Probability Theory: STAT310/MATH230; September 12, 2010

Amir Dembo

E-mail address: amir@math.stanford.edu

Department of Mathematics, Stanford University, Stanford, CA 94305.

Contents

Preface

These are the lecture notes for a year long, PhD level course in Probability Theory that I taught at Stanford University in 2004, 2006 and 2009. The goal of this course is to prepare incoming PhD students in Stanford's mathematics and statistics departments to do research in probability theory. More broadly, the goal of the text is to help the reader master the mathematical foundations of probability theory and the techniques most commonly used in proving theorems in this area. This is then applied to the rigorous study of the most fundamental classes of stochastic processes.

Towards this goal, we introduce in Chapter 1 the relevant elements from measure and integration theory, namely, the probability space and the σ-algebras of events in it, random variables viewed as measurable functions, their expectation as the corresponding Lebesgue integral, and the important concept of independence.

Utilizing these elements, we study in Chapter 2 the various notions of convergence of random variables and derive the weak and strong laws of large numbers.

Chapter 3 is devoted to the theory of weak convergence, the related concepts of distribution and characteristic functions and two important special cases: the Central Limit Theorem (in short CLT) and the Poisson approximation.

Drawing upon the framework of Chapter 1, we devote Chapter 4 to the definition, existence and properties of the conditional expectation and the associated regular conditional probability distribution.

Chapter 5 deals with filtrations, the mathematical notion of information progression in time, and with the corresponding stopping times. Results about the latter are obtained as a by product of the study of a collection of stochastic processes called martingales. Martingale representations are explored, as well as maximal inequalities, convergence theorems and various applications thereof. Aiming for a clearer and easier presentation, we focus here on the discrete time settings deferring the continuous time counterpart to Chapter 8.

Chapter 6 provides a brief introduction to the theory of Markov chains, a vast subject at the core of probability theory, to which many text books are devoted. We illustrate some of the interesting mathematical properties of such processes by examining few special cases of interest.

Chapter 7 sets the framework for studying right-continuous stochastic processes indexed by a continuous time parameter, introduces the family of Gaussian processes and rigorously constructs the Brownian motion as a Gaussian process of continuous sample path and zero-mean, stationary independent increments.

Chapter 8 expands our earlier treatment of martingales and strong Markov processes to the continuous time setting, emphasizing the role of right-continuous filtration. The mathematical structure of such processes is then illustrated both in the context of Brownian motion and that of Markov jump processes.

Building on this, in Chapter 9 we re-construct the Brownian motion via the invariance principle as the limit of certain rescaled random walks. We further delve into the rich properties of its sample path and the many applications of Brownian motion to the CLT and the Law of the Iterated Logarithm (in short, LIL).

The intended audience for this course should have prior exposure to stochastic processes, at an informal level. While students are assumed to have taken a real analysis class dealing with Riemann integration, and mastered well this material, prior knowledge of measure theory is not assumed.

It is quite clear that these notes are much influenced by the text books [**Bil95, Dur03, Wil91, KaS97**] I have been using.

I thank my students out of whose work this text materialized and my teaching assistants Su Chen, Kshitij Khare, Guoqiang Hu, Julia Salzman, Kevin Sun and Hua Zhou for their help in the assembly of the notes of more than eighty students into a coherent document. I am also much indebted to Kevin Ross, Andrea Montanari and Oana Mocioalca for their feedback on earlier drafts of these notes, to Kevin Ross for providing all the figures in this text, and to Andrea Montanari, David Siegmund and Tze Lai for contributing some of the exercises in these notes.

AMIR DEMBO

STANFORD, CALIFORNIA
APRIL 2010

CHAPTER 1

Probability, measure and integration

This chapter is devoted to the mathematical foundations of probability theory. Section 1.1 introduces the basic measure theory framework, namely, the probability space and the σ-algebras of events in it. The next building blocks are random variables, introduced in Section 1.2 as measurable functions $\omega \mapsto X(\omega)$ and their distribution.

This allows us to define in Section 1.3 the important concept of expectation as the corresponding Lebesgue integral, extending the horizon of our discussion beyond the special functions and variables with density to which elementary probability theory is limited. Section 1.4 concludes the chapter by considering independence, the most fundamental aspect that differentiates probability from (general) measure theory, and the associated product measures.

1.1. Probability spaces, measures and σ-algebras

We shall define here the probability space $(\Omega, \mathcal{F}, \mathbf{P})$ using the terminology of measure theory.

The *sample space* Ω is a set of all possible outcomes $\omega \in \Omega$ of some random experiment. Probabilities are assigned by $A \mapsto \mathbf{P}(A)$ to A in a subset $\mathcal{F}$ of all possible sets of outcomes. The *event space* $\mathcal{F}$ represents both the amount of information available as a result of the experiment conducted and the collection of all events of possible interest to us. A pleasant mathematical framework results by imposing on $\mathcal{F}$ the structural conditions of a σ-algebra, as done in Subsection 1.1.1. The most common and useful choices for this σ-algebra are then explored in Subsection 1.1.2. Subsection 1.1.3 provides fundamental supplements from measure theory, namely Dynkin's and Carathéodory's theorems and their application to the construction of Lebesgue measure.

1.1.1. The probability space $(\Omega, \mathcal{F}, \mathbf{P})$. We use 2^Ω to denote the set of all possible subsets of Ω. The event space is thus a subset $\mathcal{F}$ of 2^Ω, consisting of all allowed events, that is, those events to which we shall assign probabilities. We next define the structural conditions imposed on $\mathcal{F}$.

Definition 1.1.1. *We say that $\mathcal{F} \subseteq 2^\Omega$ is a* σ-algebra *(or a* σ-field*), if*
(a) $\Omega \in \mathcal{F}$,
(b) If $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$ as well (where $A^c = \Omega \setminus A$).
(c) If $A_i \in \mathcal{F}$ for $i = 1, 2, 3, \ldots$ then also $\bigcup_i A_i \in \mathcal{F}$.

Remark. Using *DeMorgan's law*, we know that $(\bigcup_i A_i^c)^c = \bigcap_i A_i$. Thus the following is equivalent to property (c) of Definition 1.1.1:
(c') If $A_i \in \mathcal{F}$ for $i = 1, 2, 3, \ldots$ then also $\bigcap_i A_i \in \mathcal{F}$.

Definition 1.1.2. *A pair* $(\Omega, \mathcal{F})$ *with* $\mathcal{F}$ *a* σ*-algebra of subsets of* Ω *is called a* measurable space. *Given a measurable space* $(\Omega, \mathcal{F})$*, a* measure μ *is any* countably additive *non-negative* set function *on this space. That is,* $\mu : \mathcal{F} \to [0, \infty]$*, having the properties:*
(a) $\mu(A) \geq \mu(\emptyset) = 0$ *for all* $A \in \mathcal{F}$*.*
(b) $\mu(\bigcup_n A_n) = \sum_n \mu(A_n)$ *for any countable collection of disjoint sets* $A_n \in \mathcal{F}$*.*
When in addition $\mu(\Omega) = 1$*, we call the measure* μ *a* probability measure*, and often label it by* $\mathbf{P}$ *(it is also easy to see that then* $\mathbf{P}(A) \leq 1$ *for all* $A \in \mathcal{F}$*).*

Remark. When (b) of Definition 1.1.2 is relaxed to involve only finite collections of disjoint sets A_n, we say that μ is a *finitely additive* non-negative set-function. In measure theory we sometimes consider *signed measures*, whereby μ is no longer non-negative, hence its range is $[-\infty, \infty]$, and say that such measure is *finite* when its range is $\mathbb{R}$ (i.e. no set in $\mathcal{F}$ is assigned an infinite measure).

Definition 1.1.3. *A* measure space *is a triplet* $(\Omega, \mathcal{F}, \mu)$*, with* μ *a measure on the measurable space* $(\Omega, \mathcal{F})$*. A measure space* $(\Omega, \mathcal{F}, \mathbf{P})$ *with* $\mathbf{P}$ *a probability measure is called a* probability space.

The next exercise collects some of the fundamental properties shared by all probability measures.

Exercise 1.1.4. *Let* $(\Omega, \mathcal{F}, \mathbf{P})$ *be a probability space and* A, B, A_i *events in* $\mathcal{F}$*. Prove the following properties of every probability measure.*

(a) Monotonicity. *If* $A \subseteq B$ *then* $\mathbf{P}(A) \leq \mathbf{P}(B)$*.*
(b) Sub-additivity. *If* $A \subseteq \cup_i A_i$ *then* $\mathbf{P}(A) \leq \sum_i \mathbf{P}(A_i)$*.*
(c) Continuity from below: *If* $A_i \uparrow A$*, that is,* $A_1 \subseteq A_2 \subseteq \ldots$ *and* $\cup_i A_i = A$*, then* $\mathbf{P}(A_i) \uparrow \mathbf{P}(A)$*.*
(d) Continuity from above: *If* $A_i \downarrow A$*, that is,* $A_1 \supseteq A_2 \supseteq \ldots$ *and* $\cap_i A_i = A$*, then* $\mathbf{P}(A_i) \downarrow \mathbf{P}(A)$*.*

Remark. In the more general context of measure theory, note that properties (a)-(c) of Exercise 1.1.4 hold for any measure μ, whereas the continuity from above holds whenever $\mu(A_i) < \infty$ for all i sufficiently large. Here is more on this:

Exercise 1.1.5. *Prove that a finitely additive non-negative set function* μ *on a measurable space* $(\Omega, \mathcal{F})$ *with the "continuity" property*

$$B_n \in \mathcal{F}, \quad B_n \downarrow \emptyset, \quad \mu(B_n) < \infty \qquad \Longrightarrow \quad \mu(B_n) \to 0$$

must be countably additive if $\mu(\Omega) < \infty$*. Give an example that it is not necessarily so when* $\mu(\Omega) = \infty$*.*

The σ-algebra $\mathcal{F}$ always contains at least the set Ω and its complement, the empty set $\emptyset$. Necessarily, $\mathbf{P}(\Omega) = 1$ and $\mathbf{P}(\emptyset) = 0$. So, if we take $\mathcal{F}_0 = \{\emptyset, \Omega\}$ as our σ-algebra, then we are left with no degrees of freedom in choice of $\mathbf{P}$. For this reason we call $\mathcal{F}_0$ the *trivial* σ*-algebra.* Fixing Ω, we may expect that the larger the σ-algebra we consider, the more freedom we have in choosing the probability measure. This indeed holds to some extent, that is, as long as we have no problem satisfying the requirements in the definition of a probability measure. A natural question is when should we expect the maximal possible σ-algebra $\mathcal{F} = 2^\Omega$ to be useful?

Example 1.1.6. *When the sample space* Ω *is countable we can and typically shall take* $\mathcal{F} = 2^\Omega$*. Indeed, in such situations we assign a probability* $p_\omega > 0$ *to each* $\omega \in \Omega$

making sure that $\sum_{\omega\in\Omega} p_\omega = 1$. *Then, it is easy to see that taking* $\mathbf{P}(A) = \sum_{\omega\in A} p_\omega$ *for any* $A \subseteq \Omega$ *results with a probability measure on* $(\Omega, 2^\Omega)$. *For instance, when* Ω *is finite, we can take* $p_\omega = \frac{1}{|\Omega|}$, *the uniform measure on* Ω, *whereby computing probabilities is the same as counting. Concrete examples are a single coin toss, for which we have* $\Omega_1 = \{\mathrm{H}, \mathrm{T}\}$ *(*$\omega = \mathrm{H}$ *if the coin lands on its head and* $\omega = \mathrm{T}$ *if it lands on its tail), and* $\mathcal{F}_1 = \{\emptyset, \Omega, \mathrm{H}, \mathrm{T}\}$, *or when we consider a finite number of coin tosses, say* n, *in which case* $\Omega_n = \{(\omega_1, \ldots, \omega_n) : \omega_i \in \{\mathrm{H}, \mathrm{T}\}, i = 1, \ldots, n\}$ *is the set of all possible* n*-tuples of coin tosses, while* $\mathcal{F}_n = 2^{\Omega_n}$ *is the collection of all possible sets of* n*-tuples of coin tosses. Another example pertains to the set of all non-negative integers* $\Omega = \{0, 1, 2, \ldots\}$ *and* $\mathcal{F} = 2^\Omega$, *where we get the Poisson probability measure of parameter* $\lambda > 0$ *when starting from* $p_k = \frac{\lambda^k}{k!} e^{-\lambda}$ *for* $k = 0, 1, 2, \ldots$.

When Ω is uncountable such a strategy as in Example 1.1.6 will no longer work. The problem is that if we take $p_\omega = \mathbf{P}(\{\omega\}) > 0$ for uncountably many values of ω, we shall end up with $\mathbf{P}(\Omega) = \infty$. Of course we may define everything as before on a countable subset $\widehat{\Omega}$ of Ω and demand that $\mathbf{P}(A) = \mathbf{P}(A \cap \widehat{\Omega})$ for each $A \subseteq \Omega$. Excluding such trivial cases, to genuinely use an uncountable sample space Ω we need to restrict our σ-algebra $\mathcal{F}$ to a *strict subset* of 2^Ω.

Definition 1.1.7. *We say that a probability space* $(\Omega, \mathcal{F}, \mathbf{P})$ *is* non-atomic, *or alternatively call* $\mathbf{P}$ *non-atomic if* $\mathbf{P}(A) > 0$ *implies the existence of* $B \in \mathcal{F}$, $B \subset A$ *with* $0 < \mathbf{P}(B) < \mathbf{P}(A)$.

Indeed, in contrast to the case of countable Ω, the generic uncountable sample space results with a non-atomic probability space (c.f. Exercise 1.1.27). Here is an interesting property of such spaces (see also [**Bil95**, Problem 2.19]).

Exercise 1.1.8. *Suppose* $\mathbf{P}$ *is non-atomic and* $A \in \mathcal{F}$ *with* $\mathbf{P}(A) > 0$.

(a) *Show that for every* $\epsilon > 0$, *we have* $B \subseteq A$ *such that* $0 < \mathbf{P}(B) < \epsilon$.

(b) *Prove that if* $0 < a < \mathbf{P}(A)$ *then there exists* $B \subset A$ *with* $\mathbf{P}(B) = a$.

Hint: *Fix* $\epsilon_n \downarrow 0$ *and define inductively numbers* x_n *and sets* $G_n \in \mathcal{F}$ *with* $H_0 = \emptyset$, $H_n = \cup_{k<n} G_k$, $x_n = \sup\{\mathbf{P}(G) : G \subseteq A\backslash H_n, \mathbf{P}(H_n \cup G) \leq a\}$ *and* $G_n \subseteq A\backslash H_n$ *such that* $\mathbf{P}(H_n \bigcup G_n) \leq a$ *and* $\mathbf{P}(G_n) \geq (1 - \epsilon_n)x_n$. *Consider* $B = \cup_k G_k$.

As you show next, the collection of all measures on a given space is a *convex cone*.

Exercise 1.1.9. *Given any measures* $\{\mu_n, n \geq 1\}$ *on* $(\Omega, \mathcal{F})$, *verify that* $\mu = \sum_{n=1}^\infty c_n \mu_n$ *is also a measure on this space, for any finite constants* $c_n \geq 0$.

Here are few properties of probability measures for which the conclusions of Exercise 1.1.4 are useful.

Exercise 1.1.10. *A function* $d : \mathcal{X} \times \mathcal{X} \to [0, \infty)$ *is called a* semi-metric *on the set* $\mathcal{X}$ *if* $d(x, x) = 0$, $d(x, y) = d(y, x)$ *and the triangle inequality* $d(x, z) \leq d(x, y) + d(y, z)$ *holds. With* $A\Delta B = (A \cap B^c) \cup (A^c \cap B)$ *denoting the symmetric difference of subsets* A *and* B *of* Ω, *show that for any probability space* $(\Omega, \mathcal{F}, \mathbf{P})$, *the function* $d(A, B) = \mathbf{P}(A\Delta B)$ *is a semi-metric on* $\mathcal{F}$.

Exercise 1.1.11. *Consider events* $\{A_n\}$ *in a probability space* $(\Omega, \mathcal{F}, \mathbf{P})$ *that are almost disjoint in the sense that* $\mathbf{P}(A_n \cap A_m) = 0$ *for all* $n \neq m$. *Show that then* $\mathbf{P}(\cup_{n=1}^\infty A_n) = \sum_{n=1}^\infty \mathbf{P}(A_n)$.

Exercise 1.1.12. *Suppose a random outcome N follows the Poisson probability measure of parameter $\lambda > 0$. Find a simple expression for the probability that N is an even integer.*

1.1.2. Generated and Borel σ-algebras. Enumerating the sets in the σ-algebra $\mathcal{F}$ is not a realistic option for uncountable Ω. Instead, as we see next, the most common construction of σ-algebras is then by implicit means. That is, we demand that certain sets (called the *generators*) be in our σ-algebra, and take the smallest possible collection for which this holds.

Exercise 1.1.13.

(a) *Check that the intersection of (possibly uncountably many) σ-algebras is also a σ-algebra.*

(b) *Verify that for any σ-algebras $\mathcal{H} \subseteq \mathcal{G}$ and any $H \in \mathcal{H}$, the collection $\mathcal{H}^H = \{A \in \mathcal{G} : A \cap H \in \mathcal{H}\}$ is a σ-algebra.*

(c) *Show that $H \mapsto \mathcal{H}^H$ is non-increasing with respect to set inclusions, with $\mathcal{H}^\Omega = \mathcal{H}$ and $\mathcal{H}^\emptyset = \mathcal{G}$. Deduce that $\mathcal{H}^{H \cup H'} = \mathcal{H}^H \cap \mathcal{H}^{H'}$ for any pair $H, H' \in \mathcal{H}$.*

In view of part (a) of this exercise we have the following definition.

Definition 1.1.14. *Given a collection of subsets $A_\alpha \subseteq \Omega$ (not necessarily countable), we denote the smallest σ-algebra $\mathcal{F}$ such that $A_\alpha \in \mathcal{F}$ for all $\alpha \in \Gamma$ either by $\sigma(\{A_\alpha\})$ or by $\sigma(A_\alpha, \alpha \in \Gamma)$, and call $\sigma(\{A_\alpha\})$ the* σ-algebra generated *by the sets A_α. That is,*

$$\sigma(\{A_\alpha\}) = \bigcap \{\mathcal{G} : \ \mathcal{G} \subseteq 2^\Omega \textit{ is a } \sigma - \text{algebra}, \quad A_\alpha \in \mathcal{G} \quad \forall \alpha \in \Gamma\}.$$

Example 1.1.15. *Suppose $\Omega = \mathbb{S}$ is a topological space (that is, $\mathbb{S}$ is equipped with a notion of open subsets, or topology). An example of a generated σ-algebra is the* Borel σ-algebra *on $\mathbb{S}$ defined as $\sigma(\{O \subseteq \mathbb{S} \text{ open}\})$ and denoted by $\mathcal{B}_\mathbb{S}$. Of special importance is $\mathcal{B}_\mathbb{R}$ which we also denote by $\mathcal{B}$.*

Different sets of generators may result with the same σ-algebra. For example, taking $\Omega = \{1, 2, 3\}$ it is easy to see that $\sigma(\{1\}) = \sigma(\{2, 3\}) = \{\emptyset, \{1\}, \{2, 3\}, \{1, 2, 3\}\}$. A σ-algebra $\mathcal{F}$ is *countably generated* if there exists a countable collection of sets that generates it. Exercise 1.1.17 shows that $\mathcal{B}_\mathbb{R}$ is countably generated, but as you show next, there exist non countably generated σ-algebras even on $\Omega = \mathbb{R}$.

Exercise 1.1.16. *Let $\mathcal{F}$ consist of all $A \subseteq \Omega$ such that either A is a countable set or A^c is a countable set.*

(a) *Verify that $\mathcal{F}$ is a σ-algebra.*

(b) *Show that $\mathcal{F}$ is countably generated if and only if Ω is a countable set.*

Recall that if a collection of sets $\mathcal{A}$ is a subset of a σ-algebra $\mathcal{G}$, then also $\sigma(\mathcal{A}) \subseteq \mathcal{G}$. Consequently, to show that $\sigma(\{A_\alpha\}) = \sigma(\{B_\beta\})$ for two different sets of generators $\{A_\alpha\}$ and $\{B_\beta\}$, we only need to show that $A_\alpha \in \sigma(\{B_\beta\})$ for each α and that $B_\beta \in \sigma(\{A_\alpha\})$ for each β. For instance, considering $\mathcal{B}_\mathbb{Q} = \sigma(\{(a, b) : a < b \in \mathbb{Q}\})$, we have by this approach that $\mathcal{B}_\mathbb{Q} = \sigma(\{(a, b) : a < b \in \mathbb{R}\})$, as soon as we show that any interval (a, b) is in $\mathcal{B}_\mathbb{Q}$. To see this fact, note that for any real $a < b$ there are rational numbers $q_n < r_n$ such that $q_n \downarrow a$ and $r_n \uparrow b$, hence $(a, b) = \cup_n (q_n, r_n) \in \mathcal{B}_\mathbb{Q}$. Expanding on this, the next exercise provides useful alternative definitions of $\mathcal{B}$.

Exercise 1.1.17. *Verify the alternative definitions of the* Borel σ-algebra $\mathcal{B}$*:*

$$\begin{aligned}\sigma(\{(a,b) : a < b \in \mathbb{R}\}) &= \sigma(\{[a,b] : a < b \in \mathbb{R}\}) = \sigma(\{(-\infty,b] : b \in \mathbb{R}\}) \\ &= \sigma(\{(-\infty,b] : b \in \mathbb{Q}\}) = \sigma(\{O \subseteq \mathbb{R} \text{ open }\})\end{aligned}$$

If $A \subseteq \mathbb{R}$ is in $\mathcal{B}$ of Example 1.1.15, we say that A is a *Borel set*. In particular, all open (closed) subsets of $\mathbb{R}$ are Borel sets, as are many other sets. However,

Proposition 1.1.18. *There exists a subset of* $\mathbb{R}$ *that is not in* $\mathcal{B}$*. That is, not all subsets of* $\mathbb{R}$ *are Borel sets.*

Proof. See [**Bil95**, page 45]. □

Example 1.1.19. *Another classical example of an uncountable* Ω *is relevant for studying the experiment with an infinite number of coin tosses, that is,* $\Omega_\infty = \Omega_1^{\mathbb{N}}$ *for* $\Omega_1 = \{\mathrm{H}, \mathrm{T}\}$ *(indeed, setting* $\mathrm{H} = 1$ *and* $\mathrm{T} = 0$*, each infinite sequence* $\omega \in \Omega_\infty$ *is in correspondence with a unique real number* $x \in [0,1]$ *with* ω *being the binary expansion of* x*, see Exercise 1.2.13). The* σ*-algebra should at least allow us to consider any possible outcome of a finite number of coin tosses. The natural* σ*-algebra in this case is the minimal* σ*-algebra having this property, or put more formally* $\mathcal{F}_c = \sigma(\{A_{\theta,k}, \theta \in \Omega_1^k, k = 1,2,\ldots\})$*, where* $A_{\theta,k} = \{\omega \in \Omega_\infty : \omega_i = \theta_i, i = 1\ldots,k\}$ *for* $\theta = (\theta_1,\ldots,\theta_k)$*.*

The preceding example is a special case of the construction of a product of measurable spaces, which we detail now.

Example 1.1.20. *The* product *of the measurable spaces* $(\Omega_i, \mathcal{F}_i)$*,* $i = 1,\ldots,n$ *is the set* $\Omega = \Omega_1 \times \cdots \times \Omega_n$ *with the* σ*-algebra generated by* $\{A_1 \times \cdots \times A_n : A_i \in \mathcal{F}_i\}$*, denoted by* $\mathcal{F}_1 \times \cdots \mathcal{F}_n$*.*

You are now to check that the Borel σ-algebra of $\mathbb{R}^d$ is the product of d-copies of that of $\mathbb{R}$. As we see later, this helps simplifying the study of random vectors.

Exercise 1.1.21. *Show that for any* $d < \infty$*,*

$$\mathcal{B}_{\mathbb{R}^d} = \mathcal{B} \times \cdots \times \mathcal{B} = \sigma(\{(a_1,b_1) \times \cdots \times (a_d,b_d) : a_i < b_i \in \mathbb{R}, i = 1,\ldots,d\})$$

(you need to prove both *identities, with the middle term defined as in Example 1.1.20).*

Exercise 1.1.22. *Let* $\mathcal{F} = \sigma(A_\alpha, \alpha \in \Gamma)$ *where the collection of sets* A_α*,* $\alpha \in \Gamma$ *is uncountable (i.e.,* Γ *is uncountable). Prove that for each* $B \in \mathcal{F}$ *there exists a countable sub-collection* $\{A_{\alpha_j}, j = 1,2,\ldots\} \subset \{A_\alpha, \alpha \in \Gamma\}$*, such that* $B \in \sigma(\{A_{\alpha_j}, j = 1,2,\ldots\})$*.*

Often there is no explicit enumerative description of the σ-algebra generated by an infinite collection of subsets, but a notable exception is

Exercise 1.1.23. *Show that the sets in* $\mathcal{G} = \sigma(\{[a,b] : a,b \in \mathbf{Z}\})$ *are all possible unions of elements from the countable collection* $\{\{b\}, (b,b+1), b \in \mathbf{Z}\}$*, and deduce that* $\mathcal{B} \neq \mathcal{G}$*.*

Probability measures on the Borel σ-algebra of $\mathbb{R}$ are examples of *regular measures,* namely:

Exercise 1.1.24. *Show that if $\mathbf{P}$ is a probability measure on $(\mathbb{R}, \mathcal{B})$ then for any $A \in \mathcal{B}$ and $\epsilon > 0$, there exists an open set G containing A such that $\mathbf{P}(A) + \epsilon > \mathbf{P}(G)$.*

Here is more information about $\mathcal{B}_{\mathbb{R}^d}$.

Exercise 1.1.25. *Show that if μ is a finitely additive non-negative set function on $(\mathbb{R}^d, \mathcal{B}_{\mathbb{R}^d})$ such that $\mu(\mathbb{R}^d) = 1$ and for any Borel set A,*

$$\mu(A) = \sup\{\mu(K) : K \subseteq A,\, K \text{ compact }\},$$

then μ must be a probability measure.
Hint: *Argue by contradiction using the conclusion of Exercise 1.1.5. To this end, recall the finite intersection property (if compact $K_i \subset \mathbb{R}^d$ are such that $\bigcap_{i=1}^n K_i$ are non-empty for finite n, then the countable intersection $\bigcap_{i=1}^\infty K_i$ is also non-empty).*

1.1.3. Lebesgue measure and Carathéodory's theorem. Perhaps the most important measure on $(\mathbb{R}, \mathcal{B})$ is the *Lebesgue measure*, λ. It is the unique measure that satisfies $\lambda(F) = \sum_{k=1}^r (b_k - a_k)$ whenever $F = \bigcup_{k=1}^r (a_k, b_k]$ for some $r < \infty$ and $a_1 < b_1 < a_2 < b_2 \cdots < b_r$. Since $\lambda(\mathbb{R}) = \infty$, this is not a probability measure. However, when we restrict Ω to be the interval $(0, 1]$ we get

Example 1.1.26. *The* uniform probability measure *on $(0, 1]$, is denoted U and defined as above, now with added restrictions that $0 \le a_1$ and $b_r \le 1$. Alternatively, U is the restriction of the measure λ to the sub-σ-algebra $\mathcal{B}_{(0,1]}$ of $\mathcal{B}$.*

Exercise 1.1.27. *Show that $((0, 1], \mathcal{B}_{(0,1]}, U)$ is a non-atomic probability space and deduce that $(\mathbb{R}, \mathcal{B}, \lambda)$ is a non-atomic measure space.*

Note that any countable union of sets of probability zero has probability zero, but this is not the case for an uncountable union. For example, $U(\{x\}) = 0$ for every $x \in \mathbb{R}$, but $U(\mathbb{R}) = 1$.
As we have seen in Example 1.1.26 it is often impossible to explicitly specify the value of a measure on all sets of the σ-algebra $\mathcal{F}$. Instead, we wish to specify its values on a much smaller and better behaved collection of generators $\mathcal{A}$ of $\mathcal{F}$ and use Carathéodory's theorem to guarantee the existence of a unique measure on $\mathcal{F}$ that coincides with our specified values. To this end, we require that $\mathcal{A}$ be an algebra, that is,

Definition 1.1.28. *A collection $\mathcal{A}$ of subsets of Ω is an* algebra *(or a* field*) if*

(a) $\Omega \in \mathcal{A}$,
(b) *If $A \in \mathcal{A}$ then $A^c \in \mathcal{A}$ as well,*
(c) *If $A, B \in \mathcal{A}$ then also $A \cup B \in \mathcal{A}$.*

Remark. In view of the closure of algebra with respect to complements, we could have replaced the requirement that $\Omega \in \mathcal{A}$ with the (more standard) requirement that $\emptyset \in \mathcal{A}$. As part (c) of Definition 1.1.28 amounts to closure of an algebra under finite unions (and by DeMorgan's law also finite intersections), the difference between an algebra and a σ-algebra is that a σ-algebra is also closed under countable unions.

We sometimes make use of the fact that unlike generated σ-algebras, the algebra generated by a collection of sets $\mathcal{A}$ can be explicitly presented.

Exercise 1.1.29. *The algebra generated by a given collection of subsets $\mathcal{A}$, denoted $f(\mathcal{A})$, is the intersection of all algebras of subsets of Ω containing $\mathcal{A}$.*

(a) *Verify that $f(\mathcal{A})$ is indeed an algebra and that $f(\mathcal{A})$ is minimal in the sense that if $\mathcal{G}$ is an algebra and $\mathcal{A} \subseteq \mathcal{G}$, then $f(\mathcal{A}) \subseteq \mathcal{G}$.*

(b) *Show that $f(\mathcal{A})$ is the collection of all finite disjoint unions of sets of the form $\bigcap_{j=1}^{n_i} A_{ij}$, where for each i and j either A_{ij} or A_{ij}^c are in $\mathcal{A}$.*

We next state Carathéodory's extension theorem, a key result from measure theory, and demonstrate how it applies in the context of Example 1.1.26.

Theorem 1.1.30 (Carathéodory's extension theorem). *If $\mu_0 : \mathcal{A} \mapsto [0, \infty]$ is a countably additive set function on an algebra $\mathcal{A}$ then there exists a measure μ on $(\Omega, \sigma(\mathcal{A}))$ such that $\mu = \mu_0$ on $\mathcal{A}$. Furthermore, if $\mu_0(\Omega) < \infty$ then such a measure μ is unique.*

To construct the measure U on $\mathcal{B}_{(0,1]}$ let $\Omega = (0, 1]$ and

$$\mathcal{A} = \{(a_1, b_1] \cup \cdots \cup (a_r, b_r] : 0 \le a_1 < b_1 < \cdots < a_r < b_r \le 1\,,\ r < \infty\}$$

be a collection of subsets of $(0, 1]$. It is not hard to verify that $\mathcal{A}$ is an algebra, and further that $\sigma(\mathcal{A}) = \mathcal{B}_{(0,1]}$ (c.f. Exercise 1.1.17, for a similar issue, just with $(0, 1]$ replaced by $\mathbb{R}$). With U_0 denoting the non-negative set function on $\mathcal{A}$ such that

$$U_0\Big(\bigcup_{k=1}^{r}(a_k, b_k]\Big) = \sum_{k=1}^{r}(b_k - a_k)\,, \tag{1.1.1}$$

note that $U_0((0, 1]) = 1$, hence the existence of a unique probability measure U on $((0, 1], \mathcal{B}_{(0,1]})$ such that $U(A) = U_0(A)$ for sets $A \in \mathcal{A}$ follows by Carathéodory's extension theorem, as soon as we verify that

Lemma 1.1.31. *The set function U_0 is countably additive on $\mathcal{A}$. That is, if A_k is a sequence of disjoint sets in $\mathcal{A}$ such that $\cup_k A_k = A \in \mathcal{A}$, then $U_0(A) = \sum_k U_0(A_k)$.*

The proof of Lemma 1.1.31 is based on

Exercise 1.1.32. *Show that U_0 is finitely additive on $\mathcal{A}$. That is, $U_0(\bigcup_{k=1}^n A_k) = \sum_{k=1}^n U_0(A_k)$ for any finite collection of disjoint sets $A_1, \dots, A_n \in \mathcal{A}$.*

Proof. Let $G_n = \bigcup_{k=1}^n A_k$ and $H_n = A \setminus G_n$. Then, $H_n \downarrow \emptyset$ and since $A_k, A \in \mathcal{A}$ which is an algebra it follows that G_n and hence H_n are also in $\mathcal{A}$. By definition, U_0 is finitely additive on $\mathcal{A}$, so

$$U_0(A) = U_0(H_n) + U_0(G_n) = U_0(H_n) + \sum_{k=1}^{n} U_0(A_k)\,.$$

To prove that U_0 is countably additive, it suffices to show that $U_0(H_n) \downarrow 0$, for then

$$U_0(A) = \lim_{n\to\infty} U_0(G_n) = \lim_{n\to\infty}\sum_{k=1}^{n} U_0(A_k) = \sum_{k=1}^{\infty} U_0(A_k)\,.$$

To complete the proof, we argue by contradiction, assuming that $U_0(H_n) \ge 2\varepsilon$ for some $\varepsilon > 0$ and all n, where $H_n \downarrow \emptyset$ are elements of $\mathcal{A}$. By the definition of $\mathcal{A}$ and U_0, we can find for each ℓ a set $J_\ell \in \mathcal{A}$ whose closure $\overline{J}_\ell$ is a subset of H_ℓ and $U_0(H_\ell \setminus J_\ell) \le \varepsilon 2^{-\ell}$ (for example, add to each a_k in the representation of H_ℓ the minimum of $\varepsilon 2^{-\ell}/r$ and $(b_k - a_k)/2$). With U_0 finitely additive on the algebra $\mathcal{A}$ this implies that for each n,

$$U_0\Big(\bigcup_{\ell=1}^{n}(H_\ell \setminus J_\ell)\Big) \le \sum_{\ell=1}^{n} U_0(H_\ell \setminus J_\ell) \le \varepsilon\,.$$

As $H_n \subseteq H_\ell$ for all $\ell \leq n$, we have that

$$H_n \setminus \bigcap_{\ell \leq n} J_\ell = \bigcup_{\ell \leq n} (H_n \setminus J_\ell) \subseteq \bigcup_{\ell \leq n} (H_\ell \setminus J_\ell) .$$

Hence, by finite additivity of U_0 and our assumption that $U_0(H_n) \geq 2\varepsilon$, also

$$U_0(\bigcap_{\ell \leq n} J_\ell) = U_0(H_n) - U_0(H_n \setminus \bigcap_{\ell \leq n} J_\ell) \geq U_0(H_n) - U_0(\bigcup_{\ell \leq n} (H_\ell \setminus J_\ell)) \geq \varepsilon .$$

In particular, for every n, the set $\bigcap_{\ell \leq n} J_\ell$ is non-empty and therefore so are the decreasing sets $K_n = \bigcap_{\ell \leq n} \overline{J}_\ell$. Since K_n are compact sets (by Heine-Borel theorem), the set $\cap_\ell \overline{J}_\ell$ is then non-empty as well, and since $\overline{J}_\ell$ is a subset of H_ℓ for all ℓ we arrive at $\cap_\ell H_\ell$ non-empty, contradicting our assumption that $H_n \downarrow \emptyset$. □

Remark. The proof of Lemma 1.1.31 is generic (for finite measures). Namely, any non-negative finitely additive set function μ_0 on an algebra $\mathcal{A}$ is countably additive if $\mu_0(H_n) \downarrow 0$ whenever $H_n \in \mathcal{A}$ and $H_n \downarrow \emptyset$. Further, as this proof shows, when Ω is a topological space it suffices for countable additivity of μ_0 to have for any $H \in \mathcal{A}$ a sequence $J_k \in \mathcal{A}$ such that $\overline{J}_k \subseteq H$ are compact and $\mu_0(H \setminus J_k) \to 0$ as $k \to \infty$.

Exercise 1.1.33. *Show the necessity of the assumption that $\mathcal{A}$ be an algebra in Carathéodory's extension theorem, by giving an example of two probability measures $\mu \neq \nu$ on a measurable space $(\Omega, \mathcal{F})$ such that $\mu(A) = \nu(A)$ for all $A \in \mathcal{A}$ and $\mathcal{F} = \sigma(\mathcal{A})$.*
Hint: *This can be done with $\Omega = \{1, 2, 3, 4\}$ and $\mathcal{F} = 2^\Omega$.*

It is often useful to assume that the probability space we have is complete, in the sense we make precise now.

Definition 1.1.34. *We say that a measure space $(\Omega, \mathcal{F}, \mu)$ is* complete *if any subset N of any $B \in \mathcal{F}$ with $\mu(B) = 0$ is also in $\mathcal{F}$. If further $\mu = \mathbf{P}$ is a probability measure, we say that the probability space $(\Omega, \mathcal{F}, \mathbf{P})$ is a* complete probability space.

Our next theorem states that any measure space can be completed by adding to its σ-algebra all subsets of sets of zero measure (a procedure that depends on the measure in use).

Theorem 1.1.35. *Given a measure space $(\Omega, \mathcal{F}, \mu)$, let $\mathcal{N} = \{N : N \subseteq A$ for some $A \in \mathcal{F}$ with $\mu(A) = 0\}$ denote the collection of μ-*null sets. *Then, there exists a complete measure space $(\Omega, \overline{\mathcal{F}}, \overline{\mu})$, called the* completion *of the measure space $(\Omega, \mathcal{F}, \mu)$, such that $\overline{\mathcal{F}} = \{F \cup N : F \in \mathcal{F}, N \in \mathcal{N}\}$ and $\overline{\mu} = \mu$ on $\mathcal{F}$.*

Proof. This is beyond our scope, but see a detailed proof in [**Dur03**, page 450]. In particular, $\overline{\mathcal{F}} = \sigma(\mathcal{F}, \mathcal{N})$ and $\overline{\mu}(A \cup N) = \mu(A)$ for any $N \in \mathcal{N}$ and $A \in \mathcal{F}$ (c.f. [**Bil95**, Problems 3.10 and 10.5]). □

The following collections of sets play an important role in proving the easy part of Carathéodory's theorem, the uniqueness of the extension μ.

