb{R}$ such that $c(x) = \sup\{a_n x + b_n : n \in \mathbb{N}\}$, $\forall x \in \mathbb{R}$. Hence for all n we have $c(X) \geq a_n X + b_n$ and by the monotonicity property of conditional expectation, we also have $C \geq a_n \hat{X} + b_n$ a.s. If N_n is the set of probability zero, where this inequality is violated, then also $\mathbb{P}(N) = 0$, where $N = \cup_{n=1}^\infty N_n$. Outside N we have $C \geq \sup_n (a_n \hat{X} + b_n) = c(\hat{X})$.

(vi) The statement concerning the p-norms follows upon choosing $c(x) = |x|^p$ in (v) and taking expectations. □

Let $P : L^1(\Omega, \mathcal{F}, \mathbb{P}) \to L^1(\Omega, \mathcal{G}, \mathbb{P})$ be the linear map that transforms X into $\mathbb{E}\,[X|\mathcal{G}]$. If $\hat{X}$ is a version of $\mathbb{E}\,[X|\mathcal{G}]$, then it is also a version of $\mathbb{E}\,[\hat{X}|\mathcal{G}]$. So, we get $P^2 = P$, meaning that P is a *projection*. In the next proposition we give this a geometric interpretation in a slightly narrower context.

Proposition 6.9 *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 can thus be viewed as orthogonal projections onto $\mathcal{L}^2(\Omega, \mathcal{G}, \mathbb{P})$.

Proof Exercise 6.3. □

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.

6.3 Conditional probabilities

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

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

Likewise, one can define conditional distributions of a random variable X. For a Borel set B one defines $\mathbb{P}^X(B|\mathcal{G}) := \mathbb{P}(X^{-1}[B]|\mathcal{G})$.

Of course all versions of $\mathbb{P}(F|\mathcal{G})$ are almost surely equal. Moreover, if $F_1, F_2, \ldots$ is a sequence of disjoint events, and $\hat{\mathbb{P}}(F_n)$ are versions of the conditional probabilities, then one easily shows that $\sum_{n=1}^\infty \hat{\mathbb{P}}(F_n)$ is a version of the conditional probability $\mathbb{P}(\cup_{n=1}^{\infty} F_n|\mathcal{G})$. So, if $\hat{\mathbb{P}}(\cup_{n=1}^{\infty} F_n)$ is any version of $\mathbb{P}(\cup_{n=1}^{\infty} F_n|\mathcal{G})$, then outside a set N of probability zero, we have

$$\hat{\mathbb{P}}(\cup_{n=1}^{\infty} F_n) = \sum_{n=1}^{\infty} \hat{\mathbb{P}}(F_n). \tag{6.3}$$

A problem is that the set N in general depends on the sequence of events $F_1, F_2, \ldots$ Since there are usually uncountably many of such sequences, it is not clear (and in fact not always true!) that there is one (fixed) set of probability zero such that outside this set for *all* disjoint sequences (F_n) the equality (6.3) 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 (6.3) is assumed to hold outside a set N of probability zero. Of course, if $N = \emptyset$, this relation holds everywhere. But also of $N \neq \emptyset$, this relation can be turned into one that is everywhere true. Suppose that $N \neq \emptyset$. Redefine $\hat{\mathbb{P}}$ by taking $\hat{\mathbb{P}}(F)(\omega)$ as given on N^c, but for all $\omega \in N$ we take instead $\hat{\mathbb{P}}(\cdot)(\omega)$ as any fixed probability measure on $\mathcal{F}$ (for instance a Dirac measure). Since we change the map $\hat{\mathbb{P}}(F)$ on the null set N only, we keep on having a conditional probability of F, whereas (6.3) 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 6.10 *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 $\mathbb{P}(X^{-1}[B]|\mathcal{G})$ for all $B \in \mathcal{B}$.*

Proof We split the proof into two parts. First we show the existence of a conditional distribution function, after which we show that it generates a regular conditional distribution of X given $\mathcal{G}$.

We will construct a conditional distribution function on the rational numbers. For each $q \in \mathbb{Q}$ we select a version of $\mathbb{P}(X \leq q|\mathcal{G})$, call it $G(q)$. Let $E_{rq} = \{G(r) < G(q)\}$. Assume that $r > q$. Then $\{X \leq r\} \supset \{X \leq q\}$ and hence $G(r) \geq G(q)$ a.s. and so $\mathbb{P}(E_{rq}) = 0$. Hence we obtain that $\mathbb{P}(E) = 0$, where