Definition 1.1.36. *A* π-system *is a collection $\mathcal{P}$ of sets closed under finite intersections (i.e. if $I \in \mathcal{P}$ and $J \in \mathcal{P}$ then $I \cap J \in \mathcal{P}$).*
A λ-system *is a collection $\mathcal{L}$ of sets containing Ω and $B \backslash A$ for any $A \subseteq B$ $A, B \in \mathcal{L}$,*

which is also closed under monotone increasing limits (i.e. if $A_i \in \mathcal{L}$ and $A_i \uparrow A$, then $A \in \mathcal{L}$ as well).

Obviously, an algebra is a π-system. Though an algebra may not be a λ-system,

Proposition 1.1.37. *A collection $\mathcal{F}$ of sets is a σ-algebra if and only if it is both a π-system and a λ-system.*

Proof. The fact that a σ-algebra is a λ-system is a trivial consequence of Definition 1.1.1. To prove the converse direction, suppose that $\mathcal{F}$ is both a π-system and a λ-system. Then Ω is in the λ-system $\mathcal{F}$ and so is $A^c = \Omega \setminus A$ for any $A \in \mathcal{F}$. Further, with $\mathcal{F}$ also a π-system we have that

$$A \cup B = \Omega \setminus (A^c \cap B^c) \in \mathcal{F} ,$$

for any $A, B \in \mathcal{F}$. Consequently, if $A_i \in \mathcal{F}$ then so are also $G_n = A_1 \cup \cdots \cup A_n \in \mathcal{F}$. Since $\mathcal{F}$ is a λ-system and $G_n \uparrow \bigcup_i A_i$, it follows that $\bigcup_i A_i \in \mathcal{F}$ as well, completing the verification that $\mathcal{F}$ is a σ-algebra. □

The main tool in proving the uniqueness of the extension is *Dynkin's $\pi-\lambda$ theorem*, stated next.

Theorem 1.1.38 (Dynkin's $\pi - \lambda$ theorem). *If $\mathcal{P} \subseteq \mathcal{L}$ with $\mathcal{P}$ a π-system and $\mathcal{L}$ a λ-system then $\sigma(\mathcal{P}) \subseteq \mathcal{L}$.*

Proof. A short though dense exercise in set manipulations shows that the smallest λ-system containing $\mathcal{P}$ is a π-system (for details see [**Wil91**, Section A.1.3], or either the proof of [**Dur03**, Theorem A.2.1], or that of [**Bil95**, Theorem 3.2]). By Proposition 1.1.37 it is a σ-algebra, hence contains $\sigma(\mathcal{P})$. Further, it is contained in the λ-system $\mathcal{L}$, as $\mathcal{L}$ also contains $\mathcal{P}$. □

Remark. Proposition 1.1.37 remains valid even if in the definition of λ-system we relax the condition that $B \setminus A \in \mathcal{L}$ for any $A \subseteq B$ $A, B \in \mathcal{L}$, to the condition $A^c \in \mathcal{L}$ whenever $A \in \mathcal{L}$. However, Dynkin's theorem does not hold under the latter definition.

As we show next, the uniqueness part of Carathéodory's theorem, is an immediate consequence of the $\pi - \lambda$ theorem.

Proposition 1.1.39. *If two measures μ_1 and μ_2 on $(\Omega, \sigma(\mathcal{P}))$ agree on the π-system $\mathcal{P}$ and are such that $\mu_1(\Omega) = \mu_2(\Omega) < \infty$, then $\mu_1 = \mu_2$.*

Proof. Let $\mathcal{L} = \{A \in \sigma(\mathcal{P}) : \mu_1(A) = \mu_2(A)\}$. Our assumptions imply that $\mathcal{P} \subseteq \mathcal{L}$ and that $\Omega \in \mathcal{L}$. Further, $\sigma(\mathcal{P})$ is a λ-system (by Proposition 1.1.37), and if $A \subseteq B$, $A, B \in \mathcal{L}$, then by additivity of the finite measures μ_1 and μ_2,

$$\mu_1(B \setminus A) = \mu_1(B) - \mu_1(A) = \mu_2(B) - \mu_2(A) = \mu_2(B \setminus A),$$

that is, $B \setminus A \in \mathcal{L}$. Similarly, if $A_i \uparrow A$ and $A_i \in \mathcal{L}$, then by the continuity from below of μ_1 and μ_2 (see remark following Exercise 1.1.4),

$$\mu_1(A) = \lim_{n\to\infty} \mu_1(A_n) = \lim_{n\to\infty} \mu_2(A_n) = \mu_2(A) ,$$

so that $A \in \mathcal{L}$. We conclude that $\mathcal{L}$ is a λ-system, hence by Dynkin's $\pi-\lambda$ theorem, $\sigma(\mathcal{P}) \subseteq \mathcal{L}$, that is, $\mu_1 = \mu_2$. □

Remark. With a somewhat more involved proof one can relax the condition $\mu_1(\Omega) = \mu_2(\Omega) < \infty$ to the existence of $A_n \in \mathcal{P}$ such that $A_n \uparrow \Omega$ and $\mu_1(A_n) < \infty$ (c.f. [**Dur03**, Theorem A.2.2] or [**Bil95**, Theorem 10.3] for details). Accordingly, in Carathéodory's extension theorem we can relax $\mu_0(\Omega) < \infty$ to the assumption that μ_0 is a *σ-finite measure*, that is $\mu_0(A_n) < \infty$ for some $A_n \in \mathcal{A}$ such that $A_n \uparrow \Omega$, as is the case with *Lebesgue's measure* λ on $\mathbb{R}$.

We conclude this subsection with an outline the proof of Carathéodory's extension theorem, noting that since an algebra $\mathcal{A}$ is a π-system and $\Omega \in \mathcal{A}$, the uniqueness of the extension to $\sigma(\mathcal{A})$ follows from Proposition 1.1.39. Our outline of the existence of an extension follows [**Wil91**, Section A.1.8] (for a similar treatment see [**Dur03**, Pages 446-448], or see [**Bil95**, Theorem 11.3] for the proof of a somewhat stronger result). This outline centers on the construction of the appropriate outer measure, a relaxation of the concept of measure, which we now define.

Definition 1.1.40. *An increasing, countably sub-additive, non-negative set function μ^* on a measurable space $(\Omega, \mathcal{F})$ is called an* outer measure. *That is, $\mu^* : \mathcal{F} \mapsto [0, \infty]$, having the properties:*
(a) $\mu^*(\emptyset) = 0$ and $\mu^*(A_1) \leq \mu^*(A_2)$ for any $A_1, A_2 \in \mathcal{F}$ with $A_1 \subseteq A_2$.
(b) $\mu^*(\bigcup_n A_n) \leq \sum_n \mu^*(A_n)$ for any countable collection of sets $A_n \in \mathcal{F}$.

In the first step of the proof we define the increasing, non-negative set function

$$\mu^*(E) = \inf\{\sum_{n=1}^{\infty} \mu_0(A_n) : E \subseteq \bigcup_n A_n, A_n \in \mathcal{A}\},$$

for $E \in \mathcal{F} = 2^{\Omega}$, and prove that it is countably sub-additive, hence an outer measure on $\mathcal{F}$.

By definition, $\mu^*(A) \leq \mu_0(A)$ for any $A \in \mathcal{A}$. In the second step we prove that if in addition $A \subseteq \bigcup_n A_n$ for $A_n \in \mathcal{A}$, then the countable additivity of μ_0 on $\mathcal{A}$ results with $\mu_0(A) \leq \sum_n \mu_0(A_n)$. Consequently, $\mu^* = \mu_0$ on the algebra $\mathcal{A}$.

The third step uses the countable additivity of μ_0 on $\mathcal{A}$ to show that for any $A \in \mathcal{A}$ the outer measure μ^* is additive when splitting subsets of Ω by intersections with A and A^c. That is, we show that any element of $\mathcal{A}$ is a μ^*-measurable set, as defined next.

Definition 1.1.41. *Let λ be a non-negative set function on a measurable space $(\Omega, \mathcal{F})$, with $\lambda(\emptyset) = 0$. We say that $A \in \mathcal{F}$ is a λ-measurable set if $\lambda(F) = \lambda(F \cap A) + \lambda(F \cap A^c)$ for all $F \in \mathcal{F}$.*

The fourth step consists of proving the following general lemma.

Lemma 1.1.42 (Carathéodory's lemma). *Let μ^* be an outer measure on a measurable space $(\Omega, \mathcal{F})$. Then the μ^*-measurable sets in $\mathcal{F}$ form a σ-algebra $\mathcal{G}$ on which μ^* is countably additive, so that $(\Omega, \mathcal{G}, \mu^*)$ is a measure space.*

In the current setting, with $\mathcal{A}$ contained in the σ-algebra $\mathcal{G}$, it follows that $\sigma(\mathcal{A}) \subseteq \mathcal{G}$ on which μ^* is a measure. Thus, the restriction μ of μ^* to $\sigma(\mathcal{A})$ is the stated measure that coincides with μ_0 on $\mathcal{A}$.

Remark. In the setting of Carathéodory's extension theorem for finite measures, we have that the σ-algebra $\mathcal{G}$ of all μ^*-measurable sets is the completion of $\sigma(\mathcal{A})$ with respect to μ (c.f. [**Bil95**, Page 45] or [**Dur03**, Theorem A.3.2]). In the context of Lebesgue's measure U on $\mathcal{B}_{(0,1]}$, this is the σ-algebra $\overline{\mathcal{B}}_{(0,1]}$ of all Lebesgue measurable subsets of $(0,1]$. Associated with it are the *Lebesgue measurable* functions $f : (0,1] \mapsto \mathbb{R}$ for which $f^{-1}(B) \in \overline{\mathcal{B}}_{(0,1]}$ for all $B \in \mathcal{B}$. However, as noted for example in [**Dur03**, Theorem A.3.4], the non Borel set constructed in the proof of Proposition 1.1.18 is also non Lebesgue measurable.

The following concept of a monotone class of sets is a considerable relaxation of that of a λ-system (hence also of a σ-algebra, see Proposition 1.1.37).

DEFINITION 1.1.43. *A* monotone class *is a collection $\mathcal{M}$ of sets closed under both monotone increasing and monotone decreasing limits (i.e. if $A_i \in \mathcal{M}$ and either $A_i \uparrow A$ or $A_i \downarrow A$, then $A \in \mathcal{M}$).*

When starting from an algebra instead of a π-system, one may save effort by applying Halmos's monotone class theorem instead of Dynkin's $\pi - \lambda$ theorem.

THEOREM 1.1.44 (HALMOS'S MONOTONE CLASS THEOREM). *If $\mathcal{A} \subseteq \mathcal{M}$ with $\mathcal{A}$ an algebra and $\mathcal{M}$ a monotone class then $\sigma(\mathcal{A}) \subseteq \mathcal{M}$.*

PROOF. Clearly, any algebra which is a monotone class must be a σ-algebra. Another short though dense exercise in set manipulations shows that the intersection $m(\mathcal{A})$ of all monotone classes containing an algebra $\mathcal{A}$ is both an algebra and a monotone class (see the proof of [**Bil95**, Theorem 3.4]). Consequently, $m(\mathcal{A})$ is a σ-algebra. Since $\mathcal{A} \subseteq m(\mathcal{A})$ this implies that $\sigma(\mathcal{A}) \subseteq m(\mathcal{A})$ and we complete the proof upon noting that $m(\mathcal{A}) \subseteq \mathcal{M}$. □

EXERCISE 1.1.45. *We say that a subset V of $\{1, 2, 3, \cdots\}$ has* Cesáro density $\gamma(V)$ *and write $V \in$ CES if the limit*

$$\gamma(V) = \lim_{n\to\infty} n^{-1} |V \cap \{1, 2, 3, \cdots, n\}| ,$$

exists. Give an example of sets $V_1 \in$ CES and $V_2 \in$ CES for which $V_1 \cap V_2 \notin$ CES. Thus, CES *is not an algebra.*

Here is an alternative specification of the concept of algebra.

EXERCISE 1.1.46.

(a) *Suppose that $\Omega \in \mathcal{A}$ and that $A \cap B^c \in \mathcal{A}$ whenever $A, B \in \mathcal{A}$. Show that $\mathcal{A}$ is an algebra.*
(b) *Give an example of a collection $\mathcal{C}$ of subsets of Ω such that $\Omega \in \mathcal{C}$, if $A \in \mathcal{C}$ then $A^c \in \mathcal{C}$ and if $A, B \in \mathcal{C}$ are* disjoint *then also $A \cup B \in \mathcal{C}$, while $\mathcal{C}$ is not an algebra.*

As we already saw, the σ-algebra structure is preserved under intersections. However, whereas the increasing union of algebras is an algebra, it is not necessarily the case for σ-algebras.

EXERCISE 1.1.47. *Suppose that $\mathcal{A}_n$ are classes of sets such that $\mathcal{A}_n \subseteq \mathcal{A}_{n+1}$.*

(a) *Show that if $\mathcal{A}_n$ are algebras then so is $\bigcup_{n=1}^\infty \mathcal{A}_n$.*
(b) *Provide an example of σ-algebras $\mathcal{A}_n$ for which $\bigcup_{n=1}^\infty \mathcal{A}_n$ is not a σ-algebra.*

1.2. Random variables and their distribution

Random variables are numerical functions $\omega \mapsto X(\omega)$ of the outcome of our random experiment. However, in order to have a successful mathematical theory, we limit our interest to the subset of *measurable* functions (or more generally, measurable mappings), as defined in Subsection 1.2.1 and study the closure properties of this collection in Subsection 1.2.2. Subsection 1.2.3 is devoted to the characterization of the collection of distribution functions induced by random variables.

1.2.1. Indicators, simple functions and random variables.

We start with the definition of random variables, first in the general case, and then restricted to $\mathbb{R}$-valued variables.

Definition 1.2.1. *A mapping $X : \Omega \mapsto \mathbb{S}$ between two measurable spaces $(\Omega, \mathcal{F})$ and $(\mathbb{S}, \mathcal{S})$ is called an $(\mathbb{S}, \mathcal{S})$-valued* Random Variable *(R.V.) if*

$$X^{-1}(B) := \{\omega : X(\omega) \in B\} \in \mathcal{F} \qquad \forall B \in \mathcal{S}.$$

Such a mapping is also called a measurable mapping.

Definition 1.2.2. *When we say that X is a random variable, or a* measurable function, *we mean an $(\mathbb{R}, \mathcal{B})$-valued random variable which is the most common type of R.V. we shall encounter. We let $m\mathcal{F}$ denote the collection of all $(\mathbb{R}, \mathcal{B})$-valued measurable mappings, so X is a R.V. if and only if $X \in m\mathcal{F}$. If in addition Ω is a topological space and $\mathcal{F} = \sigma(\{O \subseteq \Omega \text{ open }\})$ is the corresponding Borel σ-algebra, we say that $X : \Omega \mapsto \mathbb{R}$ is a* Borel *(measurable)* function. *More generally, a* random vector *is an $(\mathbb{R}^d, \mathcal{B}_{\mathbb{R}^d})$-valued R.V. for some $d < \infty$.*

The next exercise shows that a random vector is merely a finite collection of R.V. on the same probability space.

Exercise 1.2.3. *Relying on Exercise 1.1.21 and Theorem 1.2.9, show that $\underline{X} : \Omega \mapsto \mathbb{R}^d$ is a random vector if and only if $\underline{X}(\omega) = (X_1(\omega), \ldots, X_d(\omega))$ with each $X_i : \Omega \mapsto \mathbb{R}$ a R.V.*

Hint: *Note that $\underline{X}^{-1}(B_1 \times \ldots \times B_d) = \bigcap\limits_{i=1}^{d} X_i^{-1}(B_i)$.*

We now provide two important generic examples of random variables.

Example 1.2.4. *For any $A \in \mathcal{F}$ the function $I_A(\omega) = \begin{cases} 1, \omega \in A \\ 0, \omega \notin A \end{cases}$ is a R.V. Indeed, $\{\omega : I_A(\omega) \in B\}$ is for any $B \subseteq \mathbb{R}$ one of the four sets $\emptyset$, A, A^c or Ω (depending on whether $0 \in B$ or not and whether $1 \in B$ or not), all of whom are in $\mathcal{F}$. We call such R.V. also an* indicator function.

Exercise 1.2.5. *By the same reasoning check that $X(\omega) = \sum_{n=1}^{N} c_n I_{A_n}(\omega)$ is a R.V. for any finite N, non-random $c_n \in \mathbb{R}$ and sets $A_n \in \mathcal{F}$. We call any such X a* simple function, *denoted by $X \in$ SF.*

Our next proposition explains why simple functions are quite useful in probability theory.

Proposition 1.2.6. *For every R.V. $X(\omega)$ there exists a sequence of simple functions $X_n(\omega)$ such that $X_n(\omega) \to X(\omega)$ as $n \to \infty$, for each fixed $\omega \in \Omega$.*

PROOF. Let

$$f_n(x) = n\mathbf{1}_{x>n} + \sum_{k=0}^{n2^n-1} k2^{-n}\mathbf{1}_{(k2^{-n},(k+1)2^{-n}]}(x)\,,$$

noting that for R.V. $X \geq 0$, we have that $X_n = f_n(X)$ are simple functions. Since $X \geq X_{n+1} \geq X_n$ and $X(\omega) - X_n(\omega) \leq 2^{-n}$ whenever $X(\omega) \leq n$, it follows that $X_n(\omega) \to X(\omega)$ as $n \to \infty$, for each ω.

We write a general R.V. as $X(\omega) = X_+(\omega) - X_-(\omega)$ where $X_+(\omega) = \max(X(\omega), 0)$ and $X_-(\omega) = -\min(X(\omega), 0)$ are non-negative R.V.-s. By the above argument the simple functions $X_n = f_n(X_+) - f_n(X_-)$ have the convergence property we claimed. □

Note that in case $\mathcal{F} = 2^\Omega$, every mapping $X : \Omega \mapsto \mathbb{S}$ is measurable (and therefore is an $(\mathbb{S}, \mathcal{S})$-valued R.V.). The choice of the σ-algebra $\mathcal{F}$ is very important in determining the class of all $(\mathbb{S}, \mathcal{S})$-valued R.V. For example, there are non-trivial σ-algebras $\mathcal{G}$ and $\mathcal{F}$ on $\Omega = \mathbb{R}$ such that $X(\omega) = \omega$ is a measurable function for $(\Omega, \mathcal{F})$, but is non-measurable for $(\Omega, \mathcal{G})$. Indeed, one such example is when $\mathcal{F}$ is the Borel σ-algebra $\mathcal{B}$ and $\mathcal{G} = \sigma(\{[a, b] : a, b \in \mathbf{Z}\})$ (for example, the set $\{\omega : \omega \leq \alpha\}$ is not in $\mathcal{G}$ whenever $\alpha \notin \mathbf{Z}$).

Building on Proposition 1.2.6 we have the following analog of Halmos's monotone class theorem. It allows us to deduce in the sequel general properties of (bounded) measurable functions upon verifying them only for indicators of elements of π-systems.

THEOREM 1.2.7 (MONOTONE CLASS THEOREM). *Suppose $\mathcal{H}$ is a collection of $\mathbb{R}$-valued functions on Ω such that:*

(a) *The constant function 1 is an element of $\mathcal{H}$.*
(b) *$\mathcal{H}$ is a vector space over $\mathbb{R}$. That is, if $h_1, h_2 \in \mathcal{H}$ and $c_1, c_2 \in \mathbb{R}$ then $c_1h_1 + c_2h_2$ is in $\mathcal{H}$.*
(c) *If $h_n \in \mathcal{H}$ are non-negative and $h_n \uparrow h$ where h is a (bounded) real-valued function on Ω, then $h \in \mathcal{H}$.*

If $\mathcal{P}$ is a π-system and $I_A \in \mathcal{H}$ for all $A \in \mathcal{P}$, then $\mathcal{H}$ contains all (bounded) functions on Ω that are measurable with respect to $\sigma(\mathcal{P})$.

REMARK. We stated here two versions of the monotone class theorem, with the less restrictive assumption that (c) holds only for bounded h yielding the weaker conclusion about bounded elements of $m\sigma(\mathcal{P})$. In the sequel we use both versions, which as we see next, are derived by essentially the same proof. Adapting this proof you can also show that any collection $\mathcal{H}$ of non-negative functions on Ω satisfying the conditions of Theorem 1.2.7 apart from requiring (b) to hold only when $c_1h_1 + c_2h_2 \geq 0$, must contain all non-negative elements of $m\sigma(\mathcal{P})$.

PROOF. Let $\mathcal{L} = \{A \subseteq \Omega : I_A \in \mathcal{H}\}$. From (a) we have that $\Omega \in \mathcal{L}$, while (b) implies that $B \setminus A$ is in $\mathcal{L}$ whenever $A \subseteq B$ are both in $\mathcal{L}$. Further, in view of (c) the collection $\mathcal{L}$ is closed under monotone increasing limits. Consequently, $\mathcal{L}$ is a λ-system, so by Dynkin's π-λ theorem, our assumption that $\mathcal{L}$ contains $\mathcal{P}$ results with $\sigma(\mathcal{P}) \subseteq \mathcal{L}$. With $\mathcal{H}$ a vector space over $\mathbb{R}$, this in turn implies that $\mathcal{H}$ contains all simple functions with respect to the measurable space $(\Omega, \sigma(\mathcal{P}))$. In the proof of Proposition 1.2.6 we saw that any (bounded) measurable function is a difference of two (bounded) non-negative functions each of which is a monotone increasing limit of certain non-negative simple functions. Thus, from (b) and (c) we conclude that $\mathcal{H}$ contains all (bounded) measurable functions with respect to $(\Omega, \sigma(\mathcal{P}))$. □

The concept of almost sure prevails throughout probability theory.

Definition 1.2.8. *We say that two $(\mathbb{S}, \mathcal{S})$-valued R.V. X and Y defined on the same probability space $(\Omega, \mathcal{F}, \mathbf{P})$ are almost surely the same if $\mathbf{P}(\{\omega : X(\omega) \neq Y(\omega)\}) = 0$. This shall be denoted by $X \overset{a.s.}{=} Y$. More generally, same notation applies to any property of R.V. For example, $X(\omega) \geq 0$ a.s. means that $\mathbf{P}(\{\omega : X(\omega) < 0\}) = 0$. Hereafter, we shall consider X and Y such that $X \overset{a.s.}{=} Y$ to be the same $\mathbb{S}$-valued R.V. hence often omit the qualifier "a.s." when stating properties of R.V. We also use the terms* almost surely *(a.s.),* almost everywhere *(a.e.), and* with probability 1 *(w.p.1) interchangeably.*

Since the σ-algebra $\mathcal{S}$ might be huge, it is very important to note that we may verify that a given mapping is measurable without the need to check that the pre-image $X^{-1}(B)$ is in $\mathcal{F}$ for every $B \in \mathcal{S}$. Indeed, as shown next, it suffices to do this only for a collection (of our choice) of generators of $\mathcal{S}$.

Theorem 1.2.9. *If $\mathcal{S} = \sigma(\mathcal{A})$ and $X : \Omega \mapsto \mathbb{S}$ is such that $X^{-1}(A) \in \mathcal{F}$ for all $A \in \mathcal{A}$, then X is an $(\mathbb{S}, \mathcal{S})$-valued R.V.*

Proof. We first check that $\widehat{\mathcal{S}} = \{B \in \mathcal{S} : X^{-1}(B) \in \mathcal{F}\}$ is a σ-algebra. Indeed,
a). $\emptyset \in \widehat{\mathcal{S}}$ since $X^{-1}(\emptyset) = \emptyset$.
b). If $A \in \widehat{\mathcal{S}}$ then $X^{-1}(A) \in \mathcal{F}$. With $\mathcal{F}$ a σ-algebra, $X^{-1}(A^c) = \big(X^{-1}(A)\big)^c \in \mathcal{F}$. Consequently, $A^c \in \widehat{\mathcal{S}}$.
c). If $A_n \in \widehat{\mathcal{S}}$ for all n then $X^{-1}(A_n) \in \mathcal{F}$ for all n. With $\mathcal{F}$ a σ-algebra, then also $X^{-1}(\bigcup_n A_n) = \bigcup_n X^{-1}(A_n) \in \mathcal{F}$. Consequently, $\bigcup_n A_n \in \widehat{\mathcal{S}}$.
Our assumption that $\mathcal{A} \subseteq \widehat{\mathcal{S}}$, then translates to $\mathcal{S} = \sigma(\mathcal{A}) \subseteq \widehat{\mathcal{S}}$, as claimed. □

The most important σ-algebras are those generated by ($(\mathbb{S}, \mathcal{S})$-valued) random variables, as defined next.

Exercise 1.2.10. *Adapting the proof of Theorem 1.2.9, show that for any mapping $X : \Omega \mapsto \mathbb{S}$ and any σ-algebra $\mathcal{S}$ of subsets of $\mathbb{S}$, the collection $\{X^{-1}(B) : B \in \mathcal{S}\}$ is a σ-algebra. Verify that X is an $(\mathbb{S}, \mathcal{S})$-valued R.V. if and only if $\{X^{-1}(B) : B \in \mathcal{S}\} \subseteq \mathcal{F}$, in which case we denote $\{X^{-1}(B) : B \in \mathcal{S}\}$ either by $\sigma(X)$ or by $\mathcal{F}^X$ and call it the* σ-algebra generated by X*.*

To practice your understanding of generated σ-algebras, solve the next exercise, providing a convenient collection of generators for $\sigma(X)$.

Exercise 1.2.11. *If X is an $(\mathbb{S}, \mathcal{S})$-valued R.V. and $\mathcal{S} = \sigma(\mathcal{A})$ then $\sigma(X)$ is generated by the collection of sets $X^{-1}(\mathcal{A}) := \{X^{-1}(A) : A \in \mathcal{A}\}$.*

An important example of use of Exercise 1.2.11 corresponds to $(\mathbb{R}, \mathcal{B})$-valued random variables and $\mathcal{A} = \{(-\infty, x] : x \in \mathbb{R}\}$ (or even $\mathcal{A} = \{(-\infty, x] : x \in \mathbb{Q}\}$) which generates $\mathcal{B}$ (see Exercise 1.1.17), leading to the following alternative definition of the σ-algebra generated by such R.V. X.

DEFINITION 1.2.12. *Given a function $X : \Omega \mapsto \mathbb{R}$ we denote by $\sigma(X)$ or by $\mathcal{F}^X$ the* smallest *σ-algebra $\mathcal{F}$ such that $X(\omega)$ is a measurable mapping from $(\Omega, \mathcal{F})$ to $(\mathbb{R}, \mathcal{B})$. Alternatively,*

$$\sigma(X) = \sigma(\{\omega : X(\omega) \le \alpha\}, \alpha \in \mathbb{R}) = \sigma(\{\omega : X(\omega) \le q\}, q \in \mathbb{Q}) \,.$$

More generally, given a random vector $\underline{X} = (X_1, \ldots, X_n)$, that is, random variables $X_1, \ldots, X_n$ on the same probability space, let $\sigma(X_k, k \le n)$ (or $\mathcal{F}^{\mathbf{X}}_n$), denote the smallest σ-algebra $\mathcal{F}$ such that $X_k(\omega)$, $k = 1, \ldots, n$ are measurable on $(\Omega, \mathcal{F})$. Alternatively,

$$\sigma(X_k, k \le n) = \sigma(\{\omega : X_k(\omega) \le \alpha\}, \alpha \in \mathbb{R}, k \le n) \,.$$

Finally, given a possibly uncountable collection of functions $X_\gamma : \Omega \mapsto \mathbb{R}$, indexed by $\gamma \in \Gamma$, we denote by $\sigma(X_\gamma, \gamma \in \Gamma)$ (or simply by $\mathcal{F}^{\mathbf{X}}$), the smallest σ-algebra $\mathcal{F}$ such that $X_\gamma(\omega)$, $\gamma \in \Gamma$ are measurable on $(\Omega, \mathcal{F})$.

The concept of σ-algebra is needed in order to produce a rigorous mathematical theory. It further has the *crucial* role of quantifying the amount of information we have. For example, $\sigma(X)$ contains exactly those events A for which we can say whether $\omega \in A$ or not, based on the value of $X(\omega)$. Interpreting Example 1.1.19 as corresponding to sequentially tossing coins, the R.V. $X_n(\omega) = \omega_n$ gives the result of the n-th coin toss in our experiment Ω_∞ of infinitely many such tosses. The σ-algebra $\mathcal{F}_n = 2^{\Omega_n}$ of Example 1.1.6 then contains exactly the information we have upon observing the outcome of the first n coin tosses, whereas the larger σ-algebra $\mathcal{F}_c$ allows us to also study the limiting properties of this sequence (and as you show next, $\mathcal{F}_c$ is isomorphic, in the sense of Definition 1.4.24, to $\mathcal{B}_{[0,1]}$).

EXERCISE 1.2.13. *Let $\mathcal{F}_c$ denote the cylindrical σ-algebra for the set $\Omega_\infty = \{0,1\}^{\mathbb{N}}$ of infinite binary sequences, as in Example 1.1.19.*

(a) *Show that $X(\omega) = \sum_{n=1}^\infty \omega_n 2^{-n}$ is a measurable map from $(\Omega_\infty, \mathcal{F}_c)$ to $([0,1], \mathcal{B}_{[0,1]})$.*
(b) *Conversely, let $Y(x) = (\omega_1, \ldots, \omega_n, \ldots)$ where for each $n \ge 1$, $\omega_n(1) = 1$ while $\omega_n(x) = I(\lfloor 2^n x \rfloor$ is an odd number) when $x \in [0,1)$. Show that $Y = X^{-1}$ is a measurable map from $([0,1], \mathcal{B}_{[0,1]})$ to $(\Omega_\infty, \mathcal{F}_c)$.*

Here are some alternatives for Definition 1.2.12.

EXERCISE 1.2.14. *Verify the following relations and show that each generating collection of sets on the right hand side is a π-system.*

(a) $\sigma(X) = \sigma(\{\omega : X(\omega) \le \alpha\}, \alpha \in \mathbb{R})$
(b) $\sigma(X_k, k \le n) = \sigma(\{\omega : X_k(\omega) \le \alpha_k, 1 \le k \le n\}, \alpha_1, \ldots, \alpha_n \in \mathbb{R})$
(c) $\sigma(X_1, X_2, \ldots) = \sigma(\{\omega : X_k(\omega) \le \alpha_k, 1 \le k \le m\}, \alpha_1, \ldots, \alpha_m \in \mathbb{R}, m \in \mathbb{N})$
(d) $\sigma(X_1, X_2, \ldots) = \sigma(\bigcup_n \sigma(X_k, k \le n))$

As you next show, when approximating a random variable by a simple function, one may also specify the latter to be based on sets in any generating algebra.

EXERCISE 1.2.15. *Suppose $(\Omega, \mathcal{F}, \mathbf{P})$ is a probability space, with $\mathcal{F} = \sigma(\mathcal{A})$ for an algebra $\mathcal{A}$.*

(a) *Show that* $\inf\{\mathbf{P}(A \Delta B) : A \in \mathcal{A}\} = 0$ *for any $B \in \mathcal{F}$ (recall that $A \Delta B = (A \cap B^c) \cup (A^c \cap B)$).*

(b) *Show that for any bounded random variable X and $\epsilon > 0$ there exists a simple function $Y = \sum_{n=1}^{N} c_n I_{A_n}$ with $A_n \in \mathcal{A}$ such that $\mathbf{P}(|X - Y| > \epsilon) < \epsilon$.*

EXERCISE 1.2.16. *Let $\mathcal{F} = \sigma(A_\alpha, \alpha \in \Gamma)$ and suppose there exist $\omega_1 \neq \omega_2 \in \Omega$ such that for any $\alpha \in \Gamma$, either $\{\omega_1, \omega_2\} \subseteq A_\alpha$ or $\{\omega_1, \omega_2\} \subseteq A_\alpha^c$.*

(a) *Show that if mapping X is measurable on $(\Omega, \mathcal{F})$ then $X(\omega_1) = X(\omega_2)$.*

(b) *Provide an explicit σ-algebra $\mathcal{F}$ of subsets of $\Omega = \{1, 2, 3\}$ and a mapping $X : \Omega \mapsto \mathbb{R}$ which is* not *a random variable on $(\Omega, \mathcal{F})$.*

We conclude with a glimpse of the canonical measurable space associated with a stochastic process $(X_t, t \in \mathbb{T})$ (for more on this, see Lemma 7.1.7).

EXERCISE 1.2.17. *Fixing a possibly uncountable collection of random variables X_t, indexed by $t \in \mathbb{T}$, let $\mathcal{F}_{\mathbb{C}}^{\mathbf{X}} = \sigma(X_t, t \in \mathbb{C})$ for each $\mathbb{C} \subseteq \mathbb{T}$. Show that*

$$\mathcal{F}_{\mathbb{T}}^{\mathbf{X}} = \bigcup_{\mathbb{C} \text{ countable}} \mathcal{F}_{\mathbb{C}}^{\mathbf{X}}$$

and that any R.V. Z on $(\Omega, \mathcal{F}_{\mathbb{T}}^{\mathbf{X}})$ is measurable on $\mathcal{F}_{\mathbb{C}}^{\mathbf{X}}$ for some countable *$\mathbb{C} \subseteq \mathbb{T}$.*

1.2.2. Closure properties of random variables. For the typical measurable space with uncountable Ω it is impractical to list all possible R.V. Instead, we state a few useful closure properties that often help us in showing that a given mapping $X(\omega)$ is indeed a R.V.

We start with closure with respect to the composition of a R.V. and a measurable mapping.

PROPOSITION 1.2.18. *If $X : \Omega \mapsto \mathbb{S}$ is an $(\mathbb{S}, \mathcal{S})$-valued R.V. and f is a measurable mapping from $(\mathbb{S}, \mathcal{S})$ to $(\mathbb{T}, \mathcal{T})$, then the composition $f(X) : \Omega \mapsto \mathbb{T}$ is a $(\mathbb{T}, \mathcal{T})$-valued R.V.*

PROOF. Considering an arbitrary $B \in \mathcal{T}$, we know that $f^{-1}(B) \in \mathcal{S}$ since f is a measurable mapping. Thus, as X is an $(\mathbb{S}, \mathcal{S})$-valued R.V. it follows that

$$[f(X)]^{-1}(B) = X^{-1}(f^{-1}(B)) \in \mathcal{F} .$$

This holds for any $B \in \mathcal{T}$, thus concluding the proof. □

In view of Exercise 1.2.3 we have the following special case of Proposition 1.2.18, corresponding to $\mathbb{S} = \mathbb{R}^n$ and $\mathbb{T} = \mathbb{R}$ equipped with the respective Borel σ-algebras.

COROLLARY 1.2.19. *Let X_i, $i = 1, \ldots, n$ be R.V. on the same measurable space $(\Omega, \mathcal{F})$ and $f : \mathbb{R}^n \mapsto \mathbb{R}$ a Borel function. Then, $f(X_1, \ldots, X_n)$ is also a R.V. on the same space.*

To appreciate the power of Corollary 1.2.19, consider the following exercise, in which you show that every continuous function is also a Borel function.

EXERCISE 1.2.20. *Suppose $(\mathbb{S}, \rho)$ is a metric space (for example, $\mathbb{S} = \mathbb{R}^n$). A function $g : \mathbb{S} \mapsto [-\infty, \infty]$ is called* lower semi-continuous *(l.s.c.) if $\liminf_{\rho(y,x) \downarrow 0} g(y) \geq g(x)$, for all $x \in \mathbb{S}$. A function g is said to be* upper semi-continuous*(u.s.c.) if $-g$ is l.s.c.*

(a) *Show that if g is l.s.c. then $\{x : g(x) \leq b\}$ is closed for each $b \in \mathbb{R}$.*

(b) *Conclude that semi-continuous functions are Borel measurable.*

(c) *Conclude that continuous functions are Borel measurable.*

A concrete application of Corollary 1.2.19 shows that any linear combination of finitely many R.V.-s is a R.V.

EXAMPLE 1.2.21. *Suppose X_i are R.V.-s on the same measurable space and $c_i \in \mathbb{R}$. Then, $W_n(\omega) = \sum_{i=1}^n c_i X_i(\omega)$ are also R.V.-s. To see this, apply Corollary 1.2.19 for $f(x_1,\ldots,x_n) = \sum_{i=1}^n c_i x_i$ a continuous, hence Borel (measurable) function (by Exercise 1.2.20).*

We turn to explore the closure properties of $m\mathcal{F}$ with respect to operations of a limiting nature, starting with the following key theorem.

THEOREM 1.2.22. *Let $\overline{\mathbb{R}} = [-\infty,\infty]$ equipped with its Borel σ-algebra*

$$\mathcal{B}_{\overline{\mathbb{R}}} = \sigma\left([-\infty,b) : \ b \in \mathbb{R}\right).$$

If X_i are $\overline{\mathbb{R}}$-valued R.V.-s on the same measurable space, then

$$\inf_n X_n, \quad \sup_n X_n, \quad \liminf_{n\to\infty} X_n, \quad \limsup_{n\to\infty} X_n\,,$$

are also $\overline{\mathbb{R}}$-valued random variables.

PROOF. Pick an arbitrary $b \in \mathbb{R}$. Then,

$$\{\omega : \inf_n X_n(\omega) < b\} = \bigcup_{n=1}^{\infty} \{\omega : X_n(\omega) < b\} = \bigcup_{n=1}^{\infty} X_n^{-1}([-\infty,b)) \in \mathcal{F}.$$

Since $\mathcal{B}_{\overline{\mathbb{R}}}$ is generated by $\{[-\infty,b) : b \in \mathbb{R}\}$, it follows by Theorem 1.2.9 that $\inf_n X_n$ is an $\overline{\mathbb{R}}$-valued R.V.

Observing that $\sup_n X_n = -\inf_n(-X_n)$, we deduce from the above and Corollary 1.2.19 (for $f(x) = -x$), that $\sup_n X_n$ is also an $\overline{\mathbb{R}}$-valued R.V.

Next, recall that

$$W = \liminf_{n\to\infty} X_n = \sup_n \left[\inf_{l\geq n} X_l\right].$$

By the preceding proof we have that $Y_n = \inf_{l\geq n} X_l$ are $\overline{\mathbb{R}}$-valued R.V.-s and hence so is $W = \sup_n Y_n$.

Similarly to the arguments already used, we conclude the proof either by observing that

$$Z = \limsup_{n\to\infty} X_n = \inf_n \left[\sup_{l\geq n} X_l\right],$$

or by observing that $\limsup_n X_n = -\liminf_n(-X_n)$. □

REMARK. Since $\inf_n X_n$, $\sup_n X_n$, $\limsup_n X_n$ and $\liminf_n X_n$ may result in values $\pm\infty$ even when every X_n is $\mathbb{R}$-valued, hereafter we let $m\mathcal{F}$ also denote the collection of $\overline{\mathbb{R}}$-valued R.V.

An important corollary of this theorem deals with the existence of limits of sequences of R.V.

COROLLARY 1.2.23. *For any sequence $X_n \in m\mathcal{F}$, both*

$$\Omega_0 = \{\omega \in \Omega : \liminf_{n\to\infty} X_n(\omega) = \limsup_{n\to\infty} X_n(\omega)\}$$

and

$$\Omega_1 = \{\omega \in \Omega : \liminf_{n\to\infty} X_n(\omega) = \limsup_{n\to\infty} X_n(\omega) \in \mathbb{R}\}$$

are measurable sets, that is, $\Omega_0 \in \mathcal{F}$ and $\Omega_1 \in \mathcal{F}$.

Proof. By Theorem 1.2.22 we have that $Z = \limsup_n X_n$ and $W = \liminf_n X_n$ are two $\overline{\mathbb{R}}$-valued variables on the same space, with $Z(\omega) \geq W(\omega)$ for all ω. Hence, $\Omega_1 = \{\omega : Z(\omega) - W(\omega) = 0, Z(\omega) \in \mathbb{R}, W(\omega) \in \mathbb{R}\}$ is measurable (apply Corollary 1.2.19 for $f(z, w) = z - w$), as is $\Omega_0 = W^{-1}(\{\infty\}) \cup Z^{-1}(\{-\infty\}) \cup \Omega_1$. □

The following structural result is yet another consequence of Theorem 1.2.22.

Corollary 1.2.24. *For any $d < \infty$ and R.V.-s $Y_1, \dots, Y_d$ on the same measurable space $(\Omega, \mathcal{F})$ the collection $\mathcal{H} = \{h(Y_1, \dots, Y_d); h : \mathbb{R}^d \mapsto \mathbb{R}$ Borel function$\}$ is a vector space over $\mathbb{R}$ containing the constant functions, such that if $X_n \in \mathcal{H}$ are non-negative and $X_n \uparrow X$, an $\mathbb{R}$-valued function on Ω, then $X \in \mathcal{H}$.*

Proof. By Example 1.2.21 the collection of all Borel functions is a vector space over $\mathbb{R}$ which evidently contains the constant functions. Consequently, the same applies for $\mathcal{H}$. Next, suppose $X_n = h_n(Y_1, \dots, Y_d)$ for Borel functions h_n such that $0 \leq X_n(\omega) \uparrow X(\omega)$ for all $\omega \in \Omega$. Then, $\overline{h}(y) = \sup_n h_n(y)$ is by Theorem 1.2.22 an $\overline{\mathbb{R}}$-valued Borel function on $\mathbb{R}^d$, such that $X = \overline{h}(Y_1, \dots, Y_d)$. Setting $h(y) = \overline{h}(y)$ when $\overline{h}(y) \in \mathbb{R}$ and $h(y) = 0$ otherwise, it is easy to check that h is a real-valued Borel function. Moreover, with $X : \Omega \mapsto \mathbb{R}$ (finite valued), necessarily $X = h(Y_1, \dots, Y_d)$ as well, so $X \in \mathcal{H}$. □

The point-wise convergence of R.V., that is $X_n(\omega) \to X(\omega)$, for every $\omega \in \Omega$ is often too strong of a requirement, as it may fail to hold as a result of the R.V. being ill-defined for a *negligible set* of values of ω (that is, a set of zero measure). We thus define the more useful, weaker notion of almost sure convergence of random variables.

Definition 1.2.25. *We say that a sequence of random variables X_n on the same probability space $(\Omega, \mathcal{F}, \mathbf{P})$* converges almost surely *if $\mathbf{P}(\Omega_0) = 1$. We then set $X_\infty = \limsup_{n\to\infty} X_n$, and say that X_n converges almost surely to X_∞, or use the notation $X_n \overset{a.s.}{\to} X_\infty$.*

Remark. Note that in Definition 1.2.25 we allow the limit $X_\infty(\omega)$ to take the values $\pm\infty$ with positive probability. So, we say that X_n converges almost surely to a *finite limit* if $\mathbf{P}(\Omega_1) = 1$, or alternatively, if $X_\infty \in \mathbb{R}$ with probability one.

We proceed with an explicit characterization of the functions measurable with respect to a σ-algebra of the form $\sigma(Y_k, k \leq n)$.

Theorem 1.2.26. *Let $\mathcal{G} = \sigma(Y_k, k \leq n)$ for some $n < \infty$ and R.V.-s $Y_1, \dots, Y_n$ on the same measurable space $(\Omega, \mathcal{F})$. Then, $m\mathcal{G} = \{g(Y_1, \dots, Y_n) : g : \mathbb{R}^n \mapsto \mathbb{R}$ is a Borel function$\}$.*

Proof. From Corollary 1.2.19 we know that $Z = g(Y_1, \dots, Y_n)$ is in $m\mathcal{G}$ for each Borel function $g : \mathbb{R}^n \mapsto \mathbb{R}$. Turning to prove the converse result, recall part (b) of Exercise 1.2.14 that the σ-algebra $\mathcal{G}$ is generated by the π-system $\mathcal{P} = \{A_{\underline{\alpha}} : \underline{\alpha} = (\alpha_1, \dots, \alpha_n) \in \mathbb{R}^n\}$ where $I_{A_{\underline{\alpha}}} = h_{\underline{\alpha}}(Y_1, \dots, Y_n)$ for the Borel function $h_{\underline{\alpha}}(y_1, \dots, y_n) = \prod_{k=1}^n \mathbf{1}_{y_k \leq \alpha_k}$. Thus, in view of Corollary 1.2.24, we have by the monotone class theorem that $\mathcal{H} = \{g(Y_1, \dots, Y_n) : g : \mathbb{R}^n \mapsto \mathbb{R}$ is a Borel function$\}$ contains all elements of $m\mathcal{G}$. □

We conclude this sub-section with a few exercises, starting with Borel measurability of monotone functions (regardless of their continuity properties).

Exercise 1.2.27. *Show that any monotone function $g : \mathbb{R} \mapsto \mathbb{R}$ is Borel measurable.*

Next, Exercise 1.2.20 implies that the set of points at which a given function g is discontinuous, is a Borel set.

Exercise 1.2.28. *Fix an arbitrary function $g : \mathbb{S} \mapsto \mathbb{R}$.*

(a) *Show that for any $\delta > 0$ the function $g_*(x, \delta) = \inf\{g(y) : \rho(x, y) < \delta\}$ is u.s.c. and the function $g^*(x, \delta) = \sup\{g(y) : \rho(x, y) < \delta\}$ is l.s.c.*

(b) *Show that $\mathbf{D}_g = \{x : \sup_k g_*(x, k^{-1}) < \inf_k g^*(x, k^{-1})\}$ is exactly the set of points at which g is discontinuous.*

(c) *Deduce that the set $\mathbf{D}_g$ of points of discontinuity of g is a Borel set.*

Here is an alternative characterization of $\mathcal{B}$ that complements Exercise 1.2.20.

Exercise 1.2.29. *Show that if $\mathcal{F}$ is a σ-algebra of subsets of $\mathbb{R}$ then $\mathcal{B} \subseteq \mathcal{F}$ if and only if every continuous function $f : \mathbb{R} \mapsto \mathbb{R}$ is in $m\mathcal{F}$ (i.e. $\mathcal{B}$ is the smallest σ-algebra on $\mathbb{R}$ with respect to which all continuous functions are measurable).*

Exercise 1.2.30. *Suppose X_n and X_∞ are real-valued random variables and*

$$\mathbf{P}(\{\omega : \limsup_{n\to\infty} X_n(\omega) \le X_\infty(\omega)\}) = 1\,.$$

Show that for any $\varepsilon > 0$, there exists an event A with $\mathbf{P}(A) < \varepsilon$ and a non-random $N = N(\epsilon)$, sufficiently large such that $X_n(\omega) < X_\infty(\omega) + \varepsilon$ for all $n \ge N$ and every $\omega \in A^c$.

Equipped with Theorem 1.2.22 you can also strengthen Proposition 1.2.6.

Exercise 1.2.31. *Show that the class $m\mathcal{F}$ of $\overline{\mathbb{R}}$-valued measurable functions, is the smallest class containing* SF *and closed under point-wise limits.*

Finally, relying on Theorem 1.2.26 it is easy to show that a Borel function can only reduce the amount of information quantified by the corresponding generated σ-algebras, whereas such information content is invariant under invertible Borel transformations, that is

Exercise 1.2.32. *Show that $\sigma(g(Y_1, \ldots, Y_n)) \subseteq \sigma(Y_k, k \le n)$ for any Borel function $g : \mathbb{R}^n \mapsto \mathbb{R}$. Further, if $Y_1, \ldots, Y_n$ and $Z_1, \ldots, Z_m$ defined on the same probability space are such that $Z_k = g_k(Y_1, \ldots, Y_n)$, $k = 1, \ldots, m$ and $Y_i = h_i(Z_1, \ldots, Z_m)$, $i = 1, \ldots, n$ for some Borel functions $g_k : \mathbb{R}^n \mapsto \mathbb{R}$ and $h_i : \mathbb{R}^m \mapsto \mathbb{R}$, then $\sigma(Y_1, \ldots, Y_n) = \sigma(Z_1, \ldots, Z_m)$.*

1.2.3. Distribution, density and law. As defined next, every random variable X induces a probability measure on its range which is called the law of X.

Definition 1.2.33. *The* law *of a real-valued R.V. X, denoted $\mathcal{P}_X$, is the probability measure on $(\mathbb{R}, \mathcal{B})$ such that $\mathcal{P}_X(B) = \mathbf{P}(\{\omega : X(\omega) \in B\})$ for any Borel set B.*

Remark. Since X is a R.V., it follows that $\mathcal{P}_X(B)$ is well defined for all $B \in \mathcal{B}$. Further, the non-negativity of $\mathbf{P}$ implies that $\mathcal{P}_X$ is a non-negative set function on $(\mathbb{R}, \mathcal{B})$, and since $X^{-1}(\mathbb{R}) = \Omega$, also $\mathcal{P}_X(\mathbb{R}) = 1$. Consider next disjoint Borel sets B_i, observing that $X^{-1}(B_i) \in \mathcal{F}$ are disjoint subsets of Ω such that

$$X^{-1}(\bigcup_i B_i) = \bigcup_i X^{-1}(B_i)\,.$$

Thus, by the countable additivity of $\mathbf{P}$ we have that

$$\mathcal{P}_X(\bigcup_i B_i) = \mathbf{P}(\bigcup_i X^{-1}(B_i)) = \sum_i \mathbf{P}(X^{-1}(B_i)) = \sum_i \mathcal{P}_X(B_i) .$$

This shows that $\mathcal{P}_X$ is also countably additive, hence a probability measure, as claimed in Definition 1.2.33.

Note that the law $\mathcal{P}_X$ of a R.V. $X : \Omega \longrightarrow \mathbb{R}$, determines the values of the probability measure $\mathbf{P}$ on $\sigma(X)$.

Definition 1.2.34. *We write $X \overset{\mathcal{D}}{=} Y$ and say that X equals Y in law (or in distribution), if and only if $\mathcal{P}_X = \mathcal{P}_Y$.*

A good way to practice your understanding of the Definitions 1.2.33 and 1.2.34 is by verifying that if $X \overset{a.s.}{=} Y$, then also $X \overset{\mathcal{D}}{=} Y$ (that is, any two random variables we consider to be the same would indeed have the same law).

The next concept we define, the distribution function, is closely associated with the law $\mathcal{P}_X$ of the R.V.

Definition 1.2.35. *The* distribution function F_X *of a real-valued R.V. X is*

$$F_X(\alpha) = \mathbf{P}(\{\omega : X(\omega) \le \alpha\}) = \mathcal{P}_X((-\infty, \alpha]) \qquad \forall \alpha \in \mathbb{R}$$

Our next result characterizes the set of all functions $F : \mathbb{R} \mapsto [0, 1]$ that are distribution functions of some R.V.

Theorem 1.2.36. *A function $F : \mathbb{R} \mapsto [0, 1]$ is a distribution function of some R.V. if and only if*

(a) *F is non-decreasing*

(b) *$\lim_{x\to\infty} F(x) = 1$ and $\lim_{x\to-\infty} F(x) = 0$*

(c) *F is right-continuous, i.e. $\lim_{y\downarrow x} F(y) = F(x)$*

Proof. First, assuming that $F = F_X$ is a distribution function, we show that it must have the stated properties (a)-(c). Indeed, if $x \le y$ then $(-\infty, x] \subseteq (-\infty, y]$, and by the monotonicity of the probability measure $\mathcal{P}_X$ (see part (a) of Exercise 1.1.4), we have that $F_X(x) \le F_X(y)$, proving that F_X is non-decreasing. Further, $(-\infty, x] \uparrow \mathbb{R}$ as $x \uparrow \infty$, while $(-\infty, x] \downarrow \emptyset$ as $x \downarrow -\infty$, resulting with property (b) of the theorem by the continuity from below and the continuity from above of the probability measure $\mathcal{P}_X$ on $\mathbb{R}$. Similarly, since $(-\infty, y] \downarrow (-\infty, x]$ as $y \downarrow x$ we get the right continuity of F_X by yet another application of continuity from above of $\mathcal{P}_X$.

We proceed to prove the converse result, that is, assuming F has the stated properties (a)-(c), we consider the random variable $X^-(\omega) = \sup\{y : F(y) < \omega\}$ on the probability space $((0, 1], \mathcal{B}_{(0,1]}, U)$ and show that $F_{X^-} = F$. With F having property (b), we see that for any $\omega > 0$ the set $\{y : F(y) < \omega\}$ is non-empty and further if $\omega < 1$ then $X^-(\omega) < \infty$, so $X^- : (0, 1) \mapsto \mathbb{R}$ is well defined. The identity

$$\{\omega : X^-(\omega) \le x\} = \{\omega : \omega \le F(x)\} , \tag{1.2.1}$$

implies that $F_{X^-}(x) = U((0, F(x)]) = F(x)$ for all $x \in \mathbb{R}$, and further, the sets $(0, F(x)]$ are all in $\mathcal{B}_{(0,1]}$, implying that X^- is a measurable function (i.e. a R.V.).

Turning to prove (1.2.1) note that if $\omega \le F(x)$ then $x \notin \{y : F(y) < \omega\}$ and so by definition (and the monotonicity of F), $X^-(\omega) \le x$. Now suppose that $\omega > F(x)$. Since F is right continuous, this implies that $F(x + \epsilon) < \omega$ for some $\epsilon > 0$, hence by definition of X^- also $X^-(\omega) \geq x + \epsilon > x$, completing the proof of (1.2.1) and with it the proof of the theorem. $\square$

Check your understanding of the preceding proof by showing that the collection of distribution functions for $\overline{\mathbb{R}}$-valued random variables consist of all $F : \mathbb{R} \mapsto [0, 1]$ that are non-decreasing and right-continuous.

Remark. The construction of the random variable $X^-(\omega)$ in Theorem 1.2.36 is called *Skorokhod's representation.* You can, and should, verify that the random variable $X^+(\omega) = \sup\{y : F(y) \leq \omega\}$ would have worked equally well for that purpose, since $X^+(\omega) \neq X^-(\omega)$ only if $X^+(\omega) > q \geq X^-(\omega)$ for some rational q, in which case by definition $\omega \geq F(q) \geq \omega$, so there are most countably many such values of ω (hence $\mathbf{P}(X^+ \neq X^-) = 0$). We shall return to this construction when dealing with convergence in distribution in Section 3.2. An alternative approach to Theorem 1.2.36 is to adapt the construction of the probability measure of Example 1.1.26, taking here $\Omega = \mathbb{R}$ with the corresponding change to $\mathcal{A}$ and replacing the right side of (1.1.1) with $\sum_{k=1}^{r}(F(b_k) - F(a_k))$, yielding a probability measure $\mathcal{P}$ on $(\mathbb{R}, \mathcal{B})$ such that $\mathcal{P}((-\infty, \alpha]) = F(\alpha)$ for all $\alpha \in \mathbb{R}$ (c.f. [**Bil95**, Theorem 12.4]).

Our next example highlights the possible shape of the distribution function.

Example 1.2.37. *Consider Example 1.1.6 of n coin tosses, with σ-algebra $\mathcal{F}_n = 2^{\Omega_n}$, sample space $\Omega_n = \{H, T\}^n$, and the probability measure $\mathbf{P}_n(A) = \sum_{\omega \in A} p_\omega$, where $p_\omega = 2^{-n}$ for each $\omega \in \Omega_n$ (that is, $\omega = \{\omega_1, \omega_2, \cdots, \omega_n\}$ for $\omega_i \in \{H, T\}$), corresponding to independent, fair, coin tosses. Let $Y(\omega) = I_{\{\omega_1 = H\}}$ measure the outcome of the first toss. The law of this random variable is,*

$$\mathcal{P}_Y(B) = \frac{1}{2}\mathbf{1}_{\{0 \in B\}} + \frac{1}{2}\mathbf{1}_{\{1 \in B\}}$$

and its distribution function is

$$F_Y(\alpha) = \mathcal{P}_Y((-\infty, \alpha]) = \mathbf{P}_n(Y(\omega) \leq \alpha) = \begin{cases} 1, \ \alpha \geq 1 \\ \frac{1}{2}, \ 0 \leq \alpha < 1 \\ 0, \ \alpha < 0 \end{cases} . \tag{1.2.2}$$

Note that in general $\sigma(X)$ is a strict subset of the σ-algebra $\mathcal{F}$ (in Example 1.2.37 we have that $\sigma(Y)$ determines the probability measure for the first coin toss, but tells us nothing about the probability measure assigned to the remaining $n - 1$ tosses). Consequently, though the law $\mathcal{P}_X$ determines the probability measure $\mathbf{P}$ on $\sigma(X)$ it usually does not completely determine $\mathbf{P}$.

Example 1.2.37 is somewhat generic. That is, if the R.V. X is a simple function (or more generally, when the set $\{X(\omega) : \omega \in \Omega\}$ is countable and has no accumulation points), then its distribution function F_X is piecewise constant with jumps at the possible values that X takes and jump sizes that are the corresponding probabilities. Indeed, note that $(-\infty, y] \uparrow (-\infty, x)$ as $y \uparrow x$, so by the continuity from below of $\mathcal{P}_X$ it follows that

$$F_X(x^-) := \lim_{y \uparrow x} F_X(y) = \mathbf{P}(\{\omega : X(\omega) < x\}) = F_X(x) - \mathbf{P}(\{\omega : X(\omega) = x\}),$$

for any R.V. X.

A direct corollary of Theorem 1.2.36 shows that any distribution function has a collection of continuity points that is dense in $\mathbb{R}$.

Exercise 1.2.38. *Show that a distribution function F has at most countably many points of discontinuity and consequently, that for any $x \in \mathbb{R}$ there exist y_k and z_k at which F is continuous such that $z_k \downarrow x$ and $y_k \uparrow x$.*

In contrast with Example 1.2.37 the distribution function of a R.V. with a density is continuous and almost everywhere differentiable, that is,

Definition 1.2.39. *We say that a R.V. $X(\omega)$ has a* probability density function *f_X if and only if its distribution function F_X can be expressed as*

$$F_X(\alpha) = \int_{-\infty}^{\alpha} f_X(x)dx \,, \qquad \forall \alpha \in \mathbb{R} \,. \tag{1.2.3}$$

By Theorem 1.2.36 a probability density function f_X must be an integrable, Lebesgue almost everywhere non-negative function, with $\int_{\mathbb{R}} f_X(x)dx = 1$. Such F_X is continuous with $\frac{dF_X}{dx}(x) = f_X(x)$ except possibly on a set of values of x of zero Lebesgue measure.

Remark. To make Definition 1.2.39 precise we temporarily assume that probability density functions f_X are Riemann integrable and interpret the integral in (1.2.3) in this sense. In Section 1.3 we construct Lebesgue's integral and extend the scope of Definition 1.2.39 to *Lebesgue integrable* density functions $f_X \geq 0$ (in particular, accommodating Borel functions f_X). This is the setting we assume thereafter, with the right-hand-side of (1.2.3) interpreted as the integral $\overline{\lambda}(f_X; (-\infty, \alpha])$ of f_X with respect to the restriction on $(-\infty, \alpha]$ of the *completion* $\overline{\lambda}$ of the *Lebesgue measure* on $\mathbb{R}$ (c.f. Definition 1.3.59 and Example 1.3.60). Further, the function f_X is uniquely defined only as a representative of an equivalence class. That is, in this context we consider f and g to be the same function when $\overline{\lambda}(\{x : f(x) \neq g(x)\}) = 0$.

Building on Example 1.1.26 we next detail a few classical examples of R.V. that have densities.

Example 1.2.40. *The distribution function F_U of the R.V. of Example 1.1.26 is*

$$F_U(\alpha) = \mathbf{P}(U \leq \alpha) = \mathbf{P}(U \in [0, \alpha]) = \begin{cases} 1, \ \alpha > 1 \\ \alpha, \ 0 \leq \alpha \leq 1 \\ 0, \ \alpha < 0 \end{cases} \tag{1.2.4}$$

and its density is $f_U(u) = \begin{cases} 1, \ 0 \leq u \leq 1 \\ 0, \ \textit{otherwise} \end{cases}$.

The exponential distribution function *is*

$$F(x) = \begin{cases} 0, \ x \leq 0 \\ 1 - e^{-x}, \ x \geq 0 \end{cases} ,$$

corresponding to the density $f(x) = \begin{cases} 0, \ x \leq 0 \\ e^{-x}, \ x > 0 \end{cases}$ *, whereas the* standard normal distribution *has the density*

$$\phi(x) = (2\pi)^{-1/2} e^{-\frac{x^2}{2}} \,,$$

with no closed form expression for the corresponding distribution function $\Phi(x) = \int^x \phi(u)du$ in terms of elementary functions.

Every real-valued R.V. X has a distribution function but not necessarily a density. For example $X = 0$ w.p.1 has distribution function $F_X(\alpha) = \mathbf{1}_{\alpha \geq 0}$. Since F_X is discontinuous at 0, the R.V. X does not have a density.

Definition 1.2.41. *We say that a function F is a* Lebesgue singular function *if it has a zero derivative except on a set of zero Lebesgue measure.*

Since the distribution function of any R.V. is non-decreasing, from real analysis we know that it is almost everywhere differentiable. However, perhaps somewhat surprisingly, there are continuous distribution functions that are Lebesgue singular functions. Consequently, there are non-discrete random variables that do not have a density. We next provide one such example.

Example 1.2.42. *The* Cantor set *C is defined by removing $(1/3, 2/3)$ from $[0, 1]$ and then iteratively removing the middle third of each interval that remains. The* uniform distribution *on the (closed) set C corresponds to the distribution function obtained by setting $F(x) = 0$ for $x \leq 0$, $F(x) = 1$ for $x \geq 1$, $F(x) = 1/2$ for $x \in [1/3, 2/3]$, then $F(x) = 1/4$ for $x \in [1/9, 2/9]$, $F(x) = 3/4$ for $x \in [7/9, 8/9]$, and so on (which as you should check, satisfies the properties (a)-(c) of Theorem 1.2.36). From the definition, we see that $dF/dx = 0$ for almost every $x \notin C$ and that the corresponding probability measure has $\mathbf{P}(C^c) = 0$. As the Lebesgue measure of C is zero, we see that the derivative of F is zero except on a set of zero Lebesgue measure, and consequently, there is no function f for which $F(x) = \int_{-\infty}^{x} f(y)dy$ holds. Though it is somewhat more involved, you may want to check that F is everywhere continuous (c.f.* [**Bil95**, Problem 31.2]*).*

Even discrete distribution functions can be quite complex. As the next example shows, the points of discontinuity of such a function might form a (countable) dense subset of $\mathbb{R}$ (which in a sense is extreme, per Exercise 1.2.38).

Example 1.2.43. *Let $q_1, q_2, \ldots$ be an enumeration of the rational numbers and set*

$$F(x) = \sum_{i=1}^{\infty} 2^{-i} \mathbf{1}_{[q_i, \infty)}(x)$$

(where $\mathbf{1}_{[q_i,\infty)}(x) = 1$ if $x \geq q_i$ and zero otherwise). Clearly, such F is non-decreasing, with limits 0 and 1 as $x \to -\infty$ and $x \to \infty$, respectively. It is not hard to check that F is also right continuous, hence a distribution function, whereas by construction F is discontinuous at each rational number.

As we have that $\mathbf{P}(\{\omega : X(\omega) \leq \alpha\}) = F_X(\alpha)$ for the generators $\{\omega : X(\omega) \leq \alpha\}$ of $\sigma(X)$, we are not at all surprised by the following proposition.

Proposition 1.2.44. *The distribution function F_X uniquely determines the law $\mathcal{P}_X$ of X.*

Proof. Consider the collection $\pi(\mathbb{R}) = \{(-\infty, b] : b \in \mathbb{R}\}$ of subsets of $\mathbb{R}$. It is easy to see that $\pi(\mathbb{R})$ is a π-system, which generates $\mathcal{B}$ (see Exercise 1.1.17). Hence, by Proposition 1.1.39, any two probability measures on $(\mathbb{R}, \mathcal{B})$ that coincide on $\pi(\mathbb{R})$ are the same. Since the distribution function F_X specifies the restriction of such a probability measure $\mathcal{P}_X$ on $\pi(\mathbb{R})$ it thus uniquely determines the values of $\mathcal{P}_X(B)$ for all $B \in \mathcal{B}$. □

Different probability measures $\mathbf{P}$ on the measurable space $(\Omega, \mathcal{F})$ may “trivialize” different σ-algebras. That is, Definition 1.2.45. *If a σ-algebra $\mathcal{H} \subseteq \mathcal{F}$ and a probability measure $\mathbf{P}$ on $(\Omega, \mathcal{F})$ are such that $\mathbf{P}(H) \in \{0,1\}$ for all $H \in \mathcal{H}$, we call $\mathcal{H}$ a* $\mathbf{P}$-trivial *σ-algebra. Similarly, a random variable X is called* $\mathbf{P}$-trivial *or* $\mathbf{P}$-degenerate*, if there exists a non-random constant c such that $\mathbf{P}(X \neq c) = 0$.*

Using distribution functions we show next that all random variables on a $\mathbf{P}$-trivial σ-algebra are $\mathbf{P}$-trivial.

Proposition 1.2.46. *If a random variable $X \in m\mathcal{H}$ for a $\mathbf{P}$-trivial σ-algebra $\mathcal{H}$, then X is $\mathbf{P}$-trivial.*

Proof. By definition, the sets $\{\omega : X(\omega) \leq \alpha\}$ are in $\mathcal{H}$ for all $\alpha \in \mathbb{R}$. Since $\mathcal{H}$ is $\mathbf{P}$-trivial this implies that $F_X(\alpha) \in \{0,1\}$ for all $\alpha \in \mathbb{R}$. In view of Theorem 1.2.36 this is possible only if $F_X(\alpha) = \mathbf{1}_{\alpha \geq c}$ for some non-random $c \in \mathbb{R}$ (for example, set $c = \inf\{\alpha : F_X(\alpha) = 1\}$). That is, $\mathbf{P}(X \neq c) = 0$, as claimed. □

We conclude with few exercises about the support of measures on $(\mathbb{R}, \mathcal{B})$.

Exercise 1.2.47. *Let μ be a measure on $(\mathbb{R}, \mathcal{B})$. A point x is said to be in the* support *of μ if $\mu(O) > 0$ for every open neighborhood O of x. Prove that the support is a closed set whose complement is the maximal open set on which μ vanishes.*

Exercise 1.2.48. *Given an arbitrary closed set $C \subseteq \mathbb{R}$, construct a probability measure on $(\mathbb{R}, \mathcal{B})$ whose support is C.*
Hint: *Try a measure consisting of a countable collection of atoms (i.e. points of positive probability).*

As you are to check next, the discontinuity points of a distribution function are closely related to the support of the corresponding law.

Exercise 1.2.49. *The* support *of a distribution function F is the set $S_F = \{x \in \mathbb{R}$ such that $F(x+\epsilon) - F(x-\epsilon) > 0$ for all $\epsilon > 0\}$.*

(a) *Show that all points of discontinuity of $F(\cdot)$ belong to S_F, and that any isolated point of S_F (that is, $x \in S_F$ such that $(x-\delta, x+\delta) \cap S_F = \{x\}$ for some $\delta > 0$) must be a point of discontinuity of $F(\cdot)$.*

(b) *Show that the support of the law $\mathcal{P}_X$ of a random variable X, as defined in Exercise 1.2.47, is the same as the support of its distribution function F_X.*

1.3. Integration and the (mathematical) expectation

A key concept in probability theory is the mathematical expectation of random variables. In Subsection 1.3.1 we provide its definition via the framework of Lebesgue integration with respect to a measure and study properties such as monotonicity and linearity. In Subsection 1.3.2 we consider fundamental inequalities associated with the expectation. Subsection 1.3.3 is about the exchange of integration and limit operations, complemented by uniform integrability and its consequences in Subsection 1.3.4. Subsection 1.3.5 considers densities relative to arbitrary measures and relates our treatment of integration and expectation to Riemann's integral and the classical definition of the expectation for a R.V. with probability density. We conclude with Subsection 1.3.6 about moments of random variables, including their values for a few well known distributions.

1.3.1. Lebesgue integral, linearity and monotonicity. Let SF_+ denote the collection of non-negative simple functions with respect to the given measurable space $(\mathbb{S}, \mathcal{F})$ and $m\mathcal{F}_+$ denote the collection of $[0, \infty]$-valued measurable functions on this space. We next define Lebesgue's integral with respect to any measure μ on $(\mathbb{S}, \mathcal{F})$, first for $\varphi \in \mathrm{SF}_+$, then extending it to all $f \in m\mathcal{F}_+$. With the notation $\mu(f) := \int_{\mathbb{S}} f(s)d\mu(s)$ for this integral, we also denote by $\mu_0(\cdot)$ the more restrictive integral, defined only on SF_+, so as to clarify the role each of these plays in some of our proofs. We call an $\overline{\mathbb{R}}$-valued measurable function $f \in m\mathcal{F}$ for which $\mu(|f|) < \infty$, a *μ-integrable* function, and denote the collection of all μ-integrable functions by $L^1(\mathbb{S}, \mathcal{F}, \mu)$, extending the definition of the integral $\mu(f)$ to all $f \in L^1(\mathbb{S}, \mathcal{F}, \mu)$.

Definition 1.3.1. *Fix a measure space $(\mathbb{S}, \mathcal{F}, \mu)$ and define $\mu(f)$ by the following four step procedure:*

Step 1. *Define $\mu_0(I_A) := \mu(A)$ for each $A \in \mathcal{F}$.*

Step 2. *Any $\varphi \in \mathrm{SF}_+$ has a representation $\varphi = \sum_{l=1}^{n} c_l I_{A_l}$ for some finite $n < \infty$, non-random $c_l \in [0, \infty]$ and sets $A_l \in \mathcal{F}$, yielding the definition of the integral via*

$$\mu_0(\varphi) := \sum_{l=1}^{n} c_l \mu(A_l)\,,$$

where we adopt hereafter the convention that $\infty \times 0 = 0 \times \infty = 0$.

Step 3. *For $f \in m\mathcal{F}_+$ we define*

$$\mu(f) := \sup\{\mu_0(\varphi) : \varphi \in \mathrm{SF}_+, \varphi \le f\}.$$

Step 4. *For $f \in m\mathcal{F}$ let $f_+ = \max(f, 0) \in m\mathcal{F}_+$ and $f_- = -\min(f, 0) \in m\mathcal{F}_+$. We then set $\mu(f) = \mu(f_+) - \mu(f_-)$ provided either $\mu(f_+) < \infty$ or $\mu(f_-) < \infty$. In particular, this applies whenever $f \in L^1(\mathbb{S}, \mathcal{F}, \mu)$, for then $\mu(f_+) + \mu(f_-) = \mu(|f|)$ is finite, hence $\mu(f)$ is well defined and finite valued.*

We use the notation $\int_{\mathbb{S}} f(s)d\mu(s)$ for $\mu(f)$ which we call Lebesgue integral *of f with respect to the measure μ.*

The *expectation* $\mathbf{E}[X]$ of a random variable X on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ is merely Lebesgue's integral $\int X(\omega)d\mathbf{P}(\omega)$ of X with respect to $\mathbf{P}$. That is,
Step 1. $\mathbf{E}[I_A] = \mathbf{P}(A)$ for any $A \in \mathcal{F}$.

Step 2. Any $\varphi \in \mathrm{SF}_+$ has a representation $\varphi = \sum_{l=1}^{n} c_l I_{A_l}$ for some non-random $n < \infty$, $c_l \ge 0$ and sets $A_l \in \mathcal{F}$, to which corresponds

$$\mathbf{E}[\varphi] = \sum_{l=1}^{n} c_l \mathbf{E}[I_{A_l}] = \sum_{l=1}^{n} c_l \mathbf{P}(A_l)\,.$$

Step 3. For $X \in m\mathcal{F}_+$ define

$$\mathbf{E}X = \sup\{\mathbf{E}Y : Y \in \mathrm{SF}_+, Y \le X\}.$$

Step 4. Represent $X \in m\mathcal{F}$ as $X = X_+ - X_-$, where $X_+ = \max(X, 0) \in m\mathcal{F}_+$ and $X_- = -\min(X, 0) \in m\mathcal{F}_+$, with the corresponding definition

$$\mathbf{E}X = \mathbf{E}X_+ - \mathbf{E}X_-\,,$$

provided either $\mathbf{E}X_+ < \infty$ or $\mathbf{E}X_- < \infty$.

Remark. Note that we may have $\mathbf{E}X = \infty$ while $X(\omega) < \infty$ for all ω. For instance, take the random variable $X(\omega) = \omega$ for $\Omega = \{1, 2, \ldots\}$ and $\mathcal{F} = 2^\Omega$. If $\mathbf{P}(\omega = k) = ck^{-2}$ with $c = [\sum_{k=1}^\infty k^{-2}]^{-1}$ a positive, finite normalization constant, then $\mathbf{E}X = c\sum_{k=1}^\infty k^{-1} = \infty$.

Similar to the notation of μ-integrable functions introduced in the last step of the definition of Lebesgue's integral, we have the following definition for random variables.

Definition 1.3.2. *We say that a random variable X is (absolutely)* integrable*, or X has finite expectation, if $\mathbf{E}|X| < \infty$, that is, both $\mathbf{E}X_+ < \infty$ and $\mathbf{E}X_- < \infty$. Fixing $1 \le q < \infty$ we denote by $L^q(\Omega, \mathcal{F}, \mathbf{P})$ the collection of random variables X on $(\Omega, \mathcal{F})$ for which $||X||_q = [\mathbf{E}|X|^q]^{1/q} < \infty$. For example, $L^1(\Omega, \mathcal{F}, \mathbf{P})$ denotes the space of all (absolutely) integrable random-variables. We use the short notation L^q when the probability space $(\Omega, \mathcal{F}, \mathbf{P})$ is clear from the context.*

We next verify that Lebesgue's integral of each function f is assigned a unique value in Definition 1.3.1. To this end, we focus on $\mu_0 : \mathrm{SF}_+ \mapsto [0, \infty]$ of Step 2 of our definition and derive its structural properties, such as monotonicity, linearity and invariance to a change of argument on a μ-negligible set.

Lemma 1.3.3. *$\mu_0(\varphi)$ assigns a unique value to each $\varphi \in \mathrm{SF}_+$. Further,*
a). $\mu_0(\varphi) = \mu_0(\psi)$ if $\varphi, \psi \in \mathrm{SF}_+$ are such that $\mu(\{s : \varphi(s) \neq \psi(s)\}) = 0$.
b). μ_0 is linear, that is

$$\mu_0(\varphi + \psi) = \mu_0(\varphi) + \mu_0(\psi)\,, \qquad\qquad \mu_0(c\varphi) = c\mu_0(\varphi)\,,$$

for any $\varphi, \psi \in \mathrm{SF}_+$ and $c \ge 0$.
c). μ_0 is monotone, that is $\mu_0(\varphi) \le \mu_0(\psi)$ if $\varphi(s) \le \psi(s)$ for all $s \in \mathbb{S}$.