$E = \cup_{r>q} E_{rq}$. Note that E is the set where the random variables $G(q)$ fail to be increasing in the argument q. Let $F_q = \{\inf_{r>q} G(r) > G(q)\}$. Let $\{q_1, q_2, \ldots\}$ be the set of rationals strictly bigger then q and let $r_n = \inf\{q_1, \ldots, q_n\}$. Then $r_n \downarrow q$, as $n \to \infty$. Since the indicators $\mathbf{1}_{\{X \leq r_n\}}$ are bounded, we have $G(r_n) \downarrow G(q)$ a.s. It follows that $\mathbb{P}(F_q) = 0$, and then $\mathbb{P}(F) = 0$, where $F = \cup_{q \in \mathbb{Q}} F_q$. Note that F is the event on which $G(\cdot)$ is not right-continuous. Let then H be the set on which $\lim_{q\to\infty} G(q) < 1$ or $\lim_{q\to-\infty} G(q) > 0$. By a similar argument, we have $\mathbb{P}(H) = 0$. On the set $\Omega_0 := (E \cup F \cup H)^c$, the random function $G(\cdot)$ has the properties of a distribution function on the rationals. Note that $\Omega_0 \in \mathcal{G}$. Let F^0 be an arbitrary distribution function and define for $x \in \mathbb{R}$

$$\hat{F}(x) = \mathbf{1}_{\Omega_0^c} F^0(x) + \mathbf{1}_{\Omega_0} \inf_{q>x} G(q).$$

It is easy to check that $\hat{F}(\cdot)$ is a distribution function for each hidden argument ω. Moreover, $\hat{F}(x)$ is $\mathcal{G}$-measurable and since $\inf_{q>x} \mathbf{1}_{\{X \leq q\}} = \mathbf{1}_{\{X \leq x\}}$, we obtain that $\hat{F}(x)$ is a version of $\mathbb{P}(X \leq x|\mathcal{G})$. This finishes the proof of the construction of a conditional distribution function of X given $\mathcal{G}$.

For every ω, the distribution function $\hat{F}(\cdot)(\omega)$ generates a probability measure $\mathbb{P}^X(\cdot)(\omega)$ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Let $\mathcal{C}$ be the class of Borel-measurable sets B for which $\mathbb{P}^X(B)$ is a version of $\mathbb{P}(X \in B|\mathcal{G})$. It follows that all intervals $(-\infty, x]$ belong to $\mathcal{C}$. Moreover, $\mathcal{C}$ is a d-system. By virtue of Dynkin's Lemma 1.13, $\mathcal{C} = \mathcal{B}(\mathbb{R})$. □

Proposition 6.11 *Let X be a random variable and $h : \mathbb{R} \to \mathbb{R}$ be a Borel-measurable function. Let $\hat{\mathbb{P}}^X$ be a regular conditional distribution of X given $\mathcal{G}$. If $h(X) \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$, then*

$$\int h(x)\, \hat{\mathbb{P}}^X(\mathrm{d}x) \tag{6.4}$$

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 (6.4) is a version of $\mathbb{E}\,[h(X)|\mathcal{G}]$. Clearly, in view of Theorem 6.10 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 convergence to h in a monotone way. Monotone convergence for conditional expectations yields $h \in \mathcal{H}$. If h is arbitrary, we split as usual $h = h^+ - h^-$ and apply the previous step. □

Once more we emphasize that regular conditional probabilities in general don't exist. The general definition of conditional expectation would be pointless if every conditional expectation could be computed by Proposition 6.11. The good news is that in most common situations Proposition 6.11 can be applied.

In Exercise 6.8 you find an explicit expression for the regular conditional distribution of a random variable X given another random variable Y.

6.4 Exercises

6.1 Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $\mathcal{A} = \{A_1, \ldots, A_n\}$ be a partition of Ω, where the A_i belong to $\mathcal{F}$. Let $X \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ and $\mathcal{G} = \sigma(\mathcal{A})$. Show that any version of $\mathbb{E}\,[X|\mathcal{G}]$ is of the form $\sum_{i=1}^n a_i \mathbf{1}_{A_i}$ and determine the a_i.

6.2 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.

6.3 Prove Proposition 6.9.

6.4 Prove the conditional version of Fatou's lemma, Theorem 6.7(ii).

6.5 Prove the conditional Dominated Convergence theorem, Theorem 6.7(iii).

6.6 Let (X, Y) have a bivariate normal distribution with $\mathbb{E}\,X = \mu_X$, $\mathbb{E}\,Y = \mu_Y$, $\operatorname{Var} X = \sigma_X^2$, $\operatorname{Var} Y = \sigma_Y^2$ and $\operatorname{Cov}(X, Y) = c$. Let

$$\hat{X} = \mu_x + \frac{c}{\sigma_Y^2}(Y - \mu_Y).$$

Show that $\mathbb{E}\,(X - \hat{X})Y = 0$. Show also (use a special property of the bivariate normal distribution) that $\mathbb{E}\,(X - \hat{X})g(Y) = 0$ if g is a Borel-measurable function such that $\mathbb{E}\,g(Y)^2 < \infty$. Conclude that $\hat{X}$ is a version of $\mathbb{E}\,[X|Y]$.