Proof. Note that a non-negative simple function $\varphi \in \mathrm{SF}_+$ has many different representations as weighted sums of indicator functions. Suppose for example that

$$\sum_{l=1}^n c_l I_{A_l}(s) = \sum_{k=1}^m d_k I_{B_k}(s)\,, \tag{1.3.1}$$

for some $c_l \ge 0$, $d_k \ge 0$, $A_l \in \mathcal{F}$, $B_k \in \mathcal{F}$ and all $s \in \mathbb{S}$. There exists a finite partition of $\mathbb{S}$ to at most 2^{n+m} disjoint sets C_i such that each of the sets A_l and B_k is a union of some C_i, $i = 1, \ldots, 2^{n+m}$. Expressing both sides of (1.3.1) as finite weighted sums of I_{C_i}, we necessarily have for each i the same weight on both sides. Due to the (finite) additivity of μ over unions of disjoint sets C_i, we thus get after some algebra that

$$\sum_{l=1}^n c_l \mu(A_l) = \sum_{k=1}^m d_k \mu(B_k)\,. \tag{1.3.2}$$

Consequently, $\mu_0(\varphi)$ is well-defined and independent of the chosen representation for φ. Further, the conclusion (1.3.2) applies also when the two sides of (1.3.1) differ for $s \in C$ as long as $\mu(C) = 0$, hence proving the first stated property of the lemma.

Choosing the representation of $\varphi + \psi$ based on the representations of φ and ψ immediately results with the stated linearity of μ_0. Given this, if $\varphi(s) \le \psi(s)$ for all s, then $\psi = \varphi + \xi$ for some $\xi \in \mathrm{SF}_+$, implying that $\mu_0(\psi) = \mu_0(\varphi) + \mu_0(\xi) \ge \mu_0(\varphi)$, as claimed. □

Remark. The stated monotonicity of μ_0 implies that $\mu(\cdot)$ coincides with $\mu_0(\cdot)$ on SF_+. As μ_0 is uniquely defined for each $f \in \mathrm{SF}_+$ and $f = f_+$ when $f \in m\mathcal{F}_+$, it follows that $\mu(f)$ is uniquely defined for each $f \in m\mathcal{F}_+ \cup L^1(\mathbb{S}, \mathcal{F}, \mu)$.

All three properties of μ_0 (hence μ) stated in Lemma 1.3.3 for functions in SF_+ extend to all of $m\mathcal{F}_+ \cup L^1$. Indeed, the facts that $\mu(cf) = c\mu(f)$, that $\mu(f) \leq \mu(g)$ whenever $0 \leq f \leq g$, and that $\mu(f) = \mu(g)$ whenever $\mu(\{s : f(s) \neq g(s)\}) = 0$ are immediate consequences of our definition (once we have these for $f, g \in \mathrm{SF}_+$). Since $f \leq g$ implies $f_+ \leq g_+$ and $f_- \geq g_-$, the monotonicity of $\mu(\cdot)$ extends to functions in L^1 (by Step 4 of our definition). To prove that $\mu(h + g) = \mu(h) + \mu(g)$ for all $h, g \in m\mathcal{F}_+ \cup L^1$ requires an application of the *monotone convergence theorem* (in short MON), which we now state, while deferring its proof to Subsection 1.3.3.

Theorem 1.3.4 (Monotone convergence theorem). *If $0 \leq h_n(s) \uparrow h(s)$ for all $s \in \mathbb{S}$ and $h_n \in m\mathcal{F}_+$, then $\mu(h_n) \uparrow \mu(h) \leq \infty$.*

Indeed, recall that while proving Proposition 1.2.6 we constructed the sequence f_n such that for every $g \in m\mathcal{F}_+$ we have $f_n(g) \in \mathrm{SF}_+$ and $f_n(g) \uparrow g$. Specifying $g, h \in m\mathcal{F}_+$ we have that $f_n(h) + f_n(g) \in \mathrm{SF}_+$. So, by Lemma 1.3.3,

$$\mu(f_n(h)+f_n(g)) = \mu_0(f_n(h)+f_n(g)) = \mu_0(f_n(h))+\mu_0(f_n(g)) = \mu(f_n(h))+\mu(f_n(g)) .$$

Since $f_n(h) \uparrow h$ and $f_n(h) + f_n(g) \uparrow h + g$, by monotone convergence,

$$\mu(h + g) = \lim_{n\to\infty} \mu(f_n(h) + f_n(g)) = \lim_{n\to\infty} \mu(f_n(h)) + \lim_{n\to\infty} \mu(f_n(g)) = \mu(h) + \mu(g) .$$

To extend this result to $g, h \in m\mathcal{F}_+ \cup L^1$, note that $h_- + g_- = f + (h+g)_- \geq f$ for some $f \in m\mathcal{F}_+$ such that $h_+ + g_+ = f + (h+g)_+$. Since $\mu(h_-) < \infty$ and $\mu(g_-) < \infty$, by linearity and monotonicity of $\mu(\cdot)$ on $m\mathcal{F}_+$ necessarily also $\mu(f) < \infty$ and the linearity of $\mu(h+g)$ on $m\mathcal{F}_+ \cup L^1$ follows by elementary algebra. In conclusion, we have just proved that

Proposition 1.3.5. *The integral $\mu(f)$ assigns a unique value to each $f \in m\mathcal{F}_+ \cup L^1(\mathbb{S}, \mathcal{F}, \mu)$. Further,*
a). $\mu(f) = \mu(g)$ whenever $\mu(\{s : f(s) \neq g(s)\}) = 0$.
b). μ is linear, that is for any $f, h, g \in m\mathcal{F}_+ \cup L^1$ and $c \geq 0$,

$$\mu(h + g) = \mu(h) + \mu(g) , \qquad \mu(cf) = c\mu(f) .$$

c). μ is monotone, that is $\mu(f) \leq \mu(g)$ if $f(s) \leq g(s)$ for all $s \in \mathbb{S}$.

Our proof of the identity $\mu(h + g) = \mu(h) + \mu(g)$ is an example of the following general approach to proving that certain properties hold for all $h \in L^1$.

Definition 1.3.6 (Standard Machine). *To prove the validity of a certain property for all $h \in L^1(\mathbb{S}, \mathcal{F}, \mu)$, break your proof to four easier steps, following those of Definition 1.3.1.*
Step 1. *Prove the property for h which is an indicator function.*
Step 2. *Using linearity, extend the property to all* SF_+.
Step 3. *Using MON extend the property to all $h \in m\mathcal{F}_+$.*
Step 4. *Extend the property in question to $h \in L^1$ by writing $h = h_+ - h_-$ and using linearity.*

Here is another application of the standard machine.

EXERCISE 1.3.7. *Suppose that a probability measure $\mathcal{P}$ on $(\mathbb{R}, \mathcal{B})$ is such that $\mathcal{P}(B) = \lambda(f I_B)$ for the Lebesgue measure λ on $\mathbb{R}$, some non-negative Borel function $f(\cdot)$ and all $B \in \mathcal{B}$. Using the standard machine, prove that then $\mathcal{P}(h) = \lambda(fh)$ for any Borel function h such that either $h \geq 0$ or $\lambda(f|h|) < \infty$.*
Hint: *See the proof of Proposition 1.3.56.*

We shall see more applications of the standard machine later (for example, when proving Proposition 1.3.56 and Theorem 1.3.61).

We next strengthen the non-negativity and monotonicity properties of Lebesgue's integral $\mu(\cdot)$ by showing that

LEMMA 1.3.8. *If $\mu(h) = 0$ for $h \in m\mathcal{F}_+$, then $\mu(\{s : h(s) > 0\}) = 0$. Consequently, if for $f, g \in L^1(\mathbb{S}, \mathcal{F}, \mu)$ both $\mu(f) = \mu(g)$ and $\mu(\{s : f(s) > g(s)\}) = 0$, then $\mu(\{s : f(s) \neq g(s)\}) = 0$.*

PROOF. By continuity below of the measure μ we have that

$$\mu(\{s : h(s) > 0\}) = \lim_{n\to\infty} \mu(\{s : h(s) > n^{-1}\})$$

(see Exercise 1.1.4). Hence, if $\mu(\{s : h(s) > 0\}) > 0$, then for some $n < \infty$,

$$0 < n^{-1}\mu(\{s : h(s) > n^{-1}\}) = \mu_0(n^{-1} I_{h>n^{-1}}) \leq \mu(h) ,$$

where the right most inequality is a consequence of the definition of $\mu(h)$ and the fact that $h \geq n^{-1} I_{h>n^{-1}} \in \mathrm{SF}_+$. Thus, our assumption that $\mu(h) = 0$ must imply that $\mu(\{s : h(s) > 0\}) = 0$.

To prove the second part of the lemma, consider $\widetilde{h} = g - f$ which is non-negative outside a set $N \in \mathcal{F}$ such that $\mu(N) = 0$. Hence, $h = (g - f)I_{N^c} \in m\mathcal{F}_+$ and $0 = \mu(g) - \mu(f) = \mu(\widetilde{h}) = \mu(h)$ by Proposition 1.3.5, implying that $\mu(\{s : h(s) > 0\}) = 0$ by the preceding proof. The same applies for $\widetilde{h}$ and the statement of the lemma follows. □

We conclude this subsection by stating the results of Proposition 1.3.5 and Lemma 1.3.8 in terms of the expectation on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$.

THEOREM 1.3.9. *The mathematical expectation $\mathbf{E}[X]$ is well defined for every R.V. X on $(\Omega, \mathcal{F}, \mathbf{P})$ provided either $X \geq 0$ almost surely, or $X \in L^1(\Omega, \mathcal{F}, \mathbf{P})$. Further,*
(a) $\mathbf{E}X = \mathbf{E}Y$ whenever $X \overset{a.s.}{=} Y$.
(b) The expectation is a linear *operation, for if Y and Z are integrable R.V. then for any constants α, β the R.V. $\alpha Y + \beta Z$ is integrable and $\mathbf{E}(\alpha Y + \beta Z) = \alpha(\mathbf{E}Y) + \beta(\mathbf{E}Z)$. The same applies when $Y, Z \geq 0$ almost surely and $\alpha, \beta \geq 0$.*
(c) The expectation is monotone. *That is, if Y and Z are either integrable or non-negative and $Y \geq Z$ almost surely, then $\mathbf{E}Y \geq \mathbf{E}Z$. Further, if Y and Z are integrable with $Y \geq Z$ a.s. and $\mathbf{E}Y = \mathbf{E}Z$, then $Y \overset{a.s.}{=} Z$.*
(d) Constants are invariant under the expectation. That is, if $X \overset{a.s.}{=} c$ for non-random $c \in (-\infty, \infty]$, then $\mathbf{E}X = c$.

REMARK. Part (d) of the theorem relies on the fact that $\mathbf{P}$ is a probability measure, namely $\mathbf{P}(\Omega) = 1$. Indeed, it is obtained by considering the expectation of the simple function cI_Ω to which X equals with probability one.

The linearity of the expectation (i.e. part (b) of the preceding theorem), is often extremely helpful when looking for an explicit formula for it. We next provide a few examples of this.

EXERCISE 1.3.10. *Write* $(\Omega, \mathcal{F}, \mathbf{P})$ *for a random experiment whose outcome is a recording of the results of* n *independent rolls of a balanced six-sided dice (including their order). Compute the expectation of the random variable* $D(\omega)$ *which counts the number of different faces of the dice recorded in these* n *rolls.*

EXERCISE 1.3.11 (MATCHING). *In a random* matching *experiment, we apply a random permutation* π *to the integers* $\{1, 2, \ldots, n\}$, *where each of the possible* $n!$ *permutations is equally likely. Let* $Z_i = I_{\{\pi(i)=i\}}$ *be the random variable indicating whether* $i = 1, 2, \ldots, n$ *is a fixed point of the random permutation, and* $X_n = \sum_{i=1}^{n} Z_i$ *count the number of fixed points of the random permutation (i.e. the number of self-matchings). Show that* $\mathbf{E}[X_n(X_n - 1) \cdots (X_n - k + 1)] = 1$ *for* $k = 1, 2, \ldots, n$.

Similarly, here is an elementary application of the monotonicity of the expectation (i.e. part (c) of the preceding theorem).

EXERCISE 1.3.12. *Suppose an integrable random variable* X *is such that* $\mathbf{E}(XI_A) = 0$ *for each* $A \in \sigma(X)$. *Show that necessarily* $X = 0$ *almost surely.*

1.3.2. Inequalities.

The linearity of the expectation often allows us to compute $\mathbf{E}X$ even when we cannot compute the distribution function F_X. In such cases the expectation can be used to bound tail probabilities, based on the following classical inequality.

THEOREM 1.3.13 (MARKOV'S INEQUALITY). *Suppose* $\psi : \mathbb{R} \mapsto [0, \infty]$ *is a Borel function and let* $\psi_*(A) = \inf\{\psi(y) : y \in A\}$ *for any* $A \in \mathcal{B}$. *Then for any R.V.* X,

$$\psi_*(A)\mathbf{P}(X \in A) \leq \mathbf{E}(\psi(X)I_{X \in A}) \leq \mathbf{E}\psi(X).$$

PROOF. By the definition of $\psi_*(A)$ and non-negativity of ψ we have that

$$\psi_*(A)I_{x \in A} \leq \psi(x)I_{x \in A} \leq \psi(x)\,,$$

for all $x \in \mathbb{R}$. Therefore, $\psi_*(A)I_{X \in A} \leq \psi(X)I_{X \in A} \leq \psi(X)$ for every $\omega \in \Omega$. We deduce the stated inequality by the monotonicity of the expectation and the identity $\mathbf{E}(\psi_*(A)I_{X \in A}) = \psi_*(A)\mathbf{P}(X \in A)$ (due to Step 2 of Definition 1.3.1). □

We next specify three common instances of Markov's inequality.

EXAMPLE 1.3.14. *(a). Taking* $\psi(x) = x_+$ *and* $A = [a, \infty)$ *for some* $a > 0$ *we have that* $\psi_*(A) = a$. *Markov's inequality is then*

$$\mathbf{P}(X \geq a) \leq \frac{\mathbf{E}X_+}{a}\,,$$

which is particularly appealing when $X \geq 0$, *so* $\mathbf{E}X_+ = \mathbf{E}X$.
(b). Taking $\psi(x) = |x|^q$ *and* $A = (-\infty, -a] \cup [a, \infty)$ *for some* $a > 0$, *we get that* $\psi_*(A) = a^q$. *Markov's inequality is then* $a^q\mathbf{P}(|X| \geq a) \leq \mathbf{E}|X|^q$. *Considering* $q = 2$ *and* $X = Y - \mathbf{E}Y$ *for* $Y \in L^2$, *this amounts to*

$$\mathbf{P}(|Y - \mathbf{E}Y| \geq a) \leq \frac{\mathsf{Var}(Y)}{a^2}\,,$$

which we call Chebyshev's inequality *(c.f. Definition 1.3.67 for the variance and moments of random variable* Y*).*
(c). Taking $\psi(x) = e^{\theta x}$ *for some* $\theta > 0$ *and* $A = [a, \infty)$ *for some* $a \in \mathbb{R}$ *we have that* $\psi_*(A) = e^{\theta a}$. *Markov's inequality is then*

$$\mathbf{P}(X \geq a) \leq e^{-\theta a}\mathbf{E}e^{\theta X}.$$

This bound provides an exponential decay in a, at the cost of requiring X to have finite exponential moments.

In general, we cannot compute $\mathbf{E}X$ explicitly from the Definition 1.3.1 except for discrete R.V.s and for R.V.s having a probability density function. We thus appeal to the properties of the expectation listed in Theorem 1.3.9, or use various inequalities to bound one expectation by another. To this end, we start with Jensen's inequality, dealing with the effect that a convex function makes on the expectation.

Proposition 1.3.15 (Jensen's inequality). *Suppose $g(\cdot)$ is a convex function on an open interval G of $\mathbb{R}$, that is,*

$$\lambda g(x) + (1-\lambda)g(y) \geq g(\lambda x + (1-\lambda)y) \qquad \forall\ x, y \in G, \qquad 0 \leq \lambda \leq 1.$$

If X is an integrable R.V. with $\mathbf{P}(X \in G) = 1$ and $g(X)$ is also integrable, then $\mathbf{E}(g(X)) \geq g(\mathbf{E}X)$.

Proof. The convexity of $g(\cdot)$ on G implies that $g(\cdot)$ is continuous on G (hence $g(X)$ is a random variable) and the existence for each $c \in G$ of $b = b(c) \in \mathbb{R}$ such that

$$g(x) \geq g(c) + b(x-c), \qquad \forall x \in G\,. \tag{1.3.3}$$

Since G is an open interval of $\mathbb{R}$ with $\mathbf{P}(X \in G) = 1$ and X is integrable, it follows that $\mathbf{E}X \in G$. Assuming (1.3.3) holds for $c = \mathbf{E}X$, that $X \in G$ a.s., and that both X and $g(X)$ are integrable, we have by Theorem 1.3.9 that

$$\mathbf{E}(g(X)) = \mathbf{E}(g(X)I_{X\in G}) \geq \mathbf{E}[(g(c)+b(X-c))I_{X\in G}] = g(c)+b(\mathbf{E}X-c) = g(\mathbf{E}X)\,,$$

as stated. To derive (1.3.3) note that if $(c-h_2, c+h_1) \subseteq G$ for positive h_1 and h_2, then by convexity of $g(\cdot)$,

$$\frac{h_2}{h_1+h_2} g(c+h_1) + \frac{h_1}{h_1+h_2} g(c-h_2) \geq g(c)\,,$$

which amounts to $[g(c+h_1) - g(c)]/h_1 \geq [g(c) - g(c-h_2)]/h_2$. Considering the infimum over $h_1 > 0$ and the supremum over $h_2 > 0$ we deduce that

$$\inf_{h>0, c+h\in G} \frac{g(c+h)-g(c)}{h} := (D_+g)(c) \geq (D_-g)(c) := \sup_{h>0, c-h\in G} \frac{g(c)-g(c-h)}{h}\,.$$

With G an open set, obviously $(D_-g)(x) > -\infty$ and $(D_+g)(x) < \infty$ for any $x \in G$ (in particular, $g(\cdot)$ is continuous on G). Now for any $b \in [(D_-g)(c), (D_+g)(c)] \subset \mathbb{R}$ we get (1.3.3) out of the definition of D_+g and D_-g. □

Remark. Since $g(\cdot)$ is convex if and only if $-g(\cdot)$ is concave, we may as well state Jensen's inequality for concave functions, just reversing the sign of the inequality in this case. A trivial instance of Jensen's inequality happens when $X(\omega) = xI_A(\omega) + yI_{A^c}(\omega)$ for some $x, y \in \mathbb{R}$ and $A \in \mathcal{F}$ such that $\mathbf{P}(A) = \lambda$. Then,

$$\mathbf{E}X = x\mathbf{P}(A) + y\mathbf{P}(A^c) = x\lambda + y(1-\lambda)\,,$$

whereas $g(X(\omega)) = g(x)I_A(\omega) + g(y)I_{A^c}(\omega)$. So,

$$\mathbf{E}g(X) = g(x)\lambda + g(y)(1-\lambda) \geq g(x\lambda + y(1-\lambda)) = g(\mathbf{E}X)\,,$$

as g is convex.

Applying Jensen's inequality, we show that the spaces $L^q(\Omega, \mathcal{F}, \mathbf{P})$ of Definition 1.3.2 are nested in terms of the parameter $q \geq 1$.

LEMMA 1.3.16. *Fixing $Y \in m\mathcal{F}$, the mapping $q \mapsto ||Y||_q = [\mathbf{E}|Y|^q]^{1/q}$ is non-decreasing for $q > 0$. Hence, the space $L^q(\Omega, \mathcal{F}, \mathbf{P})$ is contained in $L^r(\Omega, \mathcal{F}, \mathbf{P})$ for any $r \leq q$.*

PROOF. Fix $q > r > 0$ and consider the sequence of bounded R.V. $X_n(\omega) = \{\min(|Y(\omega)|, n)\}^r$. Obviously, X_n and $X_n^{q/r}$ are both in L^1. Apply Jensen's Inequality for the convex function $g(x) = |x|^{q/r}$ and the non-negative R.V. X_n, to get that

$$(\mathbf{E}X_n)^{\frac{q}{r}} \leq \mathbf{E}(X_n^{\frac{q}{r}}) = \mathbf{E}[\{\min(|Y|, n)\}^q] \leq \mathbf{E}(|Y|^q) \, .$$

For $n \uparrow \infty$ we have that $X_n \uparrow |Y|^r$, so by monotone convergence $\mathbf{E}\left(|Y|^r\right)^{\frac{q}{r}} \leq (\mathbf{E}|Y|^q)$. Taking the $1/q$-th power yields the stated result $||Y||_r \leq ||Y||_q \leq \infty$. □

We next bound the expectation of the product of two R.V. while assuming nothing about the relation between them.

PROPOSITION 1.3.17 (HÖLDER'S INEQUALITY). *Let X, Y be two random variables on the same probability space. If $p, q > 1$ with $\frac{1}{p} + \frac{1}{q} = 1$, then*

$$\mathbf{E}|XY| \leq ||X||_p ||Y||_q \, . \tag{1.3.4}$$

REMARK. Recall that if XY is integrable then $\mathbf{E}|XY|$ is by itself an upper bound on $|[\mathbf{E}XY]|$. The special case of $p = q = 2$ in Hölder's inequality

$$\mathbf{E}|XY| \leq \sqrt{\mathbf{E}X^2}\sqrt{\mathbf{E}Y^2} \, ,$$

is called the *Cauchy-Schwarz inequality.*

PROOF. Fixing $p > 1$ and $q = p/(p-1)$ let $\lambda = ||X||_p$ and $\xi = ||Y||_q$. If $\lambda = 0$ then $|X|^p \overset{a.s.}{=} 0$ (see Theorem 1.3.9). Likewise, if $\xi = 0$ then $|Y|^q \overset{a.s.}{=} 0$. In either case, the inequality (1.3.4) trivially holds. As this inequality also trivially holds when either $\lambda = \infty$ or $\xi = \infty$, we may and shall assume hereafter that both λ and ξ are finite and strictly positive. Recall that

$$\frac{x^p}{p} + \frac{y^q}{q} - xy \geq 0, \qquad \forall x, y \geq 0$$

(c.f. [**Dur03**, Page 462] where it is proved by considering the first two derivatives in x). Taking $x = |X|/\lambda$ and $y = |Y|/\xi$, we have by linearity and monotonicity of the expectation that

$$1 = \frac{1}{p} + \frac{1}{q} = \frac{\mathbf{E}|X|^p}{\lambda^p p} + \frac{\mathbf{E}|Y|^q}{\xi^q q} \geq \frac{\mathbf{E}|XY|}{\lambda\xi} \, ,$$

yielding the stated inequality (1.3.4). □

A direct consequence of Hölder's inequality is the triangle inequality for the norm $||X||_p$ in $L^p(\Omega, \mathcal{F}, \mathbf{P})$, that is,

PROPOSITION 1.3.18 (MINKOWSKI'S INEQUALITY). *If $X, Y \in L^p(\Omega, \mathcal{F}, \mathbf{P}), p \geq 1$, then $||X + Y||_p \leq ||X||_p + ||Y||_p$.*

Proof. With $|X+Y| \leq |X|+|Y|$, by monotonicity of the expectation we have the stated inequality in case $p=1$. Considering hereafter $p>1$, it follows from Hölder's inequality (Proposition 1.3.17) that

$$\begin{aligned}\mathbf{E}|X+Y|^p &= \mathbf{E}(|X+Y||X+Y|^{p-1})\\ &\leq \mathbf{E}(|X||X+Y|^{p-1}) + \mathbf{E}(|Y||X+Y|^{p-1})\\ &\leq (\mathbf{E}|X|^p)^{\frac{1}{p}}(\mathbf{E}|X+Y|^{(p-1)q})^{\frac{1}{q}} + \ (\mathbf{E}|Y|^p)^{\frac{1}{p}}(\mathbf{E}|X+Y|^{(p-1)q})^{\frac{1}{q}}\\ &= (||X||_p + ||Y||_p)\,(\mathbf{E}|X+Y|^p)^{\frac{1}{q}}\end{aligned}$$

(recall that $(p-1)q = p$). Since $X, Y \in L^p$ and

$$|x+y|^p \leq (|x|+|y|)^p \leq 2^{p-1}(|x|^p + |y|^p), \qquad \forall x, y \in \mathbb{R}, \ \ p>1,$$

if follows that $a_p = \mathbf{E}|X+Y|^p < \infty$. There is nothing to prove unless $a_p > 0$, in which case dividing by $(a_p)^{1/q}$ we get that

$$(\mathbf{E}|X+Y|^p)^{1-\frac{1}{q}} \leq ||X||_p + ||Y||_p\,,$$

giving the stated inequality (since $1-\frac{1}{q} = \frac{1}{p}$). □

Remark. Jensen's inequality applies only for probability measures, while both Hölder's inequality $\mu(|fg|) \leq \mu(|f|^p)^{1/p}\mu(|g|^q)^{1/q}$ and Minkowski's inequality apply for any measure μ, with exactly the same proof we provided for probability measures.

To practice your understanding of Markov's inequality, solve the following exercise.

Exercise 1.3.19. *Let X be a non-negative random variable with $\mathsf{Var}(X) \leq 1/2$. Show that then $\mathbf{P}(-1+\mathbf{E}X \leq X \leq 2\mathbf{E}X) \geq 1/2$.*

To practice your understanding of the proof of Jensen's inequality, try to prove its extension to convex functions on $\mathbb{R}^n$.

Exercise 1.3.20. *Suppose $g : \mathbb{R}^n \to \mathbb{R}$ is a convex function and $X_1, X_2, \ldots, X_n$ are integrable random variables, defined on the same probability space and such that $g(X_1, \ldots, X_n)$ is integrable. Show that then $\mathbf{E}g(X_1, \ldots, X_n) \geq g(\mathbf{E}X_1, \ldots, \mathbf{E}X_n)$.*
Hint: *Use convex analysis to show that $g(\cdot)$ is continuous and further that for any $\underline{c} \in \mathbb{R}^n$ there exists $\underline{b} \in \mathbb{R}^n$ such that $g(\underline{x}) \geq g(\underline{c}) + \langle \underline{b}, \underline{x} - \underline{c}\rangle$ for all $\underline{x} \in \mathbb{R}^n$ (with $\langle \cdot, \cdot \rangle$ denoting the inner product of two vectors in $\mathbb{R}^n$).*

Exercise 1.3.21. *Let $Y \geq 0$ with $v = \mathbf{E}(Y^2) < \infty$.*

(a) *Show that for any $0 \leq a < \mathbf{E}Y$,*

$$\mathbf{P}(Y > a) \geq \frac{(\mathbf{E}Y - a)^2}{\mathbf{E}(Y^2)}$$

Hint: *Apply the Cauchy-Schwarz inequality to $YI_{Y>a}$.*

(b) *Show that $(\mathbf{E}|Y^2 - v|)^2 \leq 4v(v - (\mathbf{E}Y)^2)$.*

(c) *Derive the second* Bonferroni inequality*,*

$$\mathbf{P}(\bigcup_{i=1}^n A_i) \geq \sum_{i=1}^n \mathbf{P}(A_i) - \sum_{1 \leq j < i \leq n} \mathbf{P}(A_i \cap A_j)\,.$$

How does it compare with the bound of part (a) for $Y = \sum_{i=1}^n I_{A_i}$?

1.3.3. Convergence, limits and expectation. Asymptotic behavior is a key issue in probability theory. We thus explore here various notions of convergence of random variables and the relations among them, focusing on the integrability conditions needed for exchanging the order of limit and expectation operations. Unless explicitly stated otherwise, throughout this section we assume that all R.V. are defined on the same probability space $(\Omega, \mathcal{F}, \mathbf{P})$.

In Definition 1.2.25 we have encountered the convergence almost surely of R.V. A weaker notion of convergence is *convergence in probability* as defined next.

DEFINITION 1.3.22. *We say that R.V. X_n converge to a given R.V. X_∞ in probability, denoted $X_n \xrightarrow{p} X_\infty$, if $\mathbf{P}(\{\omega : |X_n(\omega) - X_\infty(\omega)| > \varepsilon\}) \to 0$ as $n \to \infty$, for any fixed $\varepsilon > 0$. This is equivalent to $|X_n - X_\infty| \xrightarrow{p} 0$, and is a special case of the* convergence in μ-measure *of $f_n \in m\mathcal{F}$ to $f_\infty \in m\mathcal{F}$, that is $\mu(\{s : |f_n(s) - f_\infty(s)| > \varepsilon\}) \to 0$ as $n \to \infty$, for any fixed $\varepsilon > 0$.*

Our next exercise and example clarify the relationship between convergence almost surely and convergence in probability.

EXERCISE 1.3.23. *Verify that convergence almost surely* to a finite limit *implies convergence in probability, that is if $X_n \overset{a.s.}{\to} X_\infty \in \mathbb{R}$ then $X_n \xrightarrow{p} X_\infty$.*

REMARK 1.3.24. Generalizing Definition 1.3.22, for a separable metric space $(\mathbb{S}, \rho)$ we say that $(\mathbb{S}, \mathcal{B}_{\mathbb{S}})$-valued random variables X_n converge to X_∞ in probability if and only if for every $\varepsilon > 0$, $\mathbf{P}(\rho(X_n, X_\infty) > \varepsilon) \to 0$ as $n \to \infty$ (see [**Dud89**, Section 9.2] for more details). Equipping $\mathbb{S} = \overline{\mathbb{R}}$ with a suitable metric (for example, $\rho(x,y) = |\varphi(x) - \varphi(y)|$ with $\varphi(x) = x/(1+|x|) : \overline{\mathbb{R}} \mapsto [-1,1]$), this definition removes the restriction to X_∞ finite in Exercise 1.3.23.

In general, $X_n \xrightarrow{p} X_\infty$ does not imply that $X_n \overset{a.s.}{\to} X_\infty$.

EXAMPLE 1.3.25. *Consider the probability space $((0,1], \mathcal{B}_{(0,1]}, U)$ and $X_n(\omega) = \mathbf{1}_{[t_n, t_n+s_n]}(\omega)$ with $s_n \downarrow 0$ as $n \to \infty$ slowly enough and $t_n \in [0, 1-s_n]$ are such that any $\omega \in (0,1]$ is in infinitely many intervals $[t_n, t_n+s_n]$. The latter property applies if $t_n = (i-1)/k$ and $s_n = 1/k$ when $n = k(k-1)/2 + i$, $i = 1, 2, \ldots, k$ and $k = 1, 2, \ldots$ (plot the intervals $[t_n, t_n+s_n]$ to convince yourself). Then, $X_n \xrightarrow{p} 0$ (since $s_n = U(X_n \neq 0) \to 0$), whereas fixing each $\omega \in (0,1]$, we have that $X_n(\omega) = 1$ for infinitely many values of n, hence X_n does not converge a.s. to zero.*

Associated with each space $L^q(\Omega, \mathcal{F}, \mathbf{P})$ is the notion of L^q convergence which we now define.

DEFINITION 1.3.26. *We say that X_n* converges in L^q *to X_∞, denoted $X_n \xrightarrow{L^q} X_\infty$, if $X_n, X_\infty \in L^q$ and $||X_n - X_\infty||_q \to 0$ as $n \to \infty$ (i.e., $\mathbf{E}(|X_n - X_\infty|^q) \to 0$ as $n \to \infty$.*

REMARK. For $q = 2$ we have the explicit formula

$$||X_n - X||_2^2 = \mathbf{E}(X_n^2) - 2\mathbf{E}(X_n X) + \mathbf{E}(X^2).$$

Thus, it is often easiest to check convergence in L^2.

The following immediate corollary of Lemma 1.3.16 provides an ordering of L^q convergence in terms of the parameter q.

COROLLARY 1.3.27. *If $X_n \xrightarrow{L^q} X_\infty$ and $q \geq r$, then $X_n \xrightarrow{L^r} X_\infty$.*

Next note that the L^q convergence implies the convergence of the expectation of $|X_n|^q$.

Exercise 1.3.28. *Fixing $q \geq 1$, use Minkowski's inequality (Proposition 1.3.18), to show that if $X_n \overset{L^q}{\to} X_\infty$, then $\mathbf{E}|X_n|^q \to \mathbf{E}|X_\infty|^q$ and for $q = 1, 2, 3, \ldots$ also $\mathbf{E}X_n^q \to \mathbf{E}X_\infty^q$.*

Further, it follows from Markov's inequality that the convergence in L^q implies convergence in probability (for any value of q).

Proposition 1.3.29. *If $X_n \overset{L^q}{\to} X_\infty$, then $X_n \overset{p}{\to} X_\infty$.*

Proof. Fixing $q > 0$ recall that Markov's inequality results with

$$\mathbf{P}(|Y| > \varepsilon) \leq \varepsilon^{-q}\mathbf{E}[|Y|^q]\,,$$

for any R.V. Y and any $\varepsilon > 0$ (c.f part (b) of Example 1.3.14). The assumed convergence in L^q means that $\mathbf{E}[|X_n - X_\infty|^q] \to 0$ as $n \to \infty$, so taking $Y = Y_n = X_n - X_\infty$, we necessarily have also $\mathbf{P}(|X_n - X_\infty| > \varepsilon) \to 0$ as $n \to \infty$. Since $\varepsilon > 0$ is arbitrary, we see that $X_n \overset{p}{\to} X_\infty$ as claimed. □

The converse of Proposition 1.3.29 does not hold in general. As we next demonstrate, even the stronger almost surely convergence (see Exercise 1.3.23), and having a non-random constant limit are not enough to guarantee the L^q convergence, for any $q > 0$.

Example 1.3.30. *Fixing $q > 0$, consider the probability space $((0,1], \mathcal{B}_{(0,1]}, U)$ and the R.V. $Y_n(\omega) = n^{1/q} I_{[0,n^{-1}]}(\omega)$. Since $Y_n(\omega) = 0$ for all $n \geq n_0$ and some finite $n_0 = n_0(\omega)$, it follows that $Y_n(\omega) \overset{a.s.}{\to} 0$ as $n \to \infty$. However, $\mathbf{E}[|Y_n|^q] = nU([0, n^{-1}]) = 1$ for all n, so Y_n does not converge to zero in L^q (see Exercise 1.3.28).*

Considering Example 1.3.25, where $X_n \overset{L^q}{\to} 0$ while X_n does not converge a.s. to zero, and Example 1.3.30 which exhibits the converse phenomenon, we conclude that the convergence in L^q and the a.s. convergence are in general non comparable, and neither one is a consequence of convergence in probability.

Nevertheless, a sequence X_n can have at most one limit, regardless of which convergence mode is considered.

Exercise 1.3.31. *Check that if $X_n \overset{L^q}{\to} X$ and $X_n \overset{a.s.}{\to} Y$ then $X \overset{a.s.}{=} Y$.*

Though we have just seen that in general the order of the limit and expectation operations is non-interchangeable, we examine for the remainder of this subsection various conditions which do allow for such an interchange. Note in passing that upon proving any such result under certain point-wise convergence conditions, we may with no extra effort relax these to the corresponding almost sure convergence (and the same applies for integrals with respect to measures, see part (a) of Theorem 1.3.9, or that of Proposition 1.3.5).

Turning to do just that, we first outline the results which apply in the more general measure theory setting, starting with the proof of the *monotone convergence theorem.*

Proof of Theorem 1.3.4. By part (c) of Proposition 1.3.5, the proof of which did not use Theorem 1.3.4, we know that $\mu(h_n)$ is a non-decreasing sequence that is bounded above by $\mu(h)$. It therefore suffices to show that

$$\begin{aligned}\lim_{n\to\infty}\mu(h_n) &= \sup_n\{\mu_0(\psi) : \psi\in \mathrm{SF}_+, \psi\le h_n\}\\ &\ge \sup\{\mu_0(\varphi) : \varphi\in\mathrm{SF}_+, \varphi\le h\} = \mu(h)\end{aligned} \tag{1.3.5}$$

(see Step 3 of Definition 1.3.1). That is, it suffices to find for each non-negative simple function $\varphi\le h$ a sequence of non-negative simple functions $\psi_n\le h_n$ such that $\mu_0(\psi_n)\to\mu_0(\varphi)$ as $n\to\infty$. To this end, fixing φ, we may and shall choose without loss of generality a representation $\varphi = \sum\limits_{l=1}^{m} c_l I_{A_l}$ such that $A_l\in\mathcal{F}$ are disjoint and further $c_l\mu(A_l)>0$ for $l=1,\dots,m$ (see proof of Lemma 1.3.3). Using hereafter the notation $f_*(A)=\inf\{f(s) : s\in A\}$ for $f\in m\mathcal{F}_+$ and $A\in\mathcal{F}$, the condition $\varphi(s)\le h(s)$ for all $s\in\mathbb{S}$ is equivalent to $c_l\le h_*(A_l)$ for all l, so

$$\mu_0(\varphi)\le\sum_{l=1}^{m} h_*(A_l)\mu(A_l) = V\,.$$

Suppose first that $V<\infty$, that is $0<h_*(A_l)\mu(A_l)<\infty$ for all l. In this case, fixing $\lambda<1$, consider for each n the disjoint sets $A_{l,\lambda,n}=\{s\in A_l : h_n(s)\ge\lambda h_*(A_l)\}\in\mathcal{F}$ and the corresponding

$$\psi_{\lambda,n}(s)=\sum_{l=1}^{m}\lambda h_*(A_l) I_{A_{l,\lambda,n}}(s)\in\mathrm{SF}_+\,,$$

where $\psi_{\lambda,n}(s)\le h_n(s)$ for all $s\in\mathbb{S}$. If $s\in A_l$ then $h(s)>\lambda h_*(A_l)$. Thus, $h_n\uparrow h$ implies that $A_{l,\lambda,n}\uparrow A_l$ as $n\to\infty$, for each l. Consequently, by definition of $\mu(h_n)$ and the continuity from below of μ,

$$\lim_{n\to\infty}\mu(h_n)\ge\lim_{n\to\infty}\mu_0(\psi_{\lambda,n})=\lambda V\,.$$

Taking $\lambda\uparrow 1$ we deduce that $\lim_n\mu(h_n)\ge V\ge\mu_0(\varphi)$. Next suppose that $V=\infty$, so without loss of generality we may and shall assume that $h_*(A_1)\mu(A_1)=\infty$. Fixing $x\in(0,h_*(A_1))$ let $A_{1,x,n}=\{s\in A_1 : h_n(s)\ge x\}\in\mathcal{F}$ noting that $A_{1,x,n}\uparrow A_1$ as $n\to\infty$ and $\psi_{x,n}(s)=xI_{A_{1,x,n}}(s)\le h_n(s)$ for all n and $s\in\mathbb{S}$, is a non-negative simple function. Thus, again by continuity from below of μ we have that

$$\lim_{n\to\infty}\mu(h_n)\ge\lim_{n\to\infty}\mu_0(\psi_{x,n})=x\mu(A_1)\,.$$

Taking $x\uparrow h_*(A_1)$ we deduce that $\lim_n\mu(h_n)\ge h_*(A_1)\mu(A_1)=\infty$, completing the proof of (1.3.5) and that of the theorem. □

Considering probability spaces, Theorem 1.3.4 tells us that we can exchange the order of the limit and the expectation in case of monotone upward a.s. convergence of non-negative R.V. (with the limit possibly infinite). That is,

Theorem 1.3.32 (Monotone convergence theorem). *If $X_n\ge 0$ and $X_n(\omega)\uparrow X_\infty(\omega)$ for almost every ω, then $\mathbf{E}X_n\uparrow\mathbf{E}X_\infty$.*

In Example 1.3.30 we have a point-wise convergent sequence of R.V. whose expectations exceed that of their limit. In a sense this is always the case, as stated next in Fatou's lemma (which is a direct consequence of the monotone convergence theorem).

Lemma 1.3.33 (Fatou's lemma). *For any measure space $(\mathbb{S}, \mathcal{F}, \mu)$ and any $f_n \in m\mathcal{F}$, if $f_n(s) \geq g(s)$ for some μ-integrable function g, all n and μ-almost-every $s \in \mathbb{S}$, then*

$$\liminf_{n\to\infty} \mu(f_n) \geq \mu(\liminf_{n\to\infty} f_n)\,. \tag{1.3.6}$$

Alternatively, if $f_n(s) \leq g(s)$ for all n and a.e. s, then

$$\limsup_{n\to\infty} \mu(f_n) \leq \mu(\limsup_{n\to\infty} f_n)\,. \tag{1.3.7}$$

Proof. Assume first that $f_n \in m\mathcal{F}_+$ and let $h_n(s) = \inf_{k\geq n} f_k(s)$, noting that $h_n \in m\mathcal{F}_+$ is a non-decreasing sequence, whose point-wise limit is $h(s) := \liminf_{n\to\infty} f_n(s)$. By the monotone convergence theorem, $\mu(h_n) \uparrow \mu(h)$. Since $f_n(s) \geq h_n(s)$ for all $s \in \mathbb{S}$, the monotonicity of the integral (see Proposition 1.3.5) implies that $\mu(f_n) \geq \mu(h_n)$ for all n. Considering the lim inf as $n \to \infty$ we arrive at (1.3.6).

Turning to extend this inequality to the more general setting of the lemma, note that our conditions imply that $f_n \overset{a.e.}{=} g + (f_n - g)_+$ for each n. Considering the countable union of the μ-negligible sets in which one of these identities is violated, we thus have that

$$h := \liminf_{n\to\infty} f_n \overset{a.e.}{=} g + \liminf_{n\to\infty} (f_n - g)_+\,.$$

Further, $\mu(f_n) = \mu(g) + \mu((f_n - g)_+)$ by the linearity of the integral in $m\mathcal{F}_+ \cup L^1$. Taking $n \to \infty$ and applying (1.3.6) for $(f_n - g)_+ \in m\mathcal{F}_+$ we deduce that

$$\liminf_{n\to\infty} \mu(f_n) \geq \mu(g) + \mu(\liminf_{n\to\infty}(f_n - g)_+) = \mu(g) + \mu(h - g) = \mu(h)$$

(where for the right most identity we used the linearity of the integral, as well as the fact that $-g$ is μ-integrable).

Finally, we get (1.3.7) for f_n by considering (1.3.6) for $-f_n$. □

Remark. In terms of the expectation, Fatou's lemma is the statement that if R.V. $X_n \geq X$, almost surely, for some $X \in L^1$ and all n, then

$$\liminf_{n\to\infty} \mathbf{E}(X_n) \geq \mathbf{E}(\liminf_{n\to\infty} X_n)\,, \tag{1.3.8}$$

whereas if $X_n \leq X$, almost surely, for some $X \in L^1$ and all n, then

$$\limsup_{n\to\infty} \mathbf{E}(X_n) \leq \mathbf{E}(\limsup_{n\to\infty} X_n)\,. \tag{1.3.9}$$

Some text books call (1.3.9) and (1.3.7) the *Reverse Fatou Lemma* (e.g. [**Wil91**, Section 5.4]).

Using Fatou's lemma, we can easily prove Lebesgue's dominated convergence theorem (in short DOM).

Theorem 1.3.34 (Dominated convergence theorem). *For any measure space $(\mathbb{S}, \mathcal{F}, \mu)$ and any $f_n \in m\mathcal{F}$, if for some μ-integrable function g and μ-almost-every $s \in \mathbb{S}$ both $f_n(s) \to f_\infty(s)$ as $n \to \infty$, and $|f_n(s)| \leq g(s)$ for all n, then f_∞ is μ-integrable and further $\mu(|f_n - f_\infty|) \to 0$ as $n \to \infty$.*

Proof. Up to a μ-negligible subset of $\mathbb{S}$, our assumption that $|f_n| \leq g$ and $f_n \to f_\infty$, implies that $|f_\infty| \leq g$, hence f_∞ is μ-integrable. Applying Fatou's lemma (1.3.7) for $|f_n - f_\infty| \leq 2g$ such that $\limsup_n |f_n - f_\infty| = 0$, we conclude that

$$0 \leq \limsup_{n\to\infty} \mu(|f_n - f_\infty|) \leq \mu(\limsup_{n\to\infty} |f_n - f_\infty|) = \mu(0) = 0\,,$$

as claimed. □

By Minkowski's inequality, $\mu(|f_n - f_\infty|) \to 0$ implies that $\mu(|f_n|) \to \mu(|f_\infty|)$. The dominated convergence theorem provides us with a simple sufficient condition for the converse implication in case also $f_n \to f_\infty$ a.e.

Lemma 1.3.35 (Scheffé's lemma). *If $f_n \in m\mathcal{F}$ converges a.e. to $f_\infty \in m\mathcal{F}$ and $\mu(|f_n|) \to \mu(|f_\infty|) < \infty$ then $\mu(|f_n - f_\infty|) \to 0$ as $n \to \infty$.*

Remark. In terms of expectation, Scheffé's lemma states that if $X_n \overset{a.s.}{\to} X_\infty$ and $\mathbf{E}|X_n| \to \mathbf{E}|X_\infty| < \infty$, then $X_n \overset{L^1}{\to} X_\infty$ as well.

Proof. As already noted, we may assume without loss of generality that $f_n(s) \to f_\infty(s)$ for *all* $s \in \mathbb{S}$, that is $g_n(s) = f_n(s) - f_\infty(s) \to 0$ as $n \to \infty$, for all $s \in \mathbb{S}$. Further, since $\mu(|f_n|) \to \mu(|f_\infty|) < \infty$, we may and shall assume also that f_n are $\mathbb{R}$-valued and μ-integrable for all $n \le \infty$, hence $g_n \in L^1(\mathbb{S}, \mathcal{F}, \mu)$ as well.
Suppose first that $f_n \in m\mathcal{F}_+$ for all $n \le \infty$. In this case, $0 \le (g_n)_- \le f_\infty$ for all n and s. As $(g_n)_-(s) \to 0$ for every $s \in \mathbb{S}$, applying the dominated convergence theorem we deduce that $\mu((g_n)_-) \to 0$. From the assumptions of the lemma (and the linearity of the integral on L^1), we get that $\mu(g_n) = \mu(f_n) - \mu(f_\infty) \to 0$ as $n \to \infty$. Since $|x| = x + 2x_-$ for any $x \in \mathbb{R}$, it thus follows by linearity of the integral on L^1 that $\mu(|g_n|) = \mu(g_n) + 2\mu((g_n)_-) \to 0$ for $n \to \infty$, as claimed.
In the general case of $f_n \in m\mathcal{F}$, we know that both $0 \le (f_n)_+(s) \to (f_\infty)_+(s)$ and $0 \le (f_n)_-(s) \to (f_\infty)_-(s)$ for every s, so by (1.3.6) of Fatou's lemma, we have that

$$\begin{aligned}\mu(|f_\infty|) = \mu((f_\infty)_+) + \mu((f_\infty)_-) &\le \liminf_{n\to\infty} \mu((f_n)_-) + \liminf_{n\to\infty} \mu((f_n)_+) \\ &\le \liminf_{n\to\infty} [\mu((f_n)_-) + \mu((f_n)_+)] = \lim_{n\to\infty} \mu(|f_n|) = \mu(|f_\infty|)\,.\end{aligned}$$

Hence, necessarily both $\mu((f_n)_+) \to \mu((f_\infty)_+)$ and $\mu((f_n)_-) \to \mu((f_\infty)_-)$. Since $|x - y| \le |x_+ - y_+| + |x_- - y_-|$ for all $x, y \in \mathbb{R}$ and we already proved the lemma for the non-negative $(f_n)_-$ and $(f_n)_+$, we see that

$$\lim_{n\to\infty} \mu(|f_n - f_\infty|) \le \lim_{n\to\infty} \mu(|(f_n)_+ - (f_\infty)_+|) + \lim_{n\to\infty} \mu(|(f_n)_- - (f_\infty)_-|) = 0\,,$$

concluding the proof of the lemma. □

We conclude this sub-section with quite a few exercises, starting with an alternative characterization of convergence almost surely.

Exercise 1.3.36. *Show that $X_n \overset{a.s.}{\to} 0$ if and only if for each $\varepsilon > 0$ there is n so that for each* random *integer M with $M(\omega) \ge n$ for all $\omega \in \Omega$ we have that $\mathbf{P}(\{\omega : |X_{M(\omega)}(\omega)| > \varepsilon\}) < \varepsilon$.*

Exercise 1.3.37. *Let Y_n be (real-valued) random variables on $(\Omega, \mathcal{F}, \mathbf{P})$, and N_k positive integer valued random variables on the same probability space.*

(a) *Show that $Y_{N_k}(\omega) = Y_{N_k(\omega)}(\omega)$ are random variables on $(\Omega, \mathcal{F})$.*
(b) *Show that if $Y_n \overset{a.s}{\to} Y_\infty$ and $N_k \overset{a.s.}{\to} \infty$ then $Y_{N_k} \overset{a.s.}{\to} Y_\infty$.*
(c) *Provide an example of $Y_n \overset{p}{\to} 0$ and $N_k \overset{a.s.}{\to} \infty$ such that almost surely $Y_{N_k} = 1$ for all k.*
(d) *Show that if $Y_n \overset{a.s.}{\to} Y_\infty$ and $\mathbf{P}(N_k > r) \to 1$ as $k \to \infty$, for every fixed $r < \infty$, then $Y_{N_k} \overset{p}{\to} Y_\infty$.*

In the following four exercises you find some of the many applications of the monotone convergence theorem.

EXERCISE 1.3.38. *You are now to relax the non-negativity assumption in the monotone convergence theorem.*

(a) *Show that if* $\mathbf{E}[(X_1)_-] < \infty$ *and* $X_n(\omega) \uparrow X(\omega)$ *for almost every* ω*, then* $\mathbf{E}X_n \uparrow \mathbf{E}X$*.*

(b) *Show that if in addition* $\sup_n \mathbf{E}[(X_n)_+] < \infty$*, then* $X \in L^1(\Omega, \mathcal{F}, \mathbf{P})$*.*

EXERCISE 1.3.39. *In this exercise you are to show that for any R.V.* $X \geq 0$*,*

$$\mathbf{E}X = \lim_{\delta \downarrow 0} \mathbf{E}_\delta X \quad \textit{for} \quad \mathbf{E}_\delta X = \sum_{j=0}^{\infty} j\delta \mathbf{P}(\{\omega : j\delta < X(\omega) \leq (j+1)\delta\}) \,. \tag{1.3.10}$$

First use monotone convergence to show that $\mathbf{E}_{\delta_k} X$ *converges to* $\mathbf{E}X$ *along the sequence* $\delta_k = 2^{-k}$*. Then, check that* $\mathbf{E}_\delta X \leq \mathbf{E}_\eta X + \eta$ *for any* $\delta, \eta > 0$ *and deduce from it the identity (1.3.10).*

Applying (1.3.10) verify that if X *takes at most countably many values* $\{x_1, x_2, \ldots\}$*, then* $\mathbf{E}X = \sum_i x_i \mathbf{P}(\{\omega : X(\omega) = x_i\})$ *(this applies to every R.V.* $X \geq 0$ *on a countable* Ω*). More generally, verify that such formula applies whenever the series is absolutely convergent (which amounts to* $X \in L^1$*).*

EXERCISE 1.3.40. *Use monotone convergence to show that for any sequence of non-negative R.V.* Y_n*,*

$$\mathbf{E}(\sum_{n=1}^{\infty} Y_n) = \sum_{n=1}^{\infty} \mathbf{E}Y_n \,.$$

EXERCISE 1.3.41. *Suppose* $X_n, X \in L^1(\Omega, \mathcal{F}, \mathbf{P})$ *are such that*

(a) $X_n \geq 0$ *almost surely,* $\mathbf{E}[X_n] = 1$, $\mathbf{E}[X_n \log X_n] \leq 1$, *and*

(b) $\mathbf{E}[X_n Y] \to \mathbf{E}[XY]$ *as* $n \to \infty$*, for each bounded random variable* Y *on* $(\Omega, \mathcal{F})$*.*

Show that then $X \geq 0$ *almost surely,* $\mathbf{E}[X] = 1$ *and* $\mathbf{E}[X \log X] \leq 1$*.*
Hint: *Jensen's inequality is handy for showing that* $\mathbf{E}[X \log X] \leq 1$*.*

Next come few direct applications of the dominated convergence theorem.

EXERCISE 1.3.42.

(a) *Show that for any random variable* X*, the function* $t \mapsto \mathbf{E}[e^{-|t-X|}]$ *is continuous on* $\mathbb{R}$ *(this function is sometimes called the* bilateral exponential transform*).*

(b) *Suppose* $X \geq 0$ *is such that* $\mathbf{E}X^q < \infty$ *for some* $q > 0$*. Show that then* $q^{-1}(\mathbf{E}X^q - 1) \to \mathbf{E}\log X$ *as* $q \downarrow 0$ *and deduce that also* $q^{-1} \log \mathbf{E}X^q \to \mathbf{E}\log X$ *as* $q \downarrow 0$*.*

Hint: *Fixing* $x \geq 0$ *deduce from convexity of* $q \mapsto x^q$ *that* $q^{-1}(x^q - 1) \downarrow \log x$ *as* $q \downarrow 0$*.*

EXERCISE 1.3.43. *Suppose* X *is an integrable random variable.*

(a) *Show that* $\mathbf{E}(|X| I_{\{X>n\}}) \to 0$ *as* $n \to \infty$*.*

(b) *Deduce that for any* $\varepsilon > 0$ *there exists* $\delta > 0$ *such that*

$$\sup\{\mathbf{E}[|X| I_A] : \mathbf{P}(A) \leq \delta\} \leq \varepsilon \,.$$

(c) *Provide an example of $X \geq 0$ with $\mathbf{E}X = \infty$ for which the preceding fails, that is, $\mathbf{P}(A_k) \to 0$ as $k \to \infty$ while $\mathbf{E}[XI_{A_k}]$ is bounded away from zero.*

The following generalization of the dominated convergence theorem is also left as an exercise.

Exercise 1.3.44. *Suppose $g_n(\cdot) \leq f_n(\cdot) \leq h_n(\cdot)$ are μ-integrable functions in the same measure space $(\mathbb{S}, \mathcal{F}, \mu)$ such that for μ-almost-every $s \in \mathbb{S}$ both $g_n(s) \to g_\infty(s)$, $f_n(s) \to f_\infty(s)$ and $h_n(s) \to h_\infty(s)$ as $n \to \infty$. Show that if further g_∞ and h_∞ are μ-integrable functions such that $\mu(g_n) \to \mu(g_\infty)$ and $\mu(h_n) \to \mu(h_\infty)$, then f_∞ is μ-integrable and $\mu(f_n) \to \mu(f_\infty)$.*

Finally, here is a demonstration of one of the many issues that are particularly easy to resolve with respect to the $L^2(\Omega, \mathcal{F}, \mathbf{P})$ norm.

Exercise 1.3.45. *Let $X = (X(t))_{t \in \mathbb{R}}$ be a mapping from $\mathbb{R}$ into $L^2(\Omega, \mathcal{F}, \mathbf{P})$. Show that $t \mapsto X(t)$ is a continuous mapping (with respect to the norm $\|\cdot\|_2$ on $L^2(\Omega, \mathcal{F}, \mathbf{P})$), if and only if both*

$$\mu(t) = \mathbf{E}[X(t)] \quad \text{and} \quad r(s,t) = \mathbf{E}[X(s)X(t)] - \mu(s)\mu(t)$$

are continuous real-valued functions ($r(s,t)$ is continuous as a map from $\mathbb{R}^2$ to $\mathbb{R}$).

1.3.4. L^1-convergence and uniform integrability. For probability theory, the dominated convergence theorem states that if random variables $X_n \overset{a.s.}{\to} X_\infty$ are such that $|X_n| \leq Y$ for all n and some random variable Y such that $\mathbf{E}Y < \infty$, then $X_\infty \in L^1$ and $X_n \overset{L^1}{\to} X_\infty$. Since constants have finite expectation (see part (d) of Theorem 1.3.9), we have as its corollary the *bounded convergence theorem*, that is,

Corollary 1.3.46 (Bounded Convergence). *Suppose that a.s. $|X_n(\omega)| \leq K$ for some finite non-random constant K and all n. If $X_n \overset{a.s.}{\to} X_\infty$, then $X_\infty \in L^1$ and $X_n \overset{L^1}{\to} X_\infty$.*

We next state a *uniform integrability* condition that together with convergence in probability implies the convergence in L^1.

Definition 1.3.47. *A possibly uncountable collection of R.V.-s $\{X_\alpha, \alpha \in \mathcal{I}\}$ is called* uniformly integrable *(U.I.) if*

$$\lim_{M \to \infty} \sup_\alpha \mathbf{E}[|X_\alpha| I_{|X_\alpha|>M}] = 0 \,. \tag{1.3.11}$$

Our next lemma shows that U.I. is a relaxation of the condition of dominated convergence, and that U.I. still implies the boundedness in L^1 of $\{X_\alpha, \alpha \in \mathcal{I}\}$.

Lemma 1.3.48. *If $|X_\alpha| \leq Y$ for all α and some R.V. Y such that $\mathbf{E}Y < \infty$, then the collection $\{X_\alpha\}$ is U.I. In particular, any finite collection of integrable R.V. is U.I.*

Further, if $\{X_\alpha\}$ is U.I. then $\sup_\alpha \mathbf{E}|X_\alpha| < \infty$.

Proof. By monotone convergence, $\mathbf{E}[YI_{Y \leq M}] \uparrow \mathbf{E}Y$ as $M \uparrow \infty$, for any R.V. $Y \geq 0$. Hence, if in addition $\mathbf{E}Y < \infty$, then by linearity of the expectation, $\mathbf{E}[YI_{Y>M}] \downarrow 0$ as $M \uparrow \infty$. Now, if $|X_\alpha| \leq Y$ then $|X_\alpha| I_{|X_\alpha|>M} \leq YI_{Y>M}$, hence $\mathbf{E}[|X_\alpha| I_{|X_\alpha|>M}] \leq \mathbf{E}[YI_{Y>M}]$, which does not depend on α, and for $Y \in L^1$ converges to zero when $M \to \infty$. We thus proved that if $|X_\alpha| \leq Y$ for all α and some Y such that $\mathbf{E}Y < \infty$, then $\{X_\alpha\}$ is a U.I. collection of R.V.-s

For a finite collection of R.V.-s $X_i \in L^1$, $i = 1, \ldots, k$, take $Y = |X_1| + |X_2| + \cdots + |X_k| \in L^1$ such that $|X_i| \le Y$ for $i = 1, \ldots, k$, to see that any finite collection of integrable R.V.-s is U.I.

Finally, since

$$\mathbf{E}|X_\alpha| = \mathbf{E}[|X_\alpha| I_{|X_\alpha| \le M}] + \mathbf{E}[|X_\alpha| I_{|X_\alpha|>M}] \le M + \sup_\alpha \mathbf{E}[|X_\alpha| I_{|X_\alpha|>M}],$$

we see that if $\{X_\alpha, \alpha \in \mathcal{I}\}$ is U.I. then $\sup_\alpha \mathbf{E}|X_\alpha| < \infty$. □

We next state and prove Vitali's convergence theorem for probability measures, deferring the general case to Exercise 1.3.53.

Theorem 1.3.49 (Vitali's convergence theorem). *Suppose $X_n \xrightarrow{p} X_\infty$. Then, the collection $\{X_n\}$ is U.I. if and only if $X_n \xrightarrow{L^1} X_\infty$ which in turn is equivalent to X_n being integrable for all $n \le \infty$ and $\mathbf{E}|X_n| \to \mathbf{E}|X_\infty|$.*

Remark. In view of Lemma 1.3.48, Vitali's theorem relaxes the assumed a.s. convergence $X_n \to X_\infty$ of the dominated (or bounded) convergence theorem, and of Scheffé's lemma, to that of convergence in probability.

Proof. Suppose first that $|X_n| \le M$ for some non-random finite constant M and all n. For each $\varepsilon > 0$ let $B_{n,\varepsilon} = \{\omega : |X_n(\omega) - X_\infty(\omega)| > \varepsilon\}$. The assumed convergence in probability means that $\mathbf{P}(B_{n,\varepsilon}) \to 0$ as $n \to \infty$ (see Definition 1.3.22). Since $\mathbf{P}(|X_\infty| \ge M + \varepsilon) \le \mathbf{P}(B_{n,\varepsilon})$, taking $n \to \infty$ and considering $\varepsilon = \varepsilon_k \downarrow 0$, we get by continuity from below of $\mathbf{P}$ that almost surely $|X_\infty| \le M$. So, $|X_n - X_\infty| \le 2M$ and by linearity and monotonicity of the expectation, for any n and $\varepsilon > 0$,

$$\begin{aligned}\mathbf{E}|X_n - X_\infty| &= \mathbf{E}[|X_n - X_\infty| I_{B^c_{n,\varepsilon}}] + \mathbf{E}[|X_n - X_\infty| I_{B_{n,\varepsilon}}] \\ &\le \mathbf{E}[\varepsilon I_{B^c_{n,\varepsilon}}] + \mathbf{E}[2M I_{B_{n,\varepsilon}}] \le \varepsilon + 2M\mathbf{P}(B_{n,\varepsilon}).\end{aligned}$$

Since $\mathbf{P}(B_{n,\varepsilon}) \to 0$ as $n \to \infty$, it follows that $\limsup_{n\to\infty} \mathbf{E}|X_n - X_\infty| \le \varepsilon$. Taking $\varepsilon \downarrow 0$ we deduce that $\mathbf{E}|X_n - X_\infty| \to 0$ in this case.

Moving to deal now with the general case of a collection $\{X_n\}$ that is U.I., let $\varphi_M(x) = \max(\min(x, M), -M)$. As $|\varphi_M(x) - \varphi_M(y)| \le |x-y|$ for any $x, y \in \mathbb{R}$, our assumption $X_n \xrightarrow{p} X_\infty$ implies that $\varphi_M(X_n) \xrightarrow{p} \varphi_M(X_\infty)$ for any fixed $M < \infty$. With $|\varphi_M(\cdot)| \le M$, we then have by the preceding proof of bounded convergence that $\varphi_M(X_n) \xrightarrow{L^1} \varphi_M(X_\infty)$. Further, by Minkowski's inequality, also $\mathbf{E}|\varphi_M(X_n)| \to \mathbf{E}|\varphi_M(X_\infty)|$. By Lemma 1.3.48, our assumption that $\{X_n\}$ are U.I. implies their L^1 boundedness, and since $|\varphi_M(x)| \le |x|$ for all x, we deduce that for any M,

$$\infty > c := \sup_n \mathbf{E}|X_n| \ge \lim_{n\to\infty} \mathbf{E}|\varphi_M(X_n)| = \mathbf{E}|\varphi_M(X_\infty)|. \tag{1.3.12}$$

With $|\varphi_M(X_\infty)| \uparrow |X_\infty|$ as $M \uparrow \infty$, it follows from monotone convergence that $\mathbf{E}|\varphi_M(X_\infty)| \uparrow \mathbf{E}|X_\infty|$, hence $\mathbf{E}|X_\infty| \le c < \infty$ in view of (1.3.12). Fixing $\varepsilon > 0$, choose $M = M(\varepsilon) < \infty$ large enough for $\mathbf{E}[|X_\infty| I_{|X_\infty|>M}] < \varepsilon$, and further increasing M if needed, by the U.I. condition also $\mathbf{E}[|X_n| I_{|X_n|>M}] < \varepsilon$ for all n. Considering the expectation of the inequality $|x - \varphi_M(x)| \le |x| I_{|x|>M}$ (which holds for all $x \in \mathbb{R}$), with $x = X_n$ and $x = X_\infty$, we obtain that

$$\begin{aligned}\mathbf{E}|X_n - X_\infty| &\le \mathbf{E}|X_n - \varphi_M(X_n)| + \mathbf{E}|\varphi_M(X_n) - \varphi_M(X_\infty)| + \mathbf{E}|X_\infty - \varphi_M(X_\infty)| \\ &\le 2\varepsilon + \mathbf{E}|\varphi_M(X_n) - \varphi_M(X_\infty)|.\end{aligned}$$

Recall that $\varphi_M(X_n) \xrightarrow{L^1} \varphi_M(X_\infty)$, hence $\limsup_n \mathbf{E}|X_n - X_\infty| \le 2\varepsilon$. Taking $\varepsilon \to 0$ completes the proof of L^1 convergence of X_n to X_∞.

Suppose now that $X_n \xrightarrow{L^1} X_\infty$. Then, by Jensen's inequality (for the convex function $g(x) = |x|$),

$$|\mathbf{E}|X_n| - \mathbf{E}|X_\infty|| \le \mathbf{E}[|\ |X_n| - |X_\infty|\ |] \le \mathbf{E}|X_n - X_\infty| \to 0.$$

That is, $\mathbf{E}|X_n| \to \mathbf{E}|X_\infty|$ and X_n, $n \le \infty$ are integrable.

It thus remains only to show that if $X_n \xrightarrow{p} X_\infty$, all of which are integrable and $\mathbf{E}|X_n| \to \mathbf{E}|X_\infty|$ then the collection $\{X_n\}$ is U.I. To the end, for any $M > 1$, let

$$\psi_M(x) = |x| I_{|x| \le M-1} + (M-1)(M - |x|) I_{(M-1,M]}(|x|)\,,$$

a piecewise-linear, continuous, bounded function, such that $\psi_M(x) = |x|$ for $|x| \le M-1$ and $\psi_M(x) = 0$ for $|x| \ge M$. Fixing $\epsilon > 0$, with X_∞ integrable, by dominated convergence $\mathbf{E}|X_\infty| - \mathbf{E}\psi_m(X_\infty) \le \epsilon$ for some finite $m = m(\epsilon)$. Further, as $|\psi_m(x) - \psi_m(y)| \le (m-1)|x-y|$ for any $x, y \in \mathbb{R}$, our assumption $X_n \xrightarrow{p} X_\infty$ implies that $\psi_m(X_n) \xrightarrow{p} \psi_m(X_\infty)$. Hence, by the preceding proof of bounded convergence, followed by Minkowski's inequality, we deduce that $\mathbf{E}\psi_m(X_n) \to \mathbf{E}\psi_m(X_\infty)$ as $n \to \infty$. Since $|x| I_{|x|>m} \le |x| - \psi_m(x)$ for all $x \in \mathbb{R}$, our assumption $\mathbf{E}|X_n| \to \mathbf{E}|X_\infty|$ thus implies that for some $n_0 = n_0(\epsilon)$ finite and all $n \ge n_0$ and $M \ge m(\epsilon)$,

$$\begin{aligned}\mathbf{E}[|X_n| I_{|X_n|>M}] \le \mathbf{E}[|X_n| I_{|X_n|>m}] &\le \mathbf{E}|X_n| - \mathbf{E}\psi_m(X_n) \\ &\le \mathbf{E}|X_\infty| - \mathbf{E}\psi_m(X_\infty) + \epsilon \le 2\epsilon\,.\end{aligned}$$

As each X_n is integrable, $\mathbf{E}[|X_n| I_{|X_n|>M}] \le 2\epsilon$ for some $M \ge m$ finite and all n (including also $n < n_0(\epsilon)$). The fact that such finite $M = M(\epsilon)$ exists for any $\epsilon > 0$ amounts to the collection $\{X_n\}$ being U.I. □

The following exercise builds upon the bounded convergence theorem.

Exercise 1.3.50. *Show that for any $X \ge 0$ (do not assume $\mathbf{E}(1/X) < \infty$), both*

(a) $\lim\limits_{y \to \infty} y \mathbf{E}[X^{-1} I_{X>y}] = 0$ *and*

(b) $\lim\limits_{y \downarrow 0} y \mathbf{E}[X^{-1} I_{X>y}] = 0$.

Next is an example of the advantage of Vitali's convergence theorem over the dominated convergence theorem.

Exercise 1.3.51. *On $((0,1], \mathcal{B}_{(0,1]}, U)$, let $X_n(\omega) = (n/\log n) I_{(0,n^{-1})}(\omega)$ for $n \ge 2$. Show that the collection $\{X_n\}$ is U.I. such that $X_n \xrightarrow{a.s.} 0$ and $\mathbf{E}X_n \to 0$, but there is no random variable Y with finite expectation such that $Y \ge X_n$ for all $n \ge 2$ and almost all $\omega \in (0,1]$.*

By a simple application of Vitali's convergence theorem you can derive a classical result of analysis, dealing with the convergence of Cesáro averages.

Exercise 1.3.52. *Let U_n denote a random variable whose law is the* uniform probability measure *on $(0,n]$, namely, Lebesgue measure restricted to the interval $(0,n]$ and normalized by n^{-1} to a probability measure. Show that $g(U_n) \xrightarrow{p} 0$ as $n \to \infty$, for any Borel function $g(\cdot)$ such that $|g(y)| \to 0$ as $y \to \infty$. Further, assuming that also* $\sup_y |g(y)| < \infty$, *deduce that $\mathbf{E}|g(U_n)| = n^{-1} \int_0^n |g(y)| dy \to 0$ as $n \to \infty$.*

Here is Vitali's convergence theorem for a general measure space.

EXERCISE 1.3.53. *Given a measure space $(\mathbb{S}, \mathcal{F}, \mu)$, suppose $f_n, f_\infty \in m\mathcal{F}$ with $\mu(|f_n|)$ finite and $\mu(|f_n - f_\infty| > \varepsilon) \to 0$ as $n \to \infty$, for each fixed $\varepsilon > 0$. Show that $\mu(|f_n - f_\infty|) \to 0$ as $n \to \infty$ if and only if both $\sup_n \mu(|f_n| I_{|f_n|>k}) \to 0$ and $\sup_n \mu(|f_n| I_{A_k}) \to 0$ for $k \to \infty$ and some $\{A_k\} \subseteq \mathcal{F}$ such that $\mu(A_k^c) < \infty$.*

We conclude this subsection with a useful sufficient criterion for uniform integrability and few of its consequences.

EXERCISE 1.3.54. *Let $f \geq 0$ be a Borel function such that $f(r)/r \to \infty$ as $r \to \infty$. Suppose $\mathbf{E} f(|X_\alpha|) \leq C$ for some finite non-random constant C and all $\alpha \in \mathcal{I}$. Show that then $\{X_\alpha : \alpha \in \mathcal{I}\}$ is a uniformly integrable collection of R.V.*

EXERCISE 1.3.55.

(a) *Construct random variables X_n such that $\sup_n \mathbf{E}(|X_n|) < \infty$, but the collection $\{X_n\}$ is not uniformly integrable.*

(b) *Show that if $\{X_n\}$ is a U.I. collection and $\{Y_n\}$ is a U.I. collection, then $\{X_n + Y_n\}$ is also U.I.*

(c) *Show that if $X_n \xrightarrow{p} X_\infty$ and the collection $\{X_n\}$ is uniformly integrable, then $\mathbf{E}(X_n I_A) \to \mathbf{E}(X_\infty I_A)$ as $n \to \infty$, for any measurable set A.*

1.3.5. Expectation, density and Riemann integral.

Applying the *standard machine* we now show that fixing a measure space $(\mathbb{S}, \mathcal{F}, \mu)$, each non-negative measurable function f induces a measure $f\mu$ on $(\mathbb{S}, \mathcal{F})$, with f being the natural generalization of the concept of probability density function.

PROPOSITION 1.3.56. *Fix a measure space $(\mathbb{S}, \mathcal{F}, \mu)$. Every $f \in m\mathcal{F}_+$ induces a measure $f\mu$ on $(\mathbb{S}, \mathcal{F})$ via $(f\mu)(A) = \mu(f I_A)$ for all $A \in \mathcal{F}$. These measures satisfy the composition relation $h(f\mu) = (hf)\mu$ for all $f, h \in m\mathcal{F}_+$. Further, $h \in L^1(\mathbb{S}, \mathcal{F}, f\mu)$ if and only if $fh \in L^1(\mathbb{S}, \mathcal{F}, \mu)$ and then $(f\mu)(h) = \mu(fh)$.*

PROOF. Fixing $f \in m\mathcal{F}_+$, obviously $f\mu$ is a non-negative set function on $(\mathbb{S}, \mathcal{F})$ with $(f\mu)(\emptyset) = \mu(f I_\emptyset) = \mu(0) = 0$. To check that $f\mu$ is countably additive, hence a measure, let $A = \cup_k A_k$ for a countable collection of disjoint sets $A_k \in \mathcal{F}$. Since $\sum_{k=1}^n f I_{A_k} \uparrow f I_A$, it follows by monotone convergence and linearity of the integral that,

$$\mu(f I_A) = \lim_{n\to\infty} \mu(\sum_{k=1}^n f I_{A_k}) = \lim_{n\to\infty} \sum_{k=1}^n \mu(f I_{A_k}) = \sum_k \mu(f I_{A_k})$$

Thus, $(f\mu)(A) = \sum_k (f\mu)(A_k)$ verifying that $f\mu$ is a measure.

Fixing $f \in m\mathcal{F}_+$, we turn to prove that the identity

$$(f\mu)(h I_A) = \mu(f h I_A) \qquad \forall A \in \mathcal{F}\,, \tag{1.3.13}$$

holds for any $h \in m\mathcal{F}_+$. Since the left side of (1.3.13) is the value assigned to A by the measure $h(f\mu)$ and the right side of this identity is the value assigned to the same set by the measure $(hf)\mu$, this would verify the stated composition rule $h(f\mu) = (hf)\mu$. The proof of (1.3.13) proceeds by applying the standard machine:
Step 1. If $h = I_B$ for $B \in \mathcal{F}$ we have by the definition of the integral of an indicator function that

$$(f\mu)(I_B I_A) = (f\mu)(I_{A\cap B}) = (f\mu)(A \cap B) = \mu(f I_{A\cap B}) = \mu(f I_B I_A)\,,$$

which is (1.3.13).

Step 2. Take $h \in \text{SF}_+$ represented as $h = \sum_{l=1}^n c_l I_{B_l}$ with $c_l \geq 0$ and $B_l \in \mathcal{F}$. Then, by Step 1 and the linearity of the integrals with respect to $f\mu$ and with respect to μ, we see that

$$(f\mu)(hI_A) = \sum_{l=1}^n c_l (f\mu)(I_{B_l} I_A) = \sum_{l=1}^n c_l \mu(f I_{B_l} I_A) = \mu(f \sum_{l=1}^n c_l I_{B_l} I_A) = \mu(fhI_A)\,,$$

again yielding (1.3.13).
Step 3. For any $h \in m\mathcal{F}_+$ there exist $h_n \in \text{SF}_+$ such that $h_n \uparrow h$. By Step 2 we know that $(f\mu)(h_n I_A) = \mu(f h_n I_A)$ for any $A \in \mathcal{F}$ and all n. Further, $h_n I_A \uparrow hI_A$ and $f h_n I_A \uparrow f h I_A$, so by monotone convergence (for both integrals with respect to $f\mu$ and μ),

$$(f\mu)(hI_A) = \lim_{n\to\infty} (f\mu)(h_n I_A) = \lim_{n\to\infty} \mu(f h_n I_A) = \mu(fhI_A)\,,$$

completing the proof of (1.3.13) for all $h \in m\mathcal{F}_+$.
Writing $h \in m\mathcal{F}$ as $h = h_+ - h_-$ with $h_+ = \max(h, 0) \in m\mathcal{F}_+$ and $h_- = -\min(h, 0) \in m\mathcal{F}_+$, it follows from the composition rule that

$$\int h_\pm d(f\mu) = (f\mu)(h_\pm I_{\mathbb{S}}) = h_\pm(f\mu)(\mathbb{S}) = ((h_\pm f)\mu)(\mathbb{S}) = \mu(f h_\pm I_{\mathbb{S}}) = \int f h_\pm d\mu\,.$$

Observing that $fh_\pm = (fh)_\pm$ when $f \in m\mathcal{F}_+$, we thus deduce that h is $f\mu$-integrable if and only if fh is μ-integrable in which case $\int hd(f\mu) = \int fhd\mu$, as stated. □

Fixing a measure space $(\mathbb{S}, \mathcal{F}, \mu)$, every set $D \in \mathcal{F}$ induces a σ-algebra $\mathcal{F}_D = \{A \in \mathcal{F} : A \subseteq D\}$. Let μ_D denote the *restriction* of μ to $(D, \mathcal{F}_D)$. As a corollary of Proposition 1.3.56 we express the integral with respect to μ_D in terms of the original measure μ.