6.7 Let $X, Y \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ and assume that $\mathbb{E}\,[X|Y] = Y$ and $\mathbb{E}\,[Y|X] = X$ (or rather, versions of them are a.s. equal). Show that $\mathbb{P}(X = Y) = 1$. *Hint:* Start to work on $\mathbb{E}\,(X - Y)\mathbf{1}_{\{X>z, Y\leq z\}} + \mathbb{E}\,(X - Y)\mathbf{1}_{\{X\leq z, Y\leq z\}}$ for arbitrary $z \in \mathbb{R}$.

6.8 Let X and Y be random variables and assume that (X, Y) admits a density f w.r.t. Lebesgue measure on $(\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2))$. Let f_Y be the marginal density of Y. Define $\hat{f}(x|y)$ by

$$\hat{f}(x|y) = \begin{cases} \frac{f(x,y)}{f_Y(y)} & \text{if } f_Y(y) > 0 \\ 0 & \text{else.} \end{cases}$$

Assume that $\mathbb{E}\,|h(X)| < \infty$. Put $\hat{h}(y) = \int_{\mathbb{R}} h(x)\hat{f}(x|y)\,\mathrm{d}x$. Show that $\hat{h}(Y)$ is a version of $\mathbb{E}\,[h(X)|Y]$. Show also that

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

defines a regular conditional probability on $\mathcal{B}(\mathbb{R})$ given Y. What is the exceptional set N of Section 6.3?

6.9 Consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Two random variables X and Y are called conditionally independent given a sub-σ-algebra $\mathcal{G}$ of $\mathcal{F}$ if for all bounded Borel functions $f, g : \mathbb{R} \to \mathbb{R}$ it holds that $\mathbb{E}\,[f(X)g(Y)|\mathcal{G}] = \mathbb{E}\,[f(X)|\mathcal{G}]\mathbb{E}\,[g(Y)|\mathcal{G}]$.

(a) Show that X and Y are conditionally independent given $\mathcal{G}$ iff for every bounded measurable function $f : \mathbb{R} \to \mathbb{R}$ it holds that $\mathbb{E}\,[f(X)|\sigma(Y) \vee \mathcal{G}] = \mathbb{E}\,[f(X)|\mathcal{G}]$.

(b) Show (by examples) that in general conditional independence is not implied by independence, nor vice versa.

(c) If X and Y are given random variables, give an example of a σ-algebra $\mathcal{G}$ that makes X and Y conditionally independent.

(d) Propose a definition of conditional independence of two σ-algebras $\mathcal{H}_1$ and $\mathcal{H}_2$ given $\mathcal{G}$ that is such that conditional independence of X and Y given $\mathcal{G}$ can be derived from it as a special case.

6.10 (*Hölder's inequality for conditional expectations*) Let $X \in \mathcal{L}^p(\Omega, \mathcal{F}, \mathbb{P})$ and $Y \in \mathcal{L}^q(\Omega, \mathcal{F}, \mathbb{P})$, where $p, q \in [1, \infty]$, $\frac{1}{p} + \frac{1}{q} = 1$. Then

$$\mathbb{E}\,[|XY||\mathcal{G}] \leq (\mathbb{E}\,[|X|^p|\mathcal{G}])^{1/p}(\mathbb{E}\,[|Y|^q|\mathcal{G}])^{1/q}. \tag{6.5}$$

It is sufficient to show this for nonnegative X and Y, so assume $X, Y \geq 0$ a.s.

(a) Let $U = \mathbb{E}\,[X^p|\mathcal{G}]$ and $V = \mathbb{E}\,[Y^q|\mathcal{G}]$ and $H = \{U, V > 0\}$. Suppose that

$$\mathbf{1}_H \mathbb{E}\,[XY|\mathcal{G}] \leq \mathbf{1}_H (\mathbb{E}\,[X^p|\mathcal{G}])^{1/p}(\mathbb{E}\,[Y^q|\mathcal{G}])^{1/q}]. \tag{6.6}$$

Show that Hölder's inequality (6.5) follows from (6.6).

(b) Show that

$$\mathbb{E}\,[\mathbf{1}_G \mathbf{1}_H \frac{\mathbb{E}\,[XY|\mathcal{G}]}{UV} \leq \mathbb{E}\,[\mathbf{1}_G \mathbf{1}_H]$$

holds for all $G \in \mathcal{G}$ and deduce (6.6).

6.11 Finish the proof of Theorem 6.8 (iv), i.e. show that the assertion also holds without the boundedness condition on h.

Index