Corollary 1.3.57. *Fixing $D \in \mathcal{F}$ let h_D denote the restriction of $h \in m\mathcal{F}$ to $(D, \mathcal{F}_D)$. Then, $\mu_D(h_D) = \mu(hI_D)$ for any $h \in m\mathcal{F}_+$. Further, $h_D \in L^1(D, \mathcal{F}_D, \mu_D)$ if and only if $hI_D \in L^1(\mathbb{S}, \mathcal{F}, \mu)$, in which case also $\mu_D(h_D) = \mu(hI_D)$.*

Proof. Note that the measure $I_D\mu$ of Proposition 1.3.56 coincides with μ_D on the σ-algebra $\mathcal{F}_D$ and assigns to any set $A \in \mathcal{F}$ the same value it assigns to $A \cap D \in \mathcal{F}_D$. By Definition 1.3.1 this implies that $(I_D\mu)(h) = \mu_D(h_D)$ for any $h \in m\mathcal{F}_+$. The corollary is thus a re-statement of the composition and integrability relations of Proposition 1.3.56 for $f = I_D$. □

Remark 1.3.58. Corollary 1.3.57 justifies using hereafter the notation $\int_A f d\mu$ or $\mu(f; A)$ for $\mu(fI_A)$, or writing $\mathbf{E}(X; A) = \int_A X(\omega) dP(\omega)$ for $\mathbf{E}(XI_A)$. With this notation in place, Proposition 1.3.56 states that each $Z \geq 0$ such that $\mathbf{E}Z = 1$ induces a probability measure $\mathbf{Q} = Z\mathbf{P}$ such that $\mathbf{Q}(A) = \int_A Z d\mathbf{P}$ for all $A \in \mathcal{F}$, and then $\mathbf{E}_{\mathbf{Q}}(W) := \int W d\mathbf{Q} = \mathbf{E}(ZW)$ whenever $W \geq 0$ or $ZW \in L^1(\Omega, \mathcal{F}, \mathbf{P})$ (the assumption $\mathbf{E}Z = 1$ translates to $\mathbf{Q}(\Omega) = 1$).

Proposition 1.3.56 is closely related to the probability density function of Definition 1.2.39. En-route to showing this, we first define the collection of Lebesgue integrable functions.

Definition 1.3.59. *Consider Lebesgue's measure λ on $(\mathbb{R},\mathcal{B})$ as in Section 1.1.3, and its* completion *$\overline{\lambda}$ on $(\mathbb{R},\overline{\mathcal{B}})$ (see Theorem 1.1.35). A set $B \in \overline{\mathcal{B}}$ is called* Lebesgue measurable *and $f : \mathbb{R} \mapsto \mathbb{R}$ is called* Lebesgue integrable function *if $f \in m\overline{\mathcal{B}}$, and $\overline{\lambda}(|f|) < \infty$. As we show in Proposition 1.3.64, any non-negative Riemann integrable function is also Lebesgue integrable, and the integral values coincide, justifying the notation $\int_B f(x)dx$ for $\overline{\lambda}(f;B)$, where the function f and the set B are both Lebesgue measurable.*

Example 1.3.60. *Suppose f is a non-negative Lebesgue integrable function such that $\int_{\mathbb{R}} f(x)dx = 1$. Then, $\mathcal{P} = f\overline{\lambda}$ of Proposition 1.3.56 is a probability measure on $(\mathbb{R},\overline{\mathcal{B}})$ such that $\mathcal{P}(B) = \overline{\lambda}(f;B) = \int_B f(x)dx$ for any Lebesgue measurable set B. By Theorem 1.2.36 it is easy to verify that $F(\alpha) = \mathcal{P}((-\infty,\alpha])$ is a distribution function, such that $F(\alpha) = \int_{-\infty}^{\alpha} f(x)dx$. That is, $\mathcal{P}$ is the law of a R.V. $X : \mathbb{R} \mapsto \mathbb{R}$ whose probability density function is f (c.f. Definition 1.2.39 and Proposition 1.2.44).*

Our next theorem allows us to compute expectations of functions of a R.V. X in the space $(\mathbb{R},\mathcal{B},\mathcal{P}_X)$, using the law of X (c.f. Definition 1.2.33) and calculus, instead of working on the original general probability space. One of its immediate consequences is the "obvious" fact that if $X \overset{\mathcal{D}}{=} Y$ then $\mathbf{E}h(X) = \mathbf{E}h(Y)$ for any non-negative Borel function h.

Theorem 1.3.61 (Change of variables formula). *Let $X : \Omega \mapsto \mathbb{R}$ be a random variable on $(\Omega,\mathcal{F},\mathbf{P})$ and h a Borel measurable function such that $\mathbf{E}h_+(X) < \infty$ or $\mathbf{E}h_-(X) < \infty$. Then,*

$$\int_\Omega h(X(\omega))d\mathbf{P}(\omega) = \int_{\mathbb{R}} h(x)d\mathcal{P}_X(x). \tag{1.3.14}$$

Proof. Apply the standard machine with respect to $h \in m\mathcal{B}$:
Step 1. Taking $h = I_B$ for $B \in \mathcal{B}$, note that by the definition of expectation of indicators

$$\mathbf{E}h(X) = \mathbf{E}[I_B(X(\omega))] = \mathbf{P}(\{\omega : X(\omega) \in B\}) = \mathcal{P}_X(B) = \int h(x)d\mathcal{P}_X(x).$$

Step 2. Representing $h \in \mathrm{SF}_+$ as $h = \sum_{l=1}^m c_l I_{B_l}$ for $c_l \geq 0$, the identity (1.3.14) follows from Step 1 by the linearity of the expectation in both spaces.
Step 3. For $h \in m\mathcal{B}_+$, consider $h_n \in \mathrm{SF}_+$ such that $h_n \uparrow h$. Since $h_n(X(\omega)) \uparrow h(X(\omega))$ for all ω, we get by monotone convergence on $(\Omega,\mathcal{F},\mathbf{P})$, followed by applying Step 2 for h_n, and finally monotone convergence on $(\mathbb{R},\mathcal{B},\mathcal{P}_X)$, that

$$\begin{aligned}\int_\Omega h(X(\omega))d\mathbf{P}(\omega) &= \lim_{n\to\infty} \int_\Omega h_n(X(\omega))d\mathbf{P}(\omega) \\ &= \lim_{n\to\infty} \int_{\mathbb{R}} h_n(x)d\mathcal{P}_X(x) = \int_{\mathbb{R}} h(x)d\mathcal{P}_X(x),\end{aligned}$$

as claimed.
Step 4. Write a Borel function $h(x)$ as $h_+(x) - h_-(x)$. Then, by Step 3, (1.3.14) applies for both non-negative functions h_+ and h_-. Further, at least one of these two identities involves finite quantities. So, taking their difference and using the linearity of the expectation (in both probability spaces), lead to the same result for h. □

Combining Theorem 1.3.61 with Example 1.3.60, we show that the expectation of a Borel function of a R.V. X having a density f_X can be computed by performing calculus type integration on the real line.

Corollary 1.3.62. *Suppose that the distribution function of a R.V. X is of the form (1.2.3) for some Lebesgue integrable function $f_X(x)$. Then, for any Borel measurable function $h : \mathbb{R} \mapsto \mathbb{R}$, the R.V. $h(X)$ is integrable if and only if $\int |h(x)| f_X(x)dx < \infty$, in which case $\mathbf{E}h(X) = \int h(x) f_X(x)dx$. The latter formula applies also for any non-negative Borel function $h(\cdot)$.*

Proof. Recall Example 1.3.60 that the law $\mathcal{P}_X$ of X equals to the probability measure $f_X\overline{\lambda}$. For $h \geq 0$ we thus deduce from Theorem 1.3.61 that $\mathbf{E}h(X) = f_X\overline{\lambda}(h)$, which by the composition rule of Proposition 1.3.56 is given by $\overline{\lambda}(f_X h) = \int h(x) f_X(x)dx$. The decomposition $h = h_+ - h_-$ then completes the proof of the general case. □

Our next task is to compare Lebesgue's integral (of Definition 1.3.1) with Riemann's integral. To this end recall,

Definition 1.3.63. *A function $f : (a, b] \mapsto [0, \infty]$ is* Riemann integrable *with integral $R(f) < \infty$ if for any $\varepsilon > 0$ there exists $\delta = \delta(\varepsilon) > 0$ such that $|\sum_l f(x_l)\lambda(J_l) - R(f)| < \varepsilon$, for any $x_l \in J_l$ and $\{J_l\}$ a finite partition of $(a, b]$ into disjoint subintervals whose length $\lambda(J_l) < \delta$.*

Lebesgue's integral of a function f is based on splitting its *range* to small intervals and approximating $f(s)$ by a constant on the subset of $\mathbb{S}$ for which $f(\cdot)$ falls into each such interval. As such, it accommodates an arbitrary domain $\mathbb{S}$ of the function, in contrast to Riemann's integral where the *domain* of integration is split into small rectangles – hence limited to $\mathbb{R}^d$. As we next show, even for $\mathbb{S} = (a, b]$, if $f \geq 0$ (or more generally, f bounded), is Riemann integrable, then it is also Lebesgue integrable, with the integrals coinciding in value.

Proposition 1.3.64. *If $f(x)$ is a non-negative Riemann integrable function on an interval $(a, b]$, then it is also Lebesgue integrable on $(a, b]$ and $\overline{\lambda}(f) = R(f)$.*

Proof. Let $f_*(J) = \inf\{f(x) : x \in J\}$ and $f^*(J) = \sup\{f(x) : x \in J\}$. Varying x_l over J_l we see that

$$R(f) - \varepsilon \leq \sum_l f_*(J_l)\lambda(J_l) \leq \sum_l f^*(J_l)\lambda(J_l) \leq R(f) + \varepsilon\,, \tag{1.3.15}$$

for any finite partition Π of $(a, b]$ into disjoint subintervals J_l such that $\sup_l \lambda(J_l) \leq \delta$. For any such partition, the non-negative simple functions $\ell(\Pi) = \sum_l f_*(J_l)I_{J_l}$ and $u(\Pi) = \sum_l f^*(J_l)I_{J_l}$ are such that $\ell(\Pi) \leq f \leq u(\Pi)$, whereas $R(f) - \varepsilon \leq \lambda(\ell(\Pi)) \leq \lambda(u(\Pi)) \leq R(f) + \varepsilon$, by (1.3.15). Consider the dyadic partitions Π_n of $(a, b]$ to 2^n intervals of length $(b - a)2^{-n}$ each, such that Π_{n+1} is a refinement of Π_n for each $n = 1, 2, \ldots$. Note that $u(\Pi_n)(x) \geq u(\Pi_{n+1})(x)$ for all $x \in (a, b]$ and any n, hence $u(\Pi_n))(x) \downarrow u_\infty(x)$ a Borel measurable $\overline{\mathbb{R}}$-valued function (see Exercise 1.2.31). Similarly, $\ell(\Pi_n)(x) \uparrow \ell_\infty(x)$ for all $x \in (a, b]$, with ℓ_∞ also Borel measurable, and by the monotonicity of Lebesgue's integral,

$$R(f) \leq \lim_{n\infty} \lambda(\ell(\Pi_n)) \leq \lambda(\ell_\infty) \leq \lambda(u_\infty) \leq \lim_{n\to\infty} \lambda(u(\Pi_n)) \leq R(f)\,.$$

We deduce that $\lambda(u_\infty) = \lambda(\ell_\infty) = R(f)$ for $u_\infty \geq f \geq \ell_\infty$. The set $\{x \in (a, b] : f(x) \neq \ell_\infty(x)\}$ is a subset of the Borel set $\{x \in (a, b] : u_\infty(x) > \ell_\infty(x)\}$ whose

Lebesgue measure is zero (see Lemma 1.3.8). Consequently, f is Lebesgue measurable on $(a,b]$ with $\overline{\lambda}(f) = \lambda(\ell_\infty) = R(f)$ as stated. $\square$

Here is an alternative, direct proof of the fact that $\mathbf{Q}$ in Remark 1.3.58 is a probability measure.

Exercise 1.3.65. *Suppose $\mathbf{E}|X| < \infty$ and $A = \bigcup_n A_n$ for some disjoint sets $A_n \in \mathcal{F}$.*

(a) *Show that then*

$$\sum_{n=0}^{\infty} \mathbf{E}(X; A_n) = \mathbf{E}(X; A)\,,$$

that is, the sum converges absolutely and has the value on the right.

(b) *Deduce from this that for $Z \ge 0$ with $\mathbf{E}Z$ positive and finite, $\mathbf{Q}(A) := \mathbf{E}ZI_A/\mathbf{E}Z$ is a probability measure.*

(c) *Suppose that X and Y are non-negative random variables on the same probability space $(\Omega, \mathcal{F}, \mathbf{P})$ such that $\mathbf{E}X = \mathbf{E}Y < \infty$. Deduce from the preceding that if $\mathbf{E}XI_A = \mathbf{E}YI_A$ for any A in a π-system $\mathcal{A}$ such that $\mathcal{F} = \sigma(\mathcal{A})$, then $X \overset{a.s.}{=} Y$.*

Exercise 1.3.66. *Suppose $\mathcal{P}$ is a probability measure on $(\mathbb{R}, \mathcal{B})$ and $f \ge 0$ is a Borel function such that $\mathcal{P}(B) = \int_B f(x)dx$ for $B = (-\infty, b]$, $b \in \mathbb{R}$. Using the $\pi - \lambda$ theorem show that this identity applies for all $B \in \mathcal{B}$. Building on this result, use the standard machine to directly prove Corollary 1.3.62 (without Proposition 1.3.56).*

1.3.6. Mean, variance and moments. We start with the definition of moments of a random variable.

Definition 1.3.67. *If k is a positive integer then $\mathbf{E}X^k$ is called the kth* moment *of X. When it is well defined, the first moment $m_X = \mathbf{E}X$ is called the* mean. *If $\mathbf{E}X^2 < \infty$, then the* variance *of X is defined to be*

$$\mathsf{Var}(X) = \mathbf{E}(X - m_X)^2 = \mathbf{E}X^2 - m_X^2 \le \mathbf{E}X^2\,. \tag{1.3.16}$$

Since $\mathbf{E}(aX + b) = a\mathbf{E}X + b$ (linearity of the expectation), it follows from the definition that

$$\mathsf{Var}(aX+b) = \mathbf{E}(aX + b - \mathbf{E}(aX+b))^2 = a^2\mathbf{E}(X - m_X)^2 = a^2\,\mathsf{Var}(X) \tag{1.3.17}$$

We turn to some examples, starting with R.V. having a density.

Example 1.3.68. *If X has the* exponential distribution *then*

$$\mathbf{E}X^k = \int_0^\infty x^k e^{-x} dx = k!$$

for any k (see Example 1.2.40 for its density). The mean of X is $m_X = 1$ and its variance is $\mathbf{E}X^2 - (\mathbf{E}X)^2 = 1$. For any $\lambda > 0$, it is easy to see that $T = X/\lambda$ has density $f_T(t) = \lambda e^{-\lambda t}\mathbf{1}_{t>0}$, called the exponential density *of parameter λ. By (1.3.17) it follows that $m_T = 1/\lambda$ and $\mathsf{Var}(T) = 1/\lambda^2$.*

Similarly, if X has a standard normal distribution, then by symmetry, for k odd,

$$\mathbf{E}X^k = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^k e^{-x^2/2} dx = 0\,,$$

whereas by integration by parts, the even moments satisfy the relation

$$\mathbf{E}X^{2\ell} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^{2\ell-1} x e^{-x^2/2} dx = (2\ell - 1)\mathbf{E}X^{2\ell-2}, \tag{1.3.18}$$

for $\ell = 1, 2, \ldots$. *In particular,*

$$\mathsf{Var}(X) = \mathbf{E}X^2 = 1.$$

Consider $G = \sigma X + \mu$, *where* $\sigma > 0$ *and* $\mu \in \mathbb{R}$, *whose density is*

$$f_G(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}.$$

We call the law of G the normal distribution *of mean* μ *and variance* σ^2 *(as* $\mathbf{E}G = \mu$ *and* $\mathsf{Var}(G) = \sigma^2$*).*

Next are some examples of R.V. with finite or countable set of possible values.

Example 1.3.69. *We say that* B *has a* Bernoulli distribution *of parameter* $p \in [0, 1]$ *if* $\mathbf{P}(B = 1) = 1 - \mathbf{P}(B = 0) = p$. *Clearly,*

$$\mathbf{E}B = p \cdot 1 + (1 - p) \cdot 0 = p.$$

Further, $B^2 = B$ *so* $\mathbf{E}B^2 = \mathbf{E}B = p$ *and*

$$\mathsf{Var}(B) = \mathbf{E}B^2 - (\mathbf{E}B)^2 = p - p^2 = p(1 - p).$$

Recall that N *has a* Poisson distribution *with parameter* $\lambda \geq 0$ *if*

$$\mathbf{P}(N = k) = \frac{\lambda^k}{k!} e^{-\lambda} \quad \text{for} \quad k = 0, 1, 2, \ldots$$

(where in case $\lambda = 0$, $\mathbf{P}(N = 0) = 1$*). Observe that for* $k = 1, 2, \ldots$,

$$\begin{aligned}\mathbf{E}(N(N-1)\cdots(N-k+1)) &= \sum_{n=k}^{\infty} n(n-1)\cdots(n-k+1)\frac{\lambda^n}{n!}e^{-\lambda} \\ &= \lambda^k \sum_{n=k}^{\infty} \frac{\lambda^{n-k}}{(n-k)!} e^{-\lambda} = \lambda^k.\end{aligned}$$

Using this formula, it follows that $\mathbf{E}N = \lambda$ *while*

$$\mathsf{Var}(N) = \mathbf{E}N^2 - (\mathbf{E}N)^2 = \lambda.$$

The random variable Z *is said to have a* Geometric distribution *of success probability* $p \in (0, 1)$ *if*

$$\mathbf{P}(Z = k) = p(1 - p)^{k-1} \quad \text{for} \quad k = 1, 2, \ldots$$

This is the distribution of the number of independent coin tosses needed till the first appearance of a Head, or more generally, the number of independent trials till the first occurrence in this sequence of a specific event whose probability is p. *Then,*

$$\begin{aligned}\mathbf{E}Z &= \sum_{k=1}^{\infty} kp(1-p)^{k-1} = \frac{1}{p} \\ \mathbf{E}Z(Z-1) &= \sum_{k=2}^{\infty} k(k-1)p(1-p)^{k-1} = \frac{2(1-p)}{p^2} \\ \mathsf{Var}(Z) &= \mathbf{E}Z(Z-1) + \mathbf{E}Z - (\mathbf{E}Z)^2 = \frac{1-p}{p^2}.\end{aligned}$$

Exercise 1.3.70. *Consider a counting random variable* $N_n = \sum_{i=1}^n I_{A_i}$.

(a) *Provide a formula for* $\mathsf{Var}(N_n)$ *in terms of* $\mathbf{P}(A_i)$ *and* $\mathbf{P}(A_i \cap A_j)$ *for* $i \neq j$.

(b) *Using your formula, find the variance of the number* N_n *of empty boxes when distributing at random* r *distinct balls among* n *distinct boxes, where each of the possible* n^r *assignments of balls to boxes is equally likely.*

1.4. Independence and product measures

In Subsection 1.4.1 we build-up the notion of independence, from events to random variables via σ-algebras, relating it to the structure of the joint distribution function. Subsection 1.4.2 considers finite product measures associated with the joint law of independent R.V.-s. This is followed by Kolmogorov's extension theorem which we use in order to construct infinitely many independent R.V.-s. Subsection 1.4.3 is about Fubini's theorem and its applications for computing the expectation of functions of independent R.V.

1.4.1. Definition and conditions for independence. Recall the classical definition that two events $A, B \in \mathcal{F}$ are *independent* if $\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B)$.

For example, suppose two fair dice are thrown (i.e. $\Omega = \{1,2,3,4,5,6\}^2$ with $\mathcal{F} = 2^\Omega$ and the uniform probability measure). Let $E_1 = \{\text{Sum of two is } 6\}$ and $E_2 = \{\text{first die is } 4\}$ then E_1 and E_2 are not independent since

$$\mathbf{P}(E_1) = \mathbf{P}(\{(1,5)\ (2,4)\ (3,3)\ (4,2)\ (5,1)\}) = \frac{5}{36}, \quad \mathbf{P}(E_2) = \mathbf{P}(\{\omega : \omega_1 = 4\}) = \frac{1}{6}$$

and

$$\mathbf{P}(E_1 \cap E_2) = \mathbf{P}(\{(4,2)\}) = \frac{1}{36} \neq \mathbf{P}(E_1)\mathbf{P}(E_2).$$

However one can check that E_2 and $E_3 = \{\text{sum of dice is } 7\}$ are independent.

In analogy with the independence of events we define the independence of two random vectors and more generally, that of two σ-algebras.

Definition 1.4.1. *Two* σ*-algebras* $\mathcal{H}, \mathcal{G} \subseteq \mathcal{F}$ *are independent (also denoted* $\mathbf{P}$-independent*), if*

$$\mathbf{P}(G \cap H) = \mathbf{P}(G)\mathbf{P}(H), \qquad \forall G \in \mathcal{G},\ \forall H \in \mathcal{H}\ ,$$

that is, two σ*-algebras are independent if every event in one of them is independent of every event in the other.*

The random vectors $\underline{X} = (X_1, \ldots, X_n)$ *and* $\underline{Y} = (Y_1, \ldots, Y_m)$ *on the same probability space are independent if the corresponding* σ*-algebras* $\sigma(X_1, \ldots, X_n)$ *and* $\sigma(Y_1, \ldots, Y_m)$ *are independent.*

Remark. Our definition of independence of random variables is consistent with that of independence of events. For example, if the events $A, B \in \mathcal{F}$ are independent, then so are I_A and I_B. Indeed, we need to show that $\sigma(I_A) = \{\emptyset, \Omega, A, A^c\}$ and $\sigma(I_B) = \{\emptyset, \Omega, B, B^c\}$ are independent. Since $\mathbf{P}(\emptyset) = 0$ and $\emptyset$ is invariant under intersections, whereas $\mathbf{P}(\Omega) = 1$ and all events are invariant under intersection with Ω, it suffices to consider $G \in \{A, A^c\}$ and $H \in \{B, B^c\}$. We check independence first for $G = A$ and $H = B^c$. Noting that A is the union of the disjoint events $A \cap B$ and $A \cap B^c$ we have that

$$\mathbf{P}(A \cap B^c) = \mathbf{P}(A) - \mathbf{P}(A \cap B) = \mathbf{P}(A)[1 - \mathbf{P}(B)] = \mathbf{P}(A)\mathbf{P}(B^c)\,,$$

where the middle equality is due to the assumed independence of A and B. The proof for all other choices of G and H is very similar.

More generally we define the *mutual* independence of events as follows.

Definition 1.4.2. *Events $A_i \in \mathcal{F}$ are* **P**-mutually independent *if for any $L < \infty$ and distinct indices $i_1, i_2, \ldots, i_L$,*

$$\mathbf{P}(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_L}) = \prod_{k=1}^{L} \mathbf{P}(A_{i_k}).$$

We next generalize the definition of mutual independence to σ-algebras, random variables and beyond. This definition applies to the mutual independence of both finite and infinite number of such objects.

Definition 1.4.3. *We say that the collections of events $\mathcal{A}_\alpha \subseteq \mathcal{F}$ with $\alpha \in \mathcal{I}$ (possibly an infinite index set) are **P**-mutually independent if for any $L < \infty$ and distinct $\alpha_1, \alpha_2, \ldots, \alpha_L \in \mathcal{I}$,*

$$\mathbf{P}(A_1 \cap A_2 \cap \cdots \cap A_L) = \prod_{k=1}^{L} \mathbf{P}(A_k), \qquad \forall A_k \in \mathcal{A}_{\alpha_k},\ k = 1, \ldots, L\,.$$

*We say that random variables X_α, $\alpha \in \mathcal{I}$ are **P**-mutually independent if the σ-algebras $\sigma(X_\alpha)$, $\alpha \in \mathcal{I}$ are **P**-mutually independent.*

*When the probability measure **P** in consideration is clear from the context, we say that random variables, or collections of events, are mutually independent.*

Our next theorem gives a sufficient condition for the mutual independence of a collection of σ-algebras which as we later show, greatly simplifies the task of checking independence.

Theorem 1.4.4. *Suppose $\mathcal{G}_i = \sigma(\mathcal{A}_i) \subseteq \mathcal{F}$ for $i = 1, 2, \cdots, n$ where $\mathcal{A}_i$ are π-systems. Then, a sufficient condition for the mutual independence of $\mathcal{G}_i$ is that $\mathcal{A}_i$, $i = 1, \ldots, n$ are mutually independent.*

Proof. Let $H = A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_L}$, where $i_1, i_2, \ldots, i_L$ are distinct elements from $\{1, 2, \ldots, n-1\}$ and $A_i \in \mathcal{A}_i$ for $i = 1, \ldots, n-1$. Consider the two finite measures $\mu_1(A) = \mathbf{P}(A \cap H)$ and $\mu_2(A) = \mathbf{P}(H)\mathbf{P}(A)$ on the measurable space $(\Omega, \mathcal{G}_n)$. Note that

$$\mu_1(\Omega) = \mathbf{P}(\Omega \cap H) = \mathbf{P}(H) = \mathbf{P}(H)\mathbf{P}(\Omega) = \mu_2(\Omega)\,.$$

If $A \in \mathcal{A}_n$, then by the mutual independence of $\mathcal{A}_i$, $i = 1, \ldots, n$, it follows that

$$\begin{aligned}\mu_1(A) &= \mathbf{P}(A_{i_1} \cap A_{i_2} \cap A_{i_3} \cap \cdots \cap A_{i_L} \cap A) = (\prod_{k=1}^{L} \mathbf{P}(A_{i_k}))\mathbf{P}(A) \\ &= \mathbf{P}(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_L})\mathbf{P}(A) = \mu_2(A)\ .\end{aligned}$$

Since the finite measures $\mu_1(\cdot)$ and $\mu_2(\cdot)$ agree on the π-system $\mathcal{A}_n$ and on Ω, it follows that $\mu_1 = \mu_2$ on $\mathcal{G}_n = \sigma(\mathcal{A}_n)$ (see Proposition 1.1.39). That is, $\mathbf{P}(G \cap H) = \mathbf{P}(G)\mathbf{P}(H)$ for any $G \in \mathcal{G}_n$.

Since this applies for arbitrary $A_i \in \mathcal{A}_i$, $i = 1, \ldots, n-1$, in view of Definition 1.4.3 we have just proved that if $\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_n$ are mutually independent, then $\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{G}_n$ are mutually independent.

Applying the latter relation for $\mathcal{G}_n, \mathcal{A}_1, \ldots, \mathcal{A}_{n-1}$ (which are mutually independent since Definition 1.4.3 is invariant to a permutation of the order of the collections) we get that $\mathcal{G}_n, \mathcal{A}_1, \ldots, \mathcal{A}_{n-2}, \mathcal{G}_{n-1}$ are mutually independent. After n such iterations we have the stated result. □

Because the mutual independence of the collections of events $\mathcal{A}_\alpha$, $\alpha \in \mathcal{I}$ amounts to the mutual independence of any finite number of these collections, we have the immediate consequence:

Corollary 1.4.5. *If π-systems of events $\mathcal{A}_\alpha$, $\alpha \in \mathcal{I}$, are mutually independent, then $\sigma(\mathcal{A}_\alpha)$, $\alpha \in \mathcal{I}$, are also mutually independent.*

Another immediate consequence deals with the closure of mutual independence under projections.

Corollary 1.4.6. *If the π-systems of events $\mathcal{H}_{\alpha,\beta}$, $(\alpha, \beta) \in \mathcal{J}$ are mutually independent, then the σ-algebras $\mathcal{G}_\alpha = \sigma(\cup_\beta \mathcal{H}_{\alpha,\beta})$, are also mutually independent.*

Proof. Let $\mathcal{A}_\alpha$ be the collection of sets of the form $A = \cap_{j=1}^m H_j$ where $H_j \in \mathcal{H}_{\alpha,\beta_j}$ for some $m < \infty$ and distinct $\beta_1, \ldots, \beta_m$. Since $\mathcal{H}_{\alpha,\beta}$ are π-systems, it follows that so is $\mathcal{A}_\alpha$ for each α. Since a finite intersection of sets $A_k \in \mathcal{A}_{\alpha_k}$, $k = 1, \ldots, L$ is merely a finite intersection of sets from distinct collections $\mathcal{H}_{\alpha_k, \beta_j(k)}$, the assumed mutual independence of $\mathcal{H}_{\alpha,\beta}$ implies the mutual independence of $\mathcal{A}_\alpha$. By Corollary 1.4.5, this in turn implies the mutual independence of $\sigma(\mathcal{A}_\alpha)$. To complete the proof, simply note that for any β, each $H \in \mathcal{H}_{\alpha,\beta}$ is also an element of $\mathcal{A}_\alpha$, implying that $\mathcal{G}_\alpha \subseteq \sigma(\mathcal{A}_\alpha)$. □

Relying on the preceding corollary you can now establish the following characterization of independence (which is key to proving Kolmogorov's 0-1 law).

Exercise 1.4.7. *Show that if for each $n \geq 1$ the σ-algebras $\mathcal{F}_n^{\mathbf{X}} = \sigma(X_1, \ldots, X_n)$ and $\sigma(X_{n+1})$ are $\mathbf{P}$-mutually independent then the random variables $X_1, X_2, X_3, \ldots$ are $\mathbf{P}$-mutually independent. Conversely, show that if $X_1, X_2, X_3, \ldots$ are independent, then for each $n \geq 1$ the σ-algebras $\mathcal{F}_n^{\mathbf{X}}$ and $\mathcal{T}_n^{\mathbf{X}} = \sigma(X_r, r > n)$ are independent.*

It is easy to check that a $\mathbf{P}$-trivial σ-algebra $\mathcal{H}$ is $\mathbf{P}$-independent of any other σ-algebra $\mathcal{G} \subseteq \mathcal{F}$. Conversely, as we show next, independence is a great tool for proving that a σ-algebra is $\mathbf{P}$-trivial.

Lemma 1.4.8. *If each of the σ-algebras $\mathcal{G}_k \subseteq \mathcal{G}_{k+1}$ is $\mathbf{P}$-independent of a σ-algebra $\mathcal{H} \subseteq \sigma(\bigcup_{k \geq 1} \mathcal{G}_k)$ then $\mathcal{H}$ is $\mathbf{P}$-trivial.*

Remark. In particular, if $\mathcal{H}$ is $\mathbf{P}$-independent of itself, then $\mathcal{H}$ is $\mathbf{P}$-trivial.

Proof. Since $\mathcal{G}_k \subseteq \mathcal{G}_{k+1}$ for all k and $\mathcal{G}_k$ are σ-algebras, it follows that $\mathcal{A} = \bigcup_{k \geq 1} \mathcal{G}_k$ is a π-system. The assumed $\mathbf{P}$-independence of $\mathcal{H}$ and $\mathcal{G}_k$ for each k yields the $\mathbf{P}$-independence of $\mathcal{H}$ and $\mathcal{A}$. Thus, by Theorem 1.4.4 we have that $\mathcal{H}$ and $\sigma(\mathcal{A})$ are $\mathbf{P}$-independent. Since $\mathcal{H} \subseteq \sigma(\mathcal{A})$ it follows that in particular $\mathbf{P}(H) = \mathbf{P}(H \cap H) = \mathbf{P}(H)\,\mathbf{P}(H)$ for each $H \in \mathcal{H}$. So, necessarily $\mathbf{P}(H) \in \{0, 1\}$ for all $H \in \mathcal{H}$. That is, $\mathcal{H}$ is $\mathbf{P}$-trivial. □

We next define the tail σ-algebra of a stochastic process.

Definition 1.4.9. *For a stochastic process $\{X_k\}$ we set $\mathcal{T}_n^{\mathbf{X}} = \sigma(X_r, r > n)$ and call $\mathcal{T}^{\mathbf{X}} = \cap_n \mathcal{T}_n^{\mathbf{X}}$ the* tail σ-algebra *of the process $\{X_k\}$.*

As we next see, the $\mathbf{P}$-triviality of the tail σ-algebra of independent random variables is an immediate consequence of Lemma 1.4.8. This result, due to Kolmogorov, is just one of the many *0-1 laws* that exist in probability theory.

Corollary 1.4.10 (Kolmogorov's 0-1 law). *If $\{X_k\}$ are $\mathbf{P}$-mutually independent then the corresponding tail σ-algebra $\mathcal{T}^{\mathbf{X}}$ is $\mathbf{P}$-trivial.*

Proof. Note that $\mathcal{F}_k^{\mathbf{X}} \subseteq \mathcal{F}_{k+1}^{\mathbf{X}}$ and $\mathcal{T}^{\mathbf{X}} \subseteq \mathcal{F}^{\mathbf{X}} = \sigma(X_k, k \geq 1) = \sigma(\bigcup_{k\geq 1} \mathcal{F}_k^{\mathbf{X}})$ (see Exercise 1.2.14 for the latter identity). Further, recall Exercise 1.4.7 that for any $n \geq 1$, the σ-algebras $\mathcal{T}_n^{\mathbf{X}}$ and $\mathcal{F}_n^{\mathbf{X}}$ are $\mathbf{P}$-mutually independent. Hence, each of the σ-algebras $\mathcal{F}_k^{\mathbf{X}}$ is also $\mathbf{P}$-mutually independent of the tail σ-algebra $\mathcal{T}^{\mathbf{X}}$, which by Lemma 1.4.8 is thus $\mathbf{P}$-trivial. □

Out of Corollary 1.4.6 we deduce that functions of disjoint collections of mutually independent random variables are mutually independent.

Corollary 1.4.11. *If R.V. $X_{k,j}$, $1 \leq k \leq m$, $1 \leq j \leq l(k)$ are mutually independent and $f_k : \mathbb{R}^{l(k)} \mapsto \mathbb{R}$ are Borel functions, then $Y_k = f_k(X_{k,1}, \ldots, X_{k,l(k)})$ are mutually independent random variables for $k = 1, \ldots, m$.*

Proof. We apply Corollary 1.4.6 for the index set $\mathcal{J} = \{(k,j) : 1 \leq k \leq m, 1 \leq j \leq l(k)\}$, and mutually independent π-systems $\mathcal{H}_{k,j} = \sigma(X_{k,j})$, to deduce the mutual independence of $\mathcal{G}_k = \sigma(\cup_j \mathcal{H}_{k,j})$. Recall that $\mathcal{G}_k = \sigma(X_{k,j}, 1 \leq j \leq l(k))$ and $\sigma(Y_k) \subseteq \mathcal{G}_k$ (see Definition 1.2.12 and Exercise 1.2.32). We complete the proof by noting that Y_k are mutually independent if and only if $\sigma(Y_k)$ are mutually independent. □

Our next result is an application of Theorem 1.4.4 to the independence of random variables.

Corollary 1.4.12. *Real-valued random variables $X_1, X_2, \ldots, X_m$ on the same probability space $(\Omega, \mathcal{F}, \mathbf{P})$ are mutually independent if and only if*

$$\mathbf{P}(X_1 \leq x_1, \ldots, X_m \leq x_m) = \prod_{i=1}^m \mathbf{P}(X_i \leq x_i)\,,\ \forall x_1, \ldots, x_m \in \mathbb{R}. \tag{1.4.1}$$

Proof. Let $\mathcal{A}_i$ denote the collection of subsets of Ω of the form $X_i^{-1}((-\infty, b])$ for $b \in \mathbb{R}$. Recall that $\mathcal{A}_i$ generates $\sigma(X_i)$ (see Exercise 1.2.11), whereas (1.4.1) states that the π-systems $\mathcal{A}_i$ are mutually independent (by continuity from below of $\mathbf{P}$, taking $x_i \uparrow \infty$ for $i \neq i_1, i \neq i_2, \ldots, i \neq i_L$, has the same effect as taking a subset of distinct indices $i_1, \ldots, i_L$ from $\{1, \ldots, m\}$). So, just apply Theorem 1.4.4 to conclude the proof. □

The condition (1.4.1) for mutual independence of R.V.-s is further simplified when these variables are either discrete valued, or having a density.

Exercise 1.4.13. *Suppose $(X_1, \ldots, X_m)$ are random variables and $(\mathbb{S}_1, \ldots, \mathbb{S}_m)$ are countable sets such that $\mathbf{P}(X_i \in \mathbb{S}_i) = 1$ for $i = 1, \ldots, m$. Show that if*

$$\mathbf{P}(X_1 = x_1, \ldots, X_m = x_m) = \prod_{i=1}^m \mathbf{P}(X_i = x_i)$$

whenever $x_i \in \mathbb{S}_i$, $i = 1, \ldots, m$, then $X_1, \ldots, X_m$ are mutually independent.

EXERCISE 1.4.14. *Suppose the random vector $\underline{X} = (X_1, \ldots, X_m)$ has a* joint probability density function $f_{\underline{X}}(\underline{x}) = g_1(x_1) \cdots g_m(x_m)$. *That is,*

$$\mathbf{P}((X_1, \ldots, X_m) \in A) = \int_A g_1(x_1) \cdots g_m(x_m) dx_1 \ldots dx_m \,, \qquad \forall A \in \mathcal{B}_{\mathbb{R}^m} \,,$$

where g_i are non-negative, Lebesgue integrable functions. Show that then $X_1, \ldots, X_m$ are mutually independent.

Beware that pairwise independence (of each pair A_k, A_j for $k \neq j$), does not imply mutual independence of *all* the events in question and the same applies to three or more random variables. Here is an illustrating example.

EXERCISE 1.4.15. *Consider the sample space $\Omega = \{0, 1, 2\}^2$ with probability measure on $(\Omega, 2^\Omega)$ that assigns equal probability (i.e. $1/9$) to each possible value of $\omega = (\omega_1, \omega_2) \in \Omega$. Then, $X(\omega) = \omega_1$ and $Y(\omega) = \omega_2$ are independent R.V. each taking the values $\{0, 1, 2\}$ with equal (i.e. $1/3$) probability. Define $Z_0 = X$, $Z_1 = (X + Y)\text{mod}3$ and $Z_2 = (X + 2Y)\text{mod}3$.*

(a) *Show that Z_0 is independent of Z_1, Z_0 is independent of Z_2, Z_1 is independent of Z_2, but if we know the value of Z_0 and Z_1, then we also know Z_2.*

(b) *Construct four $\{-1, 1\}$-valued random variables such that any three of them are independent but all four are not.*
Hint: *Consider products of independent random variables.*

Here is a somewhat counter intuitive example about tail σ-algebras, followed by an elaboration on the theme of Corollary 1.4.11.

EXERCISE 1.4.16. *Let $\sigma(\mathcal{A}, \mathcal{A}')$ denote the smallest σ-algebra $\mathcal{G}$ such that any function measurable on $\mathcal{A}$ or on $\mathcal{A}'$ is also measurable on $\mathcal{G}$. Let $W_0, W_1, W_2, \ldots$ be independent random variables with $\mathbf{P}(W_n = +1) = \mathbf{P}(W_n = -1) = 1/2$ for all n. For each $n \geq 1$, define $X_n := W_0 W_1 \ldots W_n$.*
(a) Prove that the variables $X_1, X_2, \ldots$ are independent.
(b) Show that $\mathcal{S} = \sigma(\mathcal{T}_0^{\mathbf{W}}, \mathcal{T}^{\mathbf{X}})$ is a strict *subset of the σ-algebra $\mathcal{F} = \cap_n \sigma(\mathcal{T}_0^{\mathbf{W}}, \mathcal{T}_n^{\mathbf{X}})$.*
Hint: *Show that $W_0 \in m\mathcal{F}$ is independent of $\mathcal{S}$.*

EXERCISE 1.4.17. *Consider random variables $(X_{i,j}, 1 \leq i, j \leq n)$ on the same probability space. Suppose that the σ-algebras $\mathcal{R}_1, \ldots, \mathcal{R}_n$ are $\mathbf{P}$-mutually independent, where $\mathcal{R}_i = \sigma(X_{i,j}, 1 \leq j \leq n)$ for $i = 1, \ldots, n$. Suppose further that the σ-algebras $\mathcal{C}_1, \ldots, \mathcal{C}_n$ are $\mathbf{P}$-mutually independent, where $\mathcal{C}_j = \sigma(X_{i,j}, 1 \leq i \leq n)$. Prove that the random variables $(X_{i,j}, 1 \leq i, j \leq n)$ must then be $\mathbf{P}$-mutually independent.*

We conclude this subsection with an application in number theory.

EXERCISE 1.4.18. *Recall Euler's zeta-function which for real $s > 1$ is given by $\zeta(s) = \sum_{k=1}^{\infty} k^{-s}$. Fixing such s, let X and Y be independent random variables with $\mathbf{P}(X = k) = \mathbf{P}(Y = k) = k^{-s}/\zeta(s)$ for $k = 1, 2, \ldots$.*

(a) *Show that the events $D_p = \{X \text{ is divisible by } p\}$, with p a prime number, are $\mathbf{P}$-mutually independent.*

(b) *By considering the event $\{X = 1\}$, provide a* probabilistic *explanation of Euler's formula $1/\zeta(s) = \prod_p (1 - 1/p^s)$.*

(c) *Show that the probability that no perfect square other than 1 divides X is precisely $1/\zeta(2s)$.*

(d) *Show that* $\mathbf{P}(G = k) = k^{-2s}/\zeta(2s)$, *where* G *is the greatest common divisor of* X *and* Y.

1.4.2. Product measures and Kolmogorov's theorem. Recall Example 1.1.20 that given two measurable spaces $(\Omega_1, \mathcal{F}_1)$ and $(\Omega_2, \mathcal{F}_2)$ the product (measurable) space $(\Omega, \mathcal{F})$ consists of $\Omega = \Omega_1 \times \Omega_2$ and $\mathcal{F} = \mathcal{F}_1 \times \mathcal{F}_2$, which is the same as $\mathcal{F} = \sigma(\mathcal{A})$ for

$$\mathcal{A} = \Big\{ \biguplus_{j=1}^{m} A_j \times B_j : A_j \in \mathcal{F}_1, B_j \in \mathcal{F}_2, m < \infty \Big\},$$

where throughout, $\biguplus$ denotes the union of disjoint subsets of Ω.

We now construct product measures on such product spaces, first for two, then for finitely many, probability (or even σ-finite) measures. As we show thereafter, these product measures are associated with the joint law of independent R.V.-s.

THEOREM 1.4.19. *Given two σ-finite measures ν_i on $(\Omega_i, \mathcal{F}_i)$, $i = 1, 2$, there exists a unique σ-finite measure μ_2 on the product space $(\Omega, \mathcal{F})$ such that*

$$\mu_2(\biguplus_{j=1}^{m} A_j \times B_j) = \sum_{j=1}^{m} \nu_1(A_j)\nu_2(B_j), \quad \forall A_j \in \mathcal{F}_1, B_j \in \mathcal{F}_2, m < \infty .$$

We denote $\mu_2 = \nu_1 \times \nu_2$ and call it the product *of the measures ν_1 and ν_2.*

PROOF. By Carathéodory's extension theorem, it suffices to show that $\mathcal{A}$ is an algebra on which μ_2 is countably additive (see Theorem 1.1.30 for the case of finite measures). To this end, note that $\Omega = \Omega_1 \times \Omega_2 \in \mathcal{A}$. Further, $\mathcal{A}$ is closed under intersections, since

$$\begin{aligned}(\biguplus_{j=1}^{m} A_j \times B_j) \bigcap (\biguplus_{i=1}^{n} C_i \times D_i) &= \biguplus_{i,j} [(A_j \times B_j) \cap (C_i \times D_i)] \\ &= \biguplus_{i,j} (A_j \cap C_i) \times (B_j \cap D_i) .\end{aligned}$$

It is also closed under complementation, for

$$(\biguplus_{j=1}^{m} A_j \times B_j)^c = \bigcap_{j=1}^{m} [(A_j^c \times B_j) \cup (A_j \times B_j^c) \cup (A_j^c \times B_j^c)] .$$

By DeMorgan's law, $\mathcal{A}$ is an algebra.

Note that countable unions of disjoint elements of $\mathcal{A}$ are also countable unions of disjoint elements of the collection $\mathcal{R} = \{A \times B : A \in \mathcal{F}_1, B \in \mathcal{F}_2\}$ of *measurable rectangles*. Hence, if we show that

$$\sum_{j=1}^{m} \nu_1(A_j)\nu_2(B_j) = \sum_{i} \nu_1(C_i)\nu_2(D_i) , \tag{1.4.2}$$

whenever $\biguplus_{j=1}^{m} A_j \times B_j = \biguplus_i (C_i \times D_i)$ for some $m < \infty$, $A_j, C_i \in \mathcal{F}_1$ and $B_j, D_i \in \mathcal{F}_2$, then we deduce that the value of $\mu_2(E)$ is independent of the representation we choose for $E \in \mathcal{A}$ in terms of measurable rectangles, and further that μ_2 is countably additive on $\mathcal{A}$. To this end, note that the preceding set identity amounts to

$$\sum_{j=1}^{m} I_{A_j}(x) I_{B_j}(y) = \sum_{i} I_{C_i}(x) I_{D_i}(y) \qquad \forall x \in \Omega_1, y \in \Omega_2 .$$

Hence, fixing $x \in \Omega_1$, we have that $\varphi(y) = \sum_{j=1}^m I_{A_j}(x) I_{B_j}(y) \in \mathrm{SF}_+$ is the monotone increasing limit of $\psi_n(y) = \sum_{i=1}^n I_{C_i}(x) I_{D_i}(y) \in \mathrm{SF}_+$ as $n \to \infty$. Thus, by linearity of the integral with respect to ν_2 and monotone convergence,

$$g(x) := \sum_{j=1}^m \nu_2(B_j) I_{A_j}(x) = \nu_2(\varphi) = \lim_{n\to\infty} \nu_2(\psi_n) = \lim_{n\to\infty} \sum_{i=1}^n I_{C_i}(x) \nu_2(D_i)\,.$$

We deduce that the non-negative $g(x) \in m\mathcal{F}_1$ is the monotone increasing limit of the non-negative measurable functions $h_n(x) = \sum_{i=1}^n \nu_2(D_i) I_{C_i}(x)$. Hence, by the same reasoning,

$$\sum_{j=1}^m \nu_2(B_j)\nu_1(A_j) = \nu_1(g) = \lim_{n\to\infty} \nu_1(h_n) = \sum_i \nu_2(D_i)\nu_1(C_i)\,,$$

proving (1.4.2) and the theorem. □

It follows from Theorem 1.4.19 by induction on n that given any finite collection of σ-finite measure spaces $(\Omega_i, \mathcal{F}_i, \nu_i)$, $i = 1, \ldots, n$, there exists a unique *product measure* $\mu_n = \nu_1 \times \cdots \times \nu_n$ on the product space $(\Omega, \mathcal{F})$ (i.e., $\Omega = \Omega_1 \times \cdots \times \Omega_n$ and $\mathcal{F} = \sigma(A_1 \times \cdots \times A_n; A_i \in \mathcal{F}_i, i = 1, \ldots, n)$), such that

$$\mu_n(A_1 \times \cdots \times A_n) = \prod_{i=1}^n \nu_i(A_i) \qquad \forall A_i \in \mathcal{F}_i, \;\; i = 1, \ldots, n. \tag{1.4.3}$$

Remark 1.4.20. A notable special case of this construction is when $\Omega_i = \mathbb{R}$ with the Borel σ-algebra and Lebesgue measure λ of Section 1.1.3. The product space is then $\mathbb{R}^n$ with its Borel σ-algebra and the product measure is λ^n, the Lebesgue measure on $\mathbb{R}^n$.

The notion of the *law* $\mathcal{P}_X$ of a real-valued random variable X as in Definition 1.2.33, naturally extends to the *joint law* $\mathcal{P}_{\underline{X}}$ of a random vector $\underline{X} = (X_1, \ldots, X_n)$ which is the probability measure $\mathcal{P}_{\underline{X}} = \mathbf{P} \circ \underline{X}^{-1}$ on $(\mathbb{R}^n, \mathcal{B}_{\mathbb{R}^n})$.

We next characterize the joint law of independent random variables $X_1, \ldots, X_n$ as the product of the laws of X_i for $i = 1, \ldots, n$.

Proposition 1.4.21. *Random variables $X_1, \ldots, X_n$ on the same probability space, having laws $\nu_i = \mathcal{P}_{X_i}$, are mutually independent if and only if their joint law is $\mu_n = \nu_1 \times \cdots \times \nu_n$.*

Proof. By Definition 1.4.3 and the identity (1.4.3), if $X_1, \ldots, X_n$ are mutually independent then for $B_i \in \mathcal{B}$,

$$\begin{aligned}\mathcal{P}_{\underline{X}}(B_1 \times \cdots \times B_n) &= \mathbf{P}(X_1 \in B_1, \ldots, X_n \in B_n)\\ &= \prod_{i=1}^n \mathbf{P}(X_i \in B_i) = \prod_{i=1}^n \nu_i(B_i) = \nu_1 \times \cdots \times \nu_n(B_1 \times \cdots \times B_n)\,.\end{aligned}$$

This shows that the law of $(X_1, \ldots, X_n)$ and the product measure μ_n agree on the collection of all measurable rectangles $B_1 \times \cdots \times B_n$, a π-system that generates $\mathcal{B}_{\mathbb{R}^n}$ (see Exercise 1.1.21). Consequently, these two probability measures agree on $\mathcal{B}_{\mathbb{R}^n}$ (c.f. Proposition 1.1.39).

Conversely, if $\mathcal{P}_{\underline{X}} = \nu_1 \times \cdots \times \nu_n$, then by same reasoning, for Borel sets B_i,

$$\begin{aligned}\mathbf{P}(\bigcap_{i=1}^{n}\{\omega : X_i(\omega) \in B_i\}) &= \mathcal{P}_{\underline{X}}(B_1 \times \cdots \times B_n) = \nu_1 \times \cdots \times \nu_n(B_1 \times \cdots \times B_n) \\ &= \prod_{i=1}^{n} \nu_i(B_i) = \prod_{i=1}^{n} \mathbf{P}(\{\omega : X_i(\omega) \in B_i\}) \,,\end{aligned}$$

which amounts to the mutual independence of $X_1, \ldots, X_n$. □

We wish to extend the construction of product measures to that of an infinite collection of independent random variables. To this end, let $\mathbf{N} = \{1, 2, \ldots\}$ denote the set of natural numbers and $\mathbb{R}^{\mathbf{N}} = \{\mathbf{x} = (x_1, x_2, \ldots) : x_i \in \mathbb{R}\}$ denote the collection of all infinite sequences of real numbers. We equip $\mathbb{R}^{\mathbf{N}}$ with the product σ-algebra $\mathcal{B}_c = \sigma(\mathcal{R})$ generated by the collection $\mathcal{R}$ of all finite dimensional measurable rectangles (also called *cylinder sets*), that is sets of the form $\{\mathbf{x} : x_1 \in B_1, \ldots, x_n \in B_n\}$, where $B_i \in \mathcal{B}$, $i = 1, \ldots, n \in \mathbf{N}$ (e.g. see Example 1.1.19).

Kolmogorov's extension theorem provides the existence of a unique probability measure $\mathbf{P}$ on $(\mathbb{R}^{\mathbf{N}}, \mathcal{B}_c)$ whose projections coincide with a given consistent sequence of probability measures μ_n on $(\mathbb{R}^n, \mathcal{B}_{\mathbb{R}^n})$.

Theorem 1.4.22 (Kolmogorov's extension theorem). *Suppose we are given probability measures μ_n on $(\mathbb{R}^n, \mathcal{B}_{\mathbb{R}^n})$ that are* consistent*, that is,*

$$\mu_{n+1}(B_1 \times \cdots \times B_n \times \mathbb{R}) = \mu_n(B_1 \times \cdots \times B_n) \qquad \forall B_i \in \mathcal{B}, \quad i = 1, \ldots, n < \infty$$

Then, there is a unique probability measure $\mathbf{P}$ on $(\mathbb{R}^{\mathbf{N}}, \mathcal{B}_c)$ such that

$$\mathbf{P}(\{\omega : \omega_i \in B_i, i = 1, \ldots, n\}) = \mu_n(B_1 \times \cdots \times B_n)\ \forall B_i \in \mathcal{B},\ i \le n < \infty \tag{1.4.4}$$

Proof. (sketch only) We take a similar approach as in the proof of Theorem 1.4.19. That is, we use (1.4.4) to define the non-negative set function $\mathbf{P}_0$ on the collection $\mathcal{R}$ of all finite dimensional measurable rectangles, where by the consistency of $\{\mu_n\}$ the value of $\mathbf{P}_0$ is independent of the specific representation chosen for a set in $\mathcal{R}$. Then, we extend $\mathbf{P}_0$ to a finitely additive set function on the algebra

$$\mathcal{A} = \Big\{ \biguplus_{j=1}^{m} E_j : E_j \in \mathcal{R}, m < \infty \Big\} \,,$$

in the same linear manner we used when proving Theorem 1.4.19. Since $\mathcal{A}$ generates $\mathcal{B}_c$ and $\mathbf{P}_0(\mathbb{R}^{\mathbf{N}}) = \mu_n(\mathbb{R}^n) = 1$, by Carathéodory's extension theorem it suffices to check that $\mathbf{P}_0$ is countably additive on $\mathcal{A}$. The countable additivity of $\mathbf{P}_0$ is verified by the method we already employed when dealing with Lebesgue's measure. That is, by the remark after Lemma 1.1.31, it suffices to prove that $\mathbf{P}_0(H_n) \downarrow 0$ whenever $H_n \in \mathcal{A}$ and $H_n \downarrow \emptyset$. The proof by contradiction of the latter, adapting the argument of Lemma 1.1.31, is based on approximating each $H \in \mathcal{A}$ by a finite union $J_k \subseteq H$ of *compact* rectangles, such that $\mathbf{P}_0(H \setminus J_k) \to 0$ as $k \to \infty$. This is done for example in [**Dur03**, Lemma A.7.2] or [**Bil95**, Page 490]. □

Example 1.4.23. *To systematically construct an infinite sequence of independent random variables $\{X_i\}$ of prescribed laws $\mathcal{P}_{X_i} = \nu_i$, we apply Kolmogorov's extension theorem for the product measures $\mu_n = \nu_1 \times \cdots \times \nu_n$ constructed following Theorem 1.4.19 (where it is by definition that the sequence μ_n is consistent). Alternatively, for infinite product measures one can take arbitrary probability spaces*

$(\Omega_i, \mathcal{F}_i, \nu_i)$ and directly show by contradiction that $\mathbf{P}_0(H_n) \downarrow 0$ whenever $H_n \in \mathcal{A}$ and $H_n \downarrow \emptyset$ (for more details, see [**Str93**, Exercise 1.1.14]*).*

Remark. As we shall find in Sections 6.1 and 7.1, Kolmogorov's extension theorem is the key to the study of *stochastic processes*, where it relates the law of the process to its finite dimensional distributions. Certain properties of $\mathbb{R}$ are key to the proof of Kolmogorov's extension theorem which indeed is false if $(\mathbb{R}, \mathcal{B})$ is replaced with an arbitrary measurable space $(\mathbb{S}, \mathcal{S})$ (see the discussions in [**Dur03**, Section 1.4c] and [**Dud89**, notes for Section 12.1]). Nevertheless, as you show next, the conclusion of this theorem applies for any *$\mathcal{B}$-isomorphic* measurable space $(\mathbb{S}, \mathcal{S})$.

Definition 1.4.24. *Two measurable spaces $(\mathbb{S}, \mathcal{S})$ and $(\mathbb{T}, \mathcal{T})$ are isomorphic if there exists a one to one and onto measurable mapping between them whose inverse is also a measurable mapping. A measurable space $(\mathbb{S}, \mathcal{S})$ is $\mathcal{B}$-isomorphic if it is isomorphic to a Borel subset $\mathbb{T}$ of $\mathbb{R}$ equipped with the* induced Borel σ-algebra *$\mathcal{T} = \{B \cap \mathbb{T} : B \in \mathcal{B}\}$.*

Here is our generalized version of Kolmogorov's extension theorem.

Corollary 1.4.25. *Given a measurable space $(\mathbb{S}, \mathcal{S})$ let $\mathbb{S}^{\mathbf{N}}$ denote the collection of all infinite sequences of elements in $\mathbb{S}$ equipped the product σ-algebra $\mathcal{S}_c$ generated by the collection of all cylinder sets of the form $\{\mathbf{s} : s_1 \in A_1, \ldots, s_n \in A_n\}$, where $A_i \in \mathcal{S}$ for $i = 1, \ldots, n$. If $(\mathbb{S}, \mathcal{S})$ is $\mathcal{B}$-isomorphic then for any consistent sequence of probability measures ν_n on $(\mathbb{S}^n, \mathcal{S}^n)$ (that is, $\nu_{n+1}(A_1 \times \cdots \times A_n \times \mathbb{S}) = \nu_n(A_1 \times \cdots \times A_n)$ for all n and $A_i \in \mathcal{S}$), there exists a unique probability measure $\mathbf{Q}$ on $(\mathbb{S}^{\mathbf{N}}, \mathcal{S}_c)$ such that for all n and $A_i \in \mathcal{S}$,*

$$\mathbf{Q}(\{\mathbf{s} : s_i \in A_i, i = 1, \ldots, n\}) = \nu_n(A_1 \times \cdots \times A_n) \,. \tag{1.4.5}$$

Next comes a guided proof of Corollary 1.4.25 out of Theorem 1.4.22.

Exercise 1.4.26.

(a) *Verify that our proof of Theorem 1.4.22 applies in case $(\mathbb{R}, \mathcal{B})$ is replaced by $\mathbb{T} \in \mathcal{B}$ equipped with the induced Borel σ-algebra $\mathcal{T}$ (with $\mathbb{R}^{\mathbf{N}}$ and $\mathcal{B}_c$ replaced by $\mathbb{T}^{\mathbf{N}}$ and $\mathcal{T}_c$, respectively).*

(b) *Fixing such $(\mathbb{T}, \mathcal{T})$ and $(\mathbb{S}, \mathcal{S})$ isomorphic to it, let $g : \mathbb{S} \mapsto \mathbb{T}$ be one to one and onto such that both g and g^{-1} are measurable. Check that the one to one and onto mappings $g_n(\mathbf{s}) = (g(s_1), \ldots, g(s_n))$ are measurable and deduce that $\mu_n(B) = \nu_n(g_n^{-1}(B))$ are consistent probability measures on $(\mathbb{T}^n, \mathcal{T}^n)$.*

(c) *Consider the one to one and onto mapping $g_\infty(\mathbf{s}) = (g(s_1), \ldots, g(s_n), \ldots)$ from $\mathbb{S}^{\mathbf{N}}$ to $\mathbb{T}^{\mathbf{N}}$ and the unique probability measure $\mathbf{P}$ on $(\mathbb{T}^{\mathbf{N}}, \mathcal{T}_c)$ for which (1.4.4) holds. Verify that $\mathcal{S}_c$ is contained in the σ-algebra of subsets A of $\mathbb{S}^{\mathbf{N}}$ for which $g_\infty(A)$ is in $\mathcal{T}_c$ and deduce that $\mathbf{Q}(A) = \mathbf{P}(g_\infty(A))$ is a probability measure on $(\mathbb{S}^{\mathbf{N}}, \mathcal{S}_c)$.*

(d) *Conclude your proof of Corollary 1.4.25 by showing that this $\mathbf{Q}$ is the unique probability measure for which (1.4.5) holds.*

Remark. Recall that Carathéodory's extension theorem applies for any σ-finite measure. It follows that, by the same proof as in the preceding exercise, any consistent sequence of σ-finite measures ν_n uniquely determines a σ-finite measure $\mathbf{Q}$ on $(\mathbb{S}^{\mathbf{N}}, \mathcal{S}_c)$ for which (1.4.5) holds, a fact which we use in later parts of this text (for example, in the study of Markov chains in Section 6.1).

Our next proposition shows that in most applications one encounters $\mathcal{B}$-isomorphic measurable spaces (for which Kolmogorov's theorem applies).

PROPOSITION 1.4.27. *If $\mathbb{S} \in \mathcal{B}_M$ for a complete separable metric space M and $\mathcal{S}$ is the restriction of $\mathcal{B}_M$ to $\mathbb{S}$ then $(\mathbb{S}, \mathcal{S})$ is $\mathcal{B}$-isomorphic.*

REMARK. While we do not provide the proof of this proposition, we note in passing that it is an immediate consequence of [**Dud89**, Theorem 13.1.1].

1.4.3. Fubini's theorem and its application.

Returning to $(\Omega, \mathcal{F}, \mu)$ which is the product of two σ-finite measure spaces, as in Theorem 1.4.19, we now prove that:

THEOREM 1.4.28 (FUBINI'S THEOREM). *Suppose $\mu = \mu_1 \times \mu_2$ is the product of the σ-finite measures μ_1 on $(\mathbb{X}, \mathfrak{X})$ and μ_2 on $(\mathbb{Y}, \mathcal{Y})$. If $h \in m\mathcal{F}$ for $\mathcal{F} = \mathfrak{X} \times \mathcal{Y}$ is such that $h \geq 0$ or $\int |h|\, d\mu < \infty$, then,*

$$\begin{aligned} \int_{\mathbb{X}\times\mathbb{Y}} h\, d\mu &= \int_{\mathbb{X}} \left[\int_{\mathbb{Y}} h(x,y)\, d\mu_2(y) \right] d\mu_1(x) \\ &= \int_{\mathbb{Y}} \left[\int_{\mathbb{X}} h(x,y)\, d\mu_1(x) \right] d\mu_2(y) \end{aligned} \tag{1.4.6}$$

REMARK. The iterated integrals on the right side of (1.4.6) are finite and well defined whenever $\int |h| d\mu < \infty$. However, for $h \notin m\mathcal{F}_+$ the inner integrals might be well defined only in the almost everywhere sense.

PROOF OF FUBINI'S THEOREM. Clearly, it suffices to prove the first identity of (1.4.6), as the second immediately follows by exchanging the roles of the two measure spaces. We thus prove Fubini's theorem by showing that

$$y \mapsto h(x,y) \in m\mathcal{Y}, \quad \forall x \in \mathbb{X}, \tag{1.4.7}$$

$$x \mapsto f_h(x) := \int_{\mathbb{Y}} h(x,y)\, d\mu_2(y) \in m\mathfrak{X}, \tag{1.4.8}$$

so the double integral on the right side of (1.4.6) is well defined and

$$\int_{\mathbb{X}\times\mathbb{Y}} h\, d\mu = \int_{\mathbb{X}} f_h(x) d\mu_1(x) \,. \tag{1.4.9}$$

We do so in three steps, first proving (1.4.7)-(1.4.9) for finite measures and bounded h, proceeding to extend these results to non-negative h and σ-finite measures, and then showing that (1.4.6) holds whenever $h \in m\mathcal{F}$ and $\int |h| d\mu$ is finite.
Step 1. Let $\mathcal{H}$ denote the collection of bounded functions on $\mathbb{X}\times\mathbb{Y}$ for which (1.4.7)–(1.4.9) hold. Assuming that both $\mu_1(\mathbb{X})$ and $\mu_2(\mathbb{Y})$ are finite, we deduce that $\mathcal{H}$ contains all bounded $h \in m\mathcal{F}$ by verifying the assumptions of the monotone class theorem (i.e. Theorem 1.2.7) for $\mathcal{H}$ and the π-system $\mathcal{R} = \{A \times B : A \in \mathfrak{X}, B \in \mathcal{Y}\}$ of measurable rectangles (which generates $\mathcal{F}$).

To this end, note that if $h = I_E$ and $E = A \times B \in \mathcal{R}$, then either $h(x,\cdot) = I_B(\cdot)$ (in case $x \in A$), or $h(x,\cdot)$ is identically zero (when $x \notin A$). With $I_B \in m\mathcal{Y}$ we thus have (1.4.7) for any such h. Further, in this case the simple function $f_h(x) = \mu_2(B) I_A(x)$ on $(\mathbb{X}, \mathfrak{X})$ is in $m\mathfrak{X}$ and

$$\int_{\mathbb{X}\times\mathbb{Y}} I_E d\mu = \mu_1 \times \mu_2(E) = \mu_2(B)\mu_1(A) = \int_{\mathbb{X}} f_h(x) d\mu_1(x) \,.$$

Consequently, $I_E \in \mathcal{H}$ for all $E \in \mathcal{R}$; in particular, the constant functions are in $\mathcal{H}$.

Next, with both $m\mathcal{Y}$ and $m\mathfrak{X}$ vector spaces over $\mathbb{R}$, by the linearity of $h \mapsto f_h$ over the vector space of bounded functions satisfying (1.4.7) and the linearity of $f_h \mapsto \mu_1(f_h)$ and $h \mapsto \mu(h)$ over the vector spaces of bounded measurable f_h and h, respectively, we deduce that $\mathcal{H}$ is also a vector space over $\mathbb{R}$.

Finally, if non-negative $h_n \in \mathcal{H}$ are such that $h_n \uparrow h$, then for each $x \in \mathbb{X}$ the mapping $y \mapsto h(x, y) = \sup_n h_n(x, y)$ is in $m\mathcal{Y}_+$ (by Theorem 1.2.22). Further, $f_{h_n} \in m\mathfrak{X}_+$ and by monotone convergence $f_{h_n} \uparrow f_h$ (for all $x \in \mathbb{X}$), so by the same reasoning $f_h \in m\mathfrak{X}_+$. Applying monotone convergence twice more, it thus follows that

$$\mu(h) = \sup_n \mu(h_n) = \sup_n \mu_1(f_{h_n}) = \mu_1(f_h)\,,$$

so h satisfies (1.4.7)–(1.4.9). In particular, if h is bounded then also $h \in \mathcal{H}$.

Step 2. Suppose now that $h \in m\mathcal{F}_+$. If μ_1 and μ_2 are finite measures, then we have shown in Step 1 that (1.4.7)–(1.4.9) hold for the bounded non-negative functions $h_n = h \wedge n$. With $h_n \uparrow h$ we have further seen that (1.4.7)-(1.4.9) hold also for the possibly unbounded h. Further, the closure of (1.4.8) and (1.4.9) with respect to monotone increasing limits of non-negative functions has been shown by monotone convergence, and as such it extends to σ-finite measures μ_1 and μ_2. Turning now to σ-finite μ_1 and μ_2, recall that there exist $E_n = A_n \times B_n \in \mathcal{R}$ such that $A_n \uparrow \mathbb{X}$, $B_n \uparrow \mathbb{Y}$, $\mu_1(A_n) < \infty$ and $\mu_2(B_n) < \infty$. As h is the monotone increasing limit of $h_n = hI_{E_n} \in m\mathcal{F}_+$ it thus suffices to verify that for each n the non-negative $f_n(x) = \int_{\mathbb{Y}} h_n(x, y) d\mu_2(y)$ is measurable with $\mu(h_n) = \mu_1(f_n)$. Fixing n and simplifying our notations to $E = E_n$, $A = A_n$ and $B = B_n$, recall Corollary 1.3.57 that $\mu(h_n) = \mu_E(h_E)$ for the restrictions h_E and μ_E of h and μ to the measurable space $(E, \mathcal{F}_E)$. Also, as $E = A \times B$ we have that $\mathcal{F}_E = \mathfrak{X}_A \times \mathcal{Y}_B$ and $\mu_E = (\mu_1)_A \times (\mu_2)_B$ for the finite measures $(\mu_1)_A$ and $(\mu_2)_B$. Finally, as $f_n(x) = f_{h_E}(x) := \int_B h_E(x, y) d(\mu_2)_B(y)$ when $x \in A$ and zero otherwise, it follows that $\mu_1(f_n) = (\mu_1)_A(f_{h_E})$. We have thus reduced our problem (for h_n), to the case of finite measures $\mu_E = (\mu_1)_A \times (\mu_2)_B$ which we have already successfully resolved.

Step 3. Write $h \in m\mathcal{F}$ as $h = h_+ - h_-$, with $h_\pm \in m\mathcal{F}_+$. By Step 2 we know that $y \mapsto h_\pm(x, y) \in m\mathcal{Y}$ for each $x \in \mathbb{X}$, hence the same applies for $y \mapsto h(x, y)$. Let $\mathbb{X}_0$ denote the subset of $\mathbb{X}$ for which $\int_{\mathbb{Y}} |h(x, y)| d\mu_2(y) < \infty$. By linearity of the integral with respect to μ_2 we have that for all $x \in \mathbb{X}_0$

$$f_h(x) = f_{h_+}(x) - f_{h_-}(x) \tag{1.4.10}$$

is finite. By Step 2 we know that $f_{h_\pm} \in m\mathfrak{X}$, hence $\mathbb{X}_0 = \{x : f_{h_+}(x) + f_{h_-}(x) < \infty\}$ is in $\mathfrak{X}$. From Step 2 we further have that $\mu_1(f_{h_\pm}) = \mu(h_\pm)$ whereby our assumption that $\int |h|\, d\mu = \mu_1(f_{h_+} + f_{h_-}) < \infty$ implies that $\mu_1(\mathbb{X}_0^c) = 0$. Let $\widetilde{f}_h(x) = f_{h_+}(x) - f_{h_-}(x)$ on $\mathbb{X}_0$ and $\widetilde{f}_h(x) = 0$ for all $x \notin \mathbb{X}_0$. Clearly, $\widetilde{f}_h \in m\mathfrak{X}$ is μ_1-almost-everywhere the same as the inner integral on the right side of (1.4.6). Moreover, in view of (1.4.10) and linearity of the integrals with respect to μ_1 and μ we deduce that

$$\mu(h) = \mu(h_+) - \mu(h_-) = \mu_1(f_{h_+}) - \mu_1(f_{h_-}) = \mu_1(\widetilde{f}_h)\,,$$

which is exactly the identity (1.4.6). □

Equipped with Fubini's theorem, we have the following simpler formula for the expectation of a Borel function h of two independent R.V.

THEOREM 1.4.29. *Suppose that X and Y are independent random variables of laws $\mu_1 = \mathcal{P}_X$ and $\mu_2 = \mathcal{P}_Y$. If $h : \mathbb{R}^2 \mapsto \mathbb{R}$ is a Borel measurable function such that $h \geq 0$ or $\mathbf{E}|h(X,Y)| < \infty$, then,*

$$\mathbf{E}h(X,Y) = \int \Big[\int h(x,y)\, d\mu_1(x) \Big]\, d\mu_2(y) \tag{1.4.11}$$

In particular, for Borel functions $f, g : \mathbb{R} \mapsto \mathbb{R}$ such that $f, g \geq 0$ or $\mathbf{E}|f(X)| < \infty$ and $\mathbf{E}|g(Y)| < \infty$,

$$\mathbf{E}(f(X)g(Y)) = \mathbf{E}f(X)\, \mathbf{E}g(Y) \tag{1.4.12}$$

PROOF. Subject to minor changes of notations, the proof of Theorem 1.3.61 applies to any $(\mathbb{S}, \mathcal{S})$-valued R.V. Considering this theorem for the random vector (X, Y) whose joint law is $\mu_1 \times \mu_2$ (c.f. Proposition 1.4.21), together with Fubini's theorem, we see that

$$\mathbf{E}h(X,Y) = \int_{\mathbb{R}^2} h(x,y)\, d(\mu_1 \times \mu_2)(x,y) = \int \Big[\int h(x,y)\, d\mu_1(x) \Big]\, d\mu_2(y)\ ,$$

which is (1.4.11). Take now $h(x,y) = f(x)g(y)$ for non-negative Borel functions $f(x)$ and $g(y)$. In this case, the iterated integral on the right side of (1.4.11) can be further simplified to,

$$\begin{aligned}\mathbf{E}(f(X)g(Y)) &= \int \Big[\int f(x)g(y)\, d\mu_1(x) \Big] d\mu_2(y) = \int g(y) [\int f(x)\, d\mu_1(x)]\, d\mu_2(y) \\ &= \int [\mathbf{E}f(X)] g(y)\, d\mu_2(y) = \mathbf{E}f(X)\, \mathbf{E}g(Y)\end{aligned}$$

(with Theorem 1.3.61 applied twice here), which is the stated identity (1.4.12).

To deal with Borel functions f and g that are not necessarily non-negative, first apply (1.4.12) for the non-negative functions $|f|$ and $|g|$ to get that $\mathbf{E}(|f(X)g(Y)|) = \mathbf{E}|f(X)|\mathbf{E}|g(Y)| < \infty$. Thus, the assumed integrability of $f(X)$ and of $g(Y)$ allows us to apply again (1.4.11) for $h(x,y) = f(x)g(y)$. Now repeat the argument we used for deriving (1.4.12) in case of non-negative Borel functions. □

Another consequence of Fubini's theorem is the following *integration by parts* formula.

LEMMA 1.4.30 (INTEGRATION BY PARTS). *Suppose $H(x) = \int_{-\infty}^{x} h(y)dy$ for a non-negative Borel function h and all $x \in \mathbb{R}$. Then, for any random variable X,*

$$\mathbf{E}H(X) = \int_{\mathbb{R}} h(y) \mathbf{P}(X > y) dy\,. \tag{1.4.13}$$

PROOF. Combining the change of variables formula (Theorem 1.3.61), with our assumption about $H(\cdot)$, we have that

$$\mathbf{E}H(X) = \int_{\mathbb{R}} H(x) d\mathcal{P}_X(x) = \int_{\mathbb{R}} \Big[\int_{\mathbb{R}} h(y) I_{x>y}\, d\lambda(y) \Big] d\mathcal{P}_X(x)\,,$$

where λ denotes Lebesgue's measure on $(\mathbb{R}, \mathcal{B})$. For each $y \in \mathbb{R}$, the expectation of the simple function $x \mapsto h(x,y) = h(y) I_{x>y}$ with respect to $(\mathbb{R}, \mathcal{B}, \mathcal{P}_X)$ is merely $h(y)\mathbf{P}(X > y)$. Thus, applying Fubini's theorem for the non-negative measurable function $h(x,y)$ on the product space $\mathbb{R} \times \mathbb{R}$ equipped with its Borel σ-algebra $\mathcal{B}_{\mathbb{R}^2}$, and the σ-finite measures $\mu_1 = \mathcal{P}_X$ and $\mu_2 = \lambda$, we have that

$$\mathbf{E}H(X) = \int_{\mathbb{R}} \Big[\int_{\mathbb{R}} h(y) I_{x>y}\, d\mathcal{P}_X(x) \Big] d\lambda(y) = \int_{\mathbb{R}} h(y) \mathbf{P}(X > y) dy\,,$$

as claimed. □

Indeed, as we see next, by combining the integration by parts formula with Hölder's inequality we can convert bounds on tail probabilities to bounds on the moments of the corresponding random variables.

LEMMA 1.4.31.

(a) *For any $r > p > 0$ and any random variable $Y \geq 0$,*

$$\mathbf{E}Y^p = \int_0^\infty py^{p-1}\mathbf{P}(Y > y)dy = \int_0^\infty py^{p-1}\mathbf{P}(Y \geq y)dy$$
$$= (1 - \frac{p}{r}) \int_0^\infty py^{p-1}\mathbf{E}[\min(Y/y, 1)^r]dy\,.$$

(b) *If $X, Y \geq 0$ are such that $\mathbf{P}(Y \geq y) \leq y^{-1}\mathbf{E}[XI_{Y \geq y}]$ for all $y > 0$, then $\|Y\|_p \leq q\|X\|_p$ for any $p > 1$ and $q = p/(p-1)$.*

(c) *Under the same hypothesis also $\mathbf{E}Y \leq 1 + \mathbf{E}[X(\log Y)_+]$.*

PROOF. (a) The first identity is merely the integration by parts formula for $h_p(y) = py^{p-1}\mathbf{1}_{y>0}$ and $H_p(x) = x^p\mathbf{1}_{x\geq 0}$ and the second identity follows by the fact that $\mathbf{P}(Y = y) = 0$ up to a (countable) set of zero Lebesgue measure. Finally, it is easy to check that $H_p(x) = \int_{\mathbb{R}} h_{p,r}(x, y)dy$ for the non-negative Borel function $h_{p,r}(x, y) = (1 - p/r)py^{p-1}\min(x/y, 1)^r\mathbf{1}_{x\geq 0}\mathbf{1}_{y>0}$ and any $r > p > 0$. Hence, replacing $h(y)I_{x>y}$ throughout the proof of Lemma 1.4.30 by $h_{p,r}(x, y)$ we find that $\mathbf{E}[H_p(X)] = \int_0^\infty \mathbf{E}[h_{p,r}(X, y)]dy$, which is exactly our third identity.
(b) In a similar manner it follows from Fubini's theorem that for $p > 1$ and any non-negative random variables X and Y

$$\mathbf{E}[XY^{p-1}] = \mathbf{E}[XH_{p-1}(Y)] = \mathbf{E}[\int_{\mathbb{R}} h_{p-1}(y)XI_{Y\geq y}dy] = \int_{\mathbb{R}} h_{p-1}(y)\mathbf{E}[XI_{Y\geq y}]dy\,.$$

Thus, with $y^{-1}h_p(y) = qh_{p-1}(y)$ our hypothesis implies that

$$\mathbf{E}Y^p = \int_{\mathbb{R}} h_p(y)\mathbf{P}(Y \geq y)dy \leq \int_{\mathbb{R}} qh_{p-1}(y)\mathbf{E}[XI_{Y\geq y}]dy = q\mathbf{E}[XY^{p-1}]\,.$$

Applying Hölder's inequality we deduce that

$$\mathbf{E}Y^p \leq q\mathbf{E}[XY^{p-1}] \leq q\|X\|_p\|Y^{p-1}\|_q = q\|X\|_p[\mathbf{E}Y^p]^{1/q}$$

where the right-most equality is due to the fact that $(p-1)q = p$. In case Y is bounded, dividing both sides of the preceding bound by $[\mathbf{E}Y^p]^{1/q}$ implies that $\|Y\|_p \leq q\|X\|_p$. To deal with the general case, let $Y_n = Y \wedge n$, $n = 1, 2, \ldots$ and note that either $\{Y_n \geq y\}$ is empty (for $n < y$) or $\{Y_n \geq y\} = \{Y \geq y\}$. Thus, our assumption implies that $\mathbf{P}(Y_n \geq y) \leq y^{-1}\mathbf{E}[XI_{Y_n\geq y}]$ for all $y > 0$ and $n \geq 1$. By the preceding argument $\|Y_n\|_p \leq q\|X\|_p$ for any n. Taking $n \to \infty$ it follows by monotone convergence that $\|Y\|_p \leq q\|X\|_p$.
(c) Considering part (a) with $p = 1$, we bound $\mathbf{P}(Y \geq y)$ by one for $y \in [0, 1]$ and by $y^{-1}\mathbf{E}[XI_{Y\geq y}]$ for $y > 1$, to get by Fubini's theorem that

$$\mathbf{E}Y = \int_0^\infty \mathbf{P}(Y \geq y)dy \leq 1 + \int_1^\infty y^{-1}\mathbf{E}[XI_{Y\geq y}]dy$$
$$= 1 + \mathbf{E}[X\int_1^\infty y^{-1}I_{Y\geq y}dy] = 1 + \mathbf{E}[X(\log Y)_+]\,.$$

□

We further have the following corollary of (1.4.12), dealing with the expectation of a product of mutually independent R.V.

Corollary 1.4.32. *Suppose that $X_1, \ldots, X_n$ are $\mathbf{P}$-mutually independent random variables such that either $X_i \geq 0$ for all i, or $\mathbf{E}|X_i| < \infty$ for all i. Then,*

$$\mathbf{E}\Big(\prod_{i=1}^{n} X_i\Big) = \prod_{i=1}^{n} \mathbf{E} X_i \,, \tag{1.4.14}$$

that is, the expectation on the left exists and has the value given on the right.

Proof. By Corollary 1.4.11 we know that $X = X_1$ and $Y = X_2 \cdots X_n$ are independent. Taking $f(x) = |x|$ and $g(y) = |y|$ in Theorem 1.4.29, we thus have that $\mathbf{E}|X_1 \cdots X_n| = \mathbf{E}|X_1| \mathbf{E}|X_2 \cdots X_n|$ for any $n \geq 2$. Applying this identity iteratively for $X_l, \ldots, X_n$, starting with $l = m$, then $l = m+1, m+2, \ldots, n-1$ leads to

$$\mathbf{E}|X_m \cdots X_n| = \prod_{k=m}^{n} \mathbf{E}|X_k| \,, \tag{1.4.15}$$

holding for any $1 \leq m \leq n$. If $X_i \geq 0$ for all i, then $|X_i| = X_i$ and we have (1.4.14) as the special case $m = 1$.

To deal with the proof in case $X_i \in L^1$ for all i, note that for $m = 2$ the identity (1.4.15) tells us that $\mathbf{E}|Y| = \mathbf{E}|X_2 \cdots X_n| < \infty$, so using Theorem 1.4.29 with $f(x) = x$ and $g(y) = y$ we have that $\mathbf{E}(X_1 \cdots X_n) = (\mathbf{E}X_1)\mathbf{E}(X_2 \cdots X_n)$. Iterating this identity for $X_l, \ldots, X_n$, starting with $l = 1$, then $l = 2, 3, \ldots, n-1$ leads to the desired result (1.4.14). □

Another application of Theorem 1.4.29 provides us with the familiar formula for the probability density function of the sum $X + Y$ of independent random variables X and Y, having densities f_X and f_Y respectively.

Corollary 1.4.33. *Suppose that R.V. X with a Borel measurable probability density function f_X and R.V. Y with a Borel measurable probability density function f_Y are independent. Then, the random variable $Z = X + Y$ has the probability density function*

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

Proof. Fixing $z \in \mathbb{R}$, apply Theorem 1.4.29 for $h(x, y) = \mathbf{1}_{(x+y \leq z)}$, to get that

$$F_Z(z) = \mathbf{P}(X + Y \leq z) = \mathbf{E}h(X, Y) = \int_{\mathbb{R}} \Big[\int_{\mathbb{R}} h(x, y) d\mathcal{P}_X(x) \Big] d\mathcal{P}_Y(y) \,.$$

Considering the inner integral for a fixed value of y, we have that

$$\int_{\mathbb{R}} h(x, y) d\mathcal{P}_X(x) = \int_{\mathbb{R}} I_{(-\infty, z-y]}(x) d\mathcal{P}_X(x) = \mathcal{P}_X((-\infty, z - y]) = \int_{-\infty}^{z-y} f_X(x) dx \,,$$

where the right most equality is by the existence of a density $f_X(x)$ for X (c.f. Definition 1.2.39). Clearly, $\int_{-\infty}^{z-y} f_X(x) dx = \int_{-\infty}^{z} f_X(x - y) dx$. Thus, applying Fubini's theorem for the Borel measurable function $g(x, y) = f_X(x - y) \geq 0$ and the product of the σ-finite Lebesgue's measure on $(-\infty, z]$ and the probability measure $\mathcal{P}_Y$, we see that

$$F_Z(z) = \int_{\mathbb{R}} \Big[\int_{-\infty}^{z} f_X(x-y)dx \Big] d\mathcal{P}_Y(y) = \int_{-\infty}^{z} \Big[\int_{\mathbb{R}} f_X(x-y) d\mathcal{P}_Y(y) \Big] dx$$

(in this application of Fubini's theorem we replace one iterated integral by another, exchanging the order of integrations). Since this applies for any $z \in \mathbb{R}$, it follows by definition that Z has the probability density

$$f_Z(z) = \int_{\mathbb{R}} f_X(z-y) d\mathcal{P}_Y(y) = \mathbf{E} f_X(z-Y) \,.$$

With Y having density f_Y, the stated formula for f_Z is a consequence of Corollary 1.3.62. □

Definition 1.4.34. *The expression $\int f(z-y)g(y)dy$ is called the* convolution *of the non-negative Borel functions f and g, denoted by $f * g(z)$. The convolution of measures μ and ν on $(\mathbb{R}, \mathcal{B})$ is the measure $\mu * \nu$ on $(\mathbb{R}, \mathcal{B})$ such that $\mu * \nu(B) = \int \mu(B-x) d\nu(x)$ for any $B \in \mathcal{B}$ (where $B - x = \{y : x + y \in B\}$).*

Corollary 1.4.33 states that if two independent random variables X and Y have densities, then so does $Z = X + Y$, whose density is the convolution of the densities of X and Y. Without assuming the existence of densities, one can show by a similar argument that the law of $X + Y$ is the convolution of the law of X and the law of Y (c.f. [**Dur03**, Theorem 1.4.9] or [**Bil95**, Page 266]).

Convolution is often used in analysis to provide a more regular approximation to a given function. Here are few of the reasons for doing so.

Exercise 1.4.35. *Suppose Borel functions f, g are such that g is a probability density and $\int |f(x)|dx$ is finite. Consider the scaled densities $g_n(\cdot) = ng(n\cdot)$, $n \geq 1$.*

(a) *Show that $f * g(y)$ is a Borel function with $\int |f * g(y)|dy \leq \int |f(x)|dx$ and if g is uniformly continuous, then so is $f * g$.*

(b) *Show that if $g(x) = 0$ whenever $|x| \geq 1$, then $f * g_n(y) \to f(y)$ as $n \to \infty$, for any continuous f and each $y \in \mathbb{R}$.*

Next you find two of the many applications of Fubini's theorem in real analysis.

Exercise 1.4.36. *Show that the set $G_f = \{(x, y) \in \mathbb{R}^2 : 0 \leq y \leq f(x)\}$ of points under the graph of a non-negative Borel function $f : \mathbb{R} \mapsto [0, \infty)$ is in $\mathcal{B}_{\mathbb{R}^2}$ and deduce the well-known formula $\lambda \times \lambda(G_f) = \int f(x) d\lambda(x)$, for its area.*

Exercise 1.4.37. *For $n \geq 2$, consider the unit sphere $S^{n-1} = \{\underline{x} \in \mathbb{R}^n : \|\underline{x}\| = 1\}$ equipped with the topology induced by $\mathbb{R}^n$. Let the* surface measure *of $A \in \mathcal{B}_{S^{n-1}}$ be $\nu(A) = n\lambda^n(C_{0,1}(A))$, for $C_{a,b}(A) = \{r\underline{x} : r \in (a, b],\ \underline{x} \in A\}$ and the n-fold product Lebesgue measure λ^n (as in Remark 1.4.20).*

(a) *Check that $C_{a,b}(A) \in \mathcal{B}_{\mathbb{R}^n}$ and deduce that $\nu(\cdot)$ is a finite measure on S^{n-1} (which is further invariant under orthogonal transformations).*

(b) *Verify that $\lambda^n(C_{a,b}(A)) = \frac{b^n - a^n}{n} \nu(A)$ and deduce that for any $B \in \mathcal{B}_{\mathbb{R}^n}$*

$$\lambda^n(B) = \int_0^\infty \Big[\int_{S^{n-1}} I_{r\underline{x} \in B}\, d\nu(\underline{x}) \Big] r^{n-1}\, d\lambda(r) \,.$$

Hint: *Recall that $\lambda^n(\gamma B) = \gamma^n \lambda^n(B)$ for any $\gamma \geq 0$ and $B \in \mathcal{B}_{\mathbb{R}^n}$.*

Combining (1.4.12) with Theorem 1.2.26 leads to the following characterization of the independence between two random vectors (compare with Definition 1.4.1).

Exercise 1.4.38. *Show that the $\mathbb{R}^n$-valued random variable $(X_1, \ldots, X_n)$ and the $\mathbb{R}^m$-valued random variable $(Y_1, \ldots, Y_m)$ are independent if and only if*

$$\mathbf{E}(h(X_1, \ldots, X_n)g(Y_1, \ldots, Y_m)) = \mathbf{E}(h(X_1, \ldots, X_n))\mathbf{E}(g(Y_1, \ldots, Y_m)),$$

for all bounded, Borel measurable functions $g : \mathbb{R}^m \mapsto \mathbb{R}$ and $h : \mathbb{R}^n \mapsto \mathbb{R}$. Then show that the assumption of $h(\cdot)$ and $g(\cdot)$ bounded can be relaxed to both $h(X_1, \ldots, X_n)$ and $g(Y_1, \ldots, Y_m)$ being in $L^1(\Omega, \mathcal{F}, \mathbf{P})$.

Here is another application of (1.4.12):

Exercise 1.4.39. *Show that $\mathbf{E}(f(X)g(X)) \geq (\mathbf{E}f(X))(\mathbf{E}g(X))$ for every random variable X and any bounded non-decreasing functions $f, g : \mathbb{R} \mapsto \mathbb{R}$.*

In the following exercise you bound the exponential moments of certain random variables.

Exercise 1.4.40. *Suppose Y is an integrable random variable such that $\mathbf{E}[e^Y]$ is finite and $\mathbf{E}[Y] = 0$.*

(a) *Show that if $|Y| \leq \kappa$ then*

$$\log \mathbf{E}[e^Y] \leq \kappa^{-2}(e^\kappa - \kappa - 1)\mathbf{E}[Y^2] .$$

Hint: *Use the Taylor expansion of $e^Y - Y - 1$.*

(b) *Show that if $\mathbf{E}[Y^2 e^Y] \leq \kappa^2 \mathbf{E}[e^Y]$ then*

$$\log \mathbf{E}[e^Y] \leq \log \cosh(\kappa) .$$

Hint: *Note that $\varphi(u) = \log \mathbf{E}[e^{uY}]$ is convex, non-negative and finite on $[0, 1]$ with $\varphi(0) = 0$ and $\varphi'(0) = 0$. Verify that $\varphi''(u) + \varphi'(u)^2 = \mathbf{E}[Y^2 e^{uY}]/\mathbf{E}[e^{uY}]$ is non-decreasing on $[0, 1]$ and $\phi(u) = \log \cosh(\kappa u)$ satisfies the differential equation $\phi''(u) + \phi'(u)^2 = \kappa^2$.*

As demonstrated next, Fubini's theorem is also handy in proving the impossibility of certain constructions.

Exercise 1.4.41. *Explain why it is impossible to have $\mathbf{P}$-mutually independent random variables $U_t(\omega)$, $t \in [0, 1]$, on the same probability space $(\Omega, \mathcal{F}, \mathbf{P})$, having each the uniform probability measure on $[-1/2, 1/2]$, such that $t \mapsto U_t(\omega)$ is a Borel function for almost every $\omega \in \Omega$.*
Hint: *Show that $\mathbf{E}[(\int_0^r U_t(\omega)dt)^2] = 0$ for all $r \in [0, 1]$.*

Random variables X and Y such that $\mathbf{E}(X^2) < \infty$ and $\mathbf{E}(Y^2) < \infty$ are called *uncorrelated* if $\mathbf{E}(XY) = \mathbf{E}(X)\mathbf{E}(Y)$. It follows from (1.4.12) that independent random variables X, Y with finite second moment are uncorrelated. While the converse is not necessarily true, it does apply for pairs of random variables that take only two different values each.

Exercise 1.4.42. *Suppose X and Y are uncorrelated random variables.*

(a) *Show that if $X = I_A$ and $Y = I_B$ for some $A, B \in \mathcal{F}$ then X and Y are also independent.*
(b) *Using this, show that if $\{a, b\}$-valued R.V. X and $\{c, d\}$-valued R.V. Y are uncorrelated, then they are also independent.*

(c) *Give an example of a pair of R.V. X and Y that are uncorrelated but not independent.*

Next come a pair of exercises utilizing Corollary 1.4.32.

Exercise 1.4.43. *Suppose X and Y are random variables on the same probability space, X has a Poisson distribution with parameter $\lambda > 0$, and Y has a Poisson distribution with parameter $\mu > \lambda$ (see Example 1.3.69).*

(a) *Show that if X and Y are independent then $\mathbf{P}(X \geq Y) \leq \exp(-(\sqrt{\mu} - \sqrt{\lambda})^2)$.*

(b) *Taking $\mu = \gamma\lambda$ for $\gamma > 1$, find $I(\gamma) > 0$ such that $\mathbf{P}(X \geq Y) \leq 2\exp(-\lambda I(\gamma))$ even when X and Y are not independent.*

Exercise 1.4.44. *Suppose X and Y are independent random variables of identical distribution such that $X > 0$ and $\mathbf{E}[X] < \infty$.*

(a) *Show that $\mathbf{E}[X^{-1}Y] > 1$ unless $X(\omega) = c$ for some non-random c and almost every $\omega \in \Omega$.*

(b) *Provide an example in which $\mathbf{E}[X^{-1}Y] = \infty$.*

We conclude this section with a concrete application of Corollary 1.4.33, computing the density of the sum of mutually independent R.V., each having the same exponential density. To this end, recall

Definition 1.4.45. *The* gamma density *with parameters $\alpha > 0$ and $\lambda > 0$ is given by*

$$f_\Gamma(s) = \Gamma(\alpha)^{-1}\lambda^\alpha s^{\alpha-1}e^{-\lambda s}\mathbf{1}_{s>0}\,,$$

where $\Gamma(\alpha) = \int_0^\infty s^{\alpha-1}e^{-s}ds$ is finite and positive. In particular, $\alpha = 1$ corresponds to the exponential density f_T of Example 1.3.68.

Exercise 1.4.46. *Suppose X has a* gamma *density of parameters α_1 and λ and Y has a* gamma *density of parameters α_2 and λ. Show that if X and Y are independent then $X + Y$ has a* gamma *density of parameters $\alpha_1 + \alpha_2$ and λ. Deduce that if $T_1, \ldots, T_n$ are mutually independent R.V. each having the exponential density of parameter λ, then $W_n = \sum_{i=1}^n T_i$ has the* gamma *density of parameters $\alpha = n$ and λ.*

CHAPTER 2

Asymptotics: the law of large numbers

Building upon the foundations of Chapter 1 we turn to deal with asymptotic theory. To this end, this chapter is devoted to degenerate limit laws, that is, situations in which a sequence of random variables converges to a non-random (constant) limit. Though not exclusively dealing with it, our focus here is on the sequence of empirical averages $n^{-1}\sum_{i=1}^{n} X_i$ as $n \to \infty$.

Section 2.1 deals with the *weak law of large numbers*, where convergence in probability (or in L^q for some $q > 1$) is considered. This is strengthened in Section 2.3 to a *strong law of large numbers*, namely, to convergence almost surely. The key tools for this improvement are the Borel-Cantelli lemmas, to which Section 2.2 is devoted.

2.1. Weak laws of large numbers

A weak law of large numbers corresponds to the situation where the normalized sums of large number of random variables converge in probability to a non-random constant. Usually, the derivation of a weak low involves the computation of variances, on which we focus in Subsection 2.1.1. However, the L^2 convergence we obtain there is of a somewhat limited scope of applicability. To remedy this, we introduce the method of *truncation* in Subsection 2.1.2 and illustrate its power in a few representative examples.

2.1.1. L^2 limits for sums of uncorrelated variables.

The key to our derivation of weak laws of large numbers is the computation of variances. As a preliminary step we define the covariance of two R.V. and extend the notion of a pair of *uncorrelated* random variables, to a (possibly infinite) family of R.V.

Definition 2.1.1. *The* covariance *of two random variables* $X, Y \in L^2(\Omega, \mathcal{F}, \mathbf{P})$ *is*

$$\mathsf{Cov}(X,Y) = \mathbf{E}[(X-\mathbf{E}X)(Y-\mathbf{E}Y)] = \mathbf{E}XY - \mathbf{E}X\mathbf{E}Y\,,$$

so in particular, $\mathsf{Cov}(X,X) = \mathsf{Var}(X)$.
We say that random variables $X_\alpha \in L^2(\Omega, \mathcal{F}, \mathbf{P})$ *are* uncorrelated *if*

$$\mathbf{E}(X_\alpha X_\beta) = \mathbf{E}(X_\alpha)\mathbf{E}(X_\beta) \qquad \forall \alpha \neq \beta\,,$$

or equivalently, if

$$\mathsf{Cov}(X_\alpha, X_\beta) = 0 \qquad \forall \alpha \neq \beta\,.$$

As we next show, the variance of the sum of finitely many uncorrelated random variables is the sum of the variances of the variables.

Lemma 2.1.2. *Suppose* $X_1, \ldots, X_n$ *are uncorrelated random variables (which necessarily are defined on the same probability space). Then,*

$$\mathsf{Var}(X_1 + \cdots + X_n) = \mathsf{Var}(X_1) + \cdots + \mathsf{Var}(X_n)\,. \tag{2.1.1}$$

Proof. Let $S_n = \sum_{i=1}^{n} X_i$. By Definition 1.3.67 of the variance and linearity of the expectation we have that

$$\mathsf{Var}(S_n) = \mathbf{E}([S_n - \mathbf{E}S_n]^2) = \mathbf{E}\big([\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \mathbf{E}X_i]^2\big) = \mathbf{E}\big([\sum_{i=1}^{n} (X_i - \mathbf{E}X_i)]^2\big) .$$

Writing the square of the sum as the sum of all possible cross-products, we get that

$$\begin{aligned}\mathsf{Var}(S_n) &= \sum_{i,j=1}^{n} \mathbf{E}[(X_i - \mathbf{E}X_i)(X_j - \mathbf{E}X_j)] \\ &= \sum_{i,j=1}^{n} \mathsf{Cov}(X_i, X_j) = \sum_{i=1}^{n} \mathsf{Cov}(X_i, X_i) = \sum_{i=1}^{n} \mathsf{Var}(X_i) ,\end{aligned}$$

where we use the fact that $\mathsf{Cov}(X_i, X_j) = 0$ for each $i \neq j$ since X_i and X_j are uncorrelated. □

Equipped with this lemma we have our

Theorem 2.1.3 (L^2 weak law of large numbers). *Consider $S_n = \sum_{i=1}^{n} X_i$ for uncorrelated random variables $X_1, \ldots, X_n, \ldots$. Suppose that $\mathsf{Var}(X_i) \le C$ and $\mathbf{E}X_i = \overline{x}$ for some finite constants C, $\overline{x}$, and all $i = 1, 2, \ldots$. Then, $n^{-1}S_n \xrightarrow{L^2} \overline{x}$ as $n \to \infty$, and hence also $n^{-1}S_n \xrightarrow{p} \overline{x}$.*

Proof. Our assumptions imply that $\mathbf{E}(n^{-1}S_n) = \overline{x}$, and further by Lemma 2.1.2 we have the bound $\mathsf{Var}(S_n) \le nC$. Recall the scaling property (1.3.17) of the variance, implying that

$$\mathbf{E}\Big[(n^{-1}S_n - \overline{x})^2\Big] = \mathsf{Var}\left(n^{-1}S_n\right) = \frac{1}{n^2}\mathsf{Var}(S_n) \le \frac{C}{n} \to 0$$

as $n \to \infty$. Thus, $n^{-1}S_n \xrightarrow{L^2} \overline{x}$ (recall Definition 1.3.26). By Proposition 1.3.29 this implies that also $n^{-1}S_n \xrightarrow{p} \overline{x}$. □

The most important special case of Theorem 2.1.3 is,

Example 2.1.4. *Suppose that $X_1, \ldots, X_n$ are independent and identically distributed (or in short,* i.i.d.*), with $\mathbf{E}X_1^2 < \infty$. Then, $\mathbf{E}X_i^2 = C$ and $\mathbf{E}X_i = m_X$ are both finite and independent of i. So, the L^2 weak law of large numbers tells us that $n^{-1}S_n \xrightarrow{L^2} m_X$, and hence also $n^{-1}S_n \xrightarrow{p} m_X$.*

Remark. As we shall see, the weaker condition $\mathbf{E}|X_i| < \infty$ suffices for the convergence in probability of $n^{-1}S_n$ to m_X. In Section 2.3 we show that it even suffices for the convergence almost surely of $n^{-1}S_n$ to m_X, a statement called the strong law of large numbers.

Exercise 2.1.5. *Show that the conclusion of the L^2 weak law of large numbers holds even for correlated X_i, provided $\mathbf{E}X_i = \overline{x}$ and $\mathsf{Cov}(X_i, X_j) \le r(|i - j|)$ for all i, j, and some bounded sequence $r(k) \to 0$ as $k \to \infty$.*

With an eye on generalizing the L^2 weak law of large numbers we observe that

Lemma 2.1.6. *If the random variables $Z_n \in L^2(\Omega, \mathcal{F}, \mathbf{P})$ and the non-random b_n are such that $b_n^{-2}\mathsf{Var}(Z_n) \to 0$ as $n \to \infty$, then $b_n^{-1}(Z_n - \mathbf{E}Z_n) \xrightarrow{L^2} 0$.*

PROOF. We have $\mathbf{E}[(b_n^{-1}(Z_n - \mathbf{E}Z_n))^2] = b_n^{-2}\,\mathsf{Var}(Z_n) \to 0$. □

EXAMPLE 2.1.7. *Let $Z_n = \sum_{k=1}^n X_k$ for uncorrelated random variables $\{X_k\}$. If $\mathsf{Var}(X_k)/k \to 0$ as $k \to \infty$, then Lemma 2.1.6 applies for Z_n and $b_n = n$, hence $n^{-1}(Z_n - \mathbf{E}Z_n) \to 0$ in L^2 (and in probability). Alternatively, if also $\mathsf{Var}(X_k) \to 0$, then Lemma 2.1.6 applies even for Z_n and $b_n = n^{-1/2}$.*

Many limit theorems involve random variables of the form $S_n = \sum\limits_{k=1}^{n} X_{n,k}$, that is, the row sums of triangular arrays of random variables $\{X_{n,k} : k = 1, \ldots, n\}$. Here are two such examples, both relying on Lemma 2.1.6.

EXAMPLE 2.1.8 (COUPON COLLECTOR'S PROBLEM). *Consider i.i.d. random variables $U_1, U_2, \ldots$, each distributed uniformly on $\{1, 2, \ldots, n\}$. Let $|\{U_1, \ldots, U_l\}|$ denote the number of distinct elements among the first l variables, and $\tau_k^n = \inf\{l : |\{U_1, \ldots, U_l\}| = k\}$ be the first time one has k distinct values. We are interested in the asymptotic behavior as $n \to \infty$ of $T_n = \tau_n^n$, the time it takes to have at least one representative of each of the n possible values.*

To motivate the name assigned to this example, think of collecting a set of n different coupons, where independently of all previous choices, each item is chosen at random in such a way that each of the possible n outcomes is equally likely. Then, T_n is the number of items one has to collect till having the complete set.

Setting $\tau_0^n = 0$, let $X_{n,k} = \tau_k^n - \tau_{k-1}^n$ denote the additional time it takes to get an item different from the first $k-1$ distinct items collected. Note that $X_{n,k}$ has a geometric distribution of success probability $q_{n,k} = 1 - \frac{k-1}{n}$, hence $\mathbf{E}X_{n,k} = q_{n,k}^{-1}$ and $\mathsf{Var}(X_{n,k}) \le q_{n,k}^{-2}$ (see Example 1.3.69). Since

$$T_n = \tau_n^n - \tau_0^n = \sum_{k=1}^n (\tau_k^n - \tau_{k-1}^n) = \sum_{k=1}^n X_{n,k}\,,$$

we have by linearity of the expectation that

$$\mathbf{E}T_n = \sum_{k=1}^n \Big(1 - \frac{k-1}{n}\Big)^{-1} = n \sum_{\ell=1}^n \ell^{-1} = n(\log n + \gamma_n)\,,$$

where $\gamma_n = \sum\limits_{\ell=1}^{n} \ell^{-1} - \int_1^n x^{-1} dx$ is between zero and one (by monotonicity of $x \mapsto x^{-1}$). Further, $X_{n,k}$ is independent of each earlier waiting time $X_{n,j}$, $j = 1, \ldots, k-1$, hence we have by Lemma 2.1.2 that

$$\mathsf{Var}(T_n) = \sum_{k=1}^n \mathsf{Var}(X_{n,k}) \le \sum_{k=1}^n \Big(1 - \frac{k-1}{n}\Big)^{-2} \le n^2 \sum_{\ell=1}^{\infty} \ell^{-2} = Cn^2\,,$$

for some $C < \infty$. Applying Lemma 2.1.6 with $b_n = n \log n$, we deduce that

$$\frac{T_n - n(\log n + \gamma_n)}{n \log n} \xrightarrow{L^2} 0\,.$$

Since $\gamma_n / \log n \to 0$, it follows that

$$\frac{T_n}{n \log n} \xrightarrow{L^2} 1\,,$$

and $T_n/(n \log n) \to 1$ in probability as well.

One possible extension of Example 2.1.8 concerns infinitely many possible coupons. That is,

EXERCISE 2.1.9. *Suppose $\{\xi_k\}$ are i.i.d. positive integer valued random variables, with $\mathbf{P}(\xi_1 = i) = p_i > 0$ for $i = 1, 2, \ldots$. Let $D_l = |\{\xi_1, \ldots, \xi_l\}|$ denote the number of distinct elements among the first l variables.*

(a) *Show that $D_n \overset{a.s.}{\to} \infty$ as $n \to \infty$.*

(b) *Show that $n^{-1}\mathbf{E}D_n \to 0$ as $n \to \infty$ and deduce that $n^{-1}D_n \overset{p}{\to} 0$.*

Hint: *Recall that $(1-p)^n \geq 1 - np$ for any $p \in [0,1]$ and $n \geq 0$.*

EXAMPLE 2.1.10 (AN OCCUPANCY PROBLEM). *Suppose we distribute at random r distinct balls among n distinct boxes, where each of the possible n^r assignments of balls to boxes is equally likely. We are interested in the asymptotic behavior of the number N_n of empty boxes when $r/n \to \alpha \in [0, \infty]$, while $n \to \infty$. To this end, let A_i denote the event that the i-th box is empty, so $N_n = \sum_{i=1}^{n} I_{A_i}$. Since $\mathbf{P}(A_i) = (1 - 1/n)^r$ for each i, it follows that $\mathbf{E}(n^{-1}N_n) = (1-1/n)^r \to e^{-\alpha}$. Further, $\mathbf{E}N_n^2 = \sum_{i,j=1}^{n} \mathbf{P}(A_i \cap A_j)$ and $\mathbf{P}(A_i \cap A_j) = (1 - 2/n)^r$ for each $i \neq j$. Hence, splitting the sum according to $i = j$ or $i \neq j$, we see that*

$$\mathsf{Var}(n^{-1}N_n) = \frac{1}{n^2}\mathbf{E}N_n^2 - (1-\frac{1}{n})^{2r} = \frac{1}{n}(1-\frac{1}{n})^r + (1-\frac{1}{n})(1-\frac{2}{n})^r - (1-\frac{1}{n})^{2r}.$$

As $n \to \infty$, the first term on the right side goes to zero, and with $r/n \to \alpha$, each of the other two terms converges to $e^{-2\alpha}$. Consequently, $\mathsf{Var}(n^{-1}N_n) \to 0$, so applying Lemma 2.1.6 for $b_n = n$ we deduce that

$$\frac{N_n}{n} \to e^{-\alpha}$$

in L^2 and in probability.

2.1.2. Weak laws and truncation.

Our next order of business is to extend the weak law of large numbers for row sums S_n in triangular arrays of independent $X_{n,k}$ which lack a finite second moment. Of course, with S_n no longer in L^2, there is no way to establish convergence in L^2. So, we aim to retain only the convergence in probability, using *truncation*. That is, we consider the row sums $\overline{S}_n$ for the truncated array $\overline{X}_{n,k} = X_{n,k}I_{|X_{n,k}| \leq b_n}$, with $b_n \to \infty$ slowly enough to control the variance of $\overline{S}_n$ and fast enough for $\mathbf{P}(S_n \neq \overline{S}_n) \to 0$. As we next show, this gives the convergence in probability for $\overline{S}_n$ which translates to same convergence result for S_n.

THEOREM 2.1.11 (WEAK LAW FOR TRIANGULAR ARRAYS). *Suppose that for each n, the random variables $X_{n,k}$, $k = 1, \ldots, n$ are pairwise independent. Let $\overline{X}_{n,k} = X_{n,k}I_{|X_{n,k}| \leq b_n}$ for non-random $b_n > 0$ such that as $n \to \infty$ both*

(a) $\sum_{k=1}^{n} \mathbf{P}(|X_{n,k}| > b_n) \to 0$,

and

(b) $b_n^{-2} \sum_{k=1}^{n} \mathsf{Var}(\overline{X}_{n,k}) \to 0$.

Then, $b_n^{-1}(S_n - a_n) \overset{p}{\to} 0$ as $n \to \infty$, where $S_n = \sum_{k=1}^{n} X_{n,k}$ and $a_n = \sum_{k=1}^{n} \mathbf{E}\overline{X}_{n,k}$.

PROOF. Let $\overline{S}_n = \sum_{k=1}^n \overline{X}_{n,k}$. Clearly, for any $\varepsilon > 0$,

$$\Big\{|\frac{S_n - a_n}{b_n}| > \varepsilon\Big\} \subseteq \Big\{S_n \neq \overline{S}_n\Big\} \bigcup \Big\{|\frac{\overline{S}_n - a_n}{b_n}| > \varepsilon\Big\} .$$

Consequently,

$$\mathbf{P}\Big(|\frac{S_n - a_n}{b_n}| > \varepsilon\Big) \le \mathbf{P}(S_n \neq \overline{S}_n) + \mathbf{P}\Big(|\frac{\overline{S}_n - a_n}{b_n}| > \varepsilon\Big) . \tag{2.1.2}$$

To bound the first term, note that our condition (a) implies that as $n \to \infty$,

$$\begin{aligned}\mathbf{P}(S_n \neq \overline{S}_n) &\le \mathbf{P}\Big(\bigcup_{k=1}^n \{X_{n,k} \neq \overline{X}_{n,k}\}\Big) \\ &\le \sum_{k=1}^n \mathbf{P}(X_{n,k} \neq \overline{X}_{n,k}) = \sum_{k=1}^n \mathbf{P}(|X_{n,k}| > b_n) \to 0 .\end{aligned}$$

Turning to bound the second term in (2.1.2), recall that pairwise independence is preserved under truncation, hence $\overline{X}_{n,k}$, $k = 1, \ldots, n$ are uncorrelated random variables (to convince yourself, apply (1.4.12) for the appropriate functions). Thus, an application of Lemma 2.1.2 yields that as $n \to \infty$,

$$\mathsf{Var}(b_n^{-1}\overline{S}_n) = b_n^{-2} \sum_{k=1}^n \mathsf{Var}(\overline{X}_{n,k}) \to 0 ,$$

by our condition (b). Since $a_n = \mathbf{E}\overline{S}_n$, from Chebyshev's inequality we deduce that for any fixed $\varepsilon > 0$,

$$\mathbf{P}\Big(|\frac{\overline{S}_n - a_n}{b_n}| > \varepsilon\Big) \le \varepsilon^{-2}\,\mathsf{Var}(b_n^{-1}\overline{S}_n) \to 0 ,$$

as $n \to \infty$. In view of (2.1.2), this completes the proof of the theorem. □

Specializing the weak law of Theorem 2.1.11 to a single sequence yields the following.

PROPOSITION 2.1.12 (WEAK LAW OF LARGE NUMBERS). *Consider i.i.d. random variables $\{X_i\}$, such that $x\mathbf{P}(|X_1| > x) \to 0$ as $x \to \infty$. Then, $n^{-1}S_n - \mu_n \xrightarrow{p} 0$, where $S_n = \sum\limits_{i=1}^n X_i$ and $\mu_n = \mathbf{E}[X_1 I_{\{|X_1| \le n\}}]$.*

PROOF. We get the result as an application of Theorem 2.1.11 for $X_{n,k} = X_k$ and $b_n = n$, in which case $a_n = n\mu_n$. Turning to verify condition (a) of this theorem, note that

$$\sum_{k=1}^n \mathbf{P}(|X_{n,k}| > n) = n\mathbf{P}(|X_1| > n) \to 0$$

as $n \to \infty$, by our assumption. Thus, all that remains to do is to verify that condition (b) of Theorem 2.1.11 holds here. This amounts to showing that as $n \to \infty$,

$$\Delta_n = n^{-2} \sum_{k=1}^n \mathsf{Var}(\overline{X}_{n,k}) = n^{-1}\,\mathsf{Var}(\overline{X}_{n,1}) \to 0 .$$

Recall that for any R.V. Z,

$$\mathsf{Var}(Z) = \mathbf{E}Z^2 - (\mathbf{E}Z)^2 \leq \mathbf{E}|Z|^2 = \int_0^\infty 2y\mathbf{P}(|Z| > y)dy$$

(see part (a) of Lemma 1.4.31 for the right identity). Considering $Z = \overline{X}_{n,1} = X_1 I_{\{|X_1| \leq n\}}$ for which $\mathbf{P}(|Z| > y) = \mathbf{P}(|X_1| > y) - \mathbf{P}(|X_1| > n) \leq \mathbf{P}(|X_1| > y)$ when $0 < y < n$ and $\mathbf{P}(|Z| > y) = 0$ when $y \geq n$, we deduce that

$$\Delta_n = n^{-1}\,\mathsf{Var}(Z) \leq n^{-1}\int_0^n g(y)dy\,,$$

where by our assumption, $g(y) = 2y\mathbf{P}(|X_1| > y) \to 0$ for $y \to \infty$. Further, the non-negative Borel function $g(y) \leq 2y$ is then uniformly bounded on $[0, \infty)$, hence $n^{-1}\int_0^n g(y)dy \to 0$ as $n \to \infty$ (c.f. Exercise 1.3.52). Verifying that $\Delta_n \to 0$, we established condition (b) of Theorem 2.1.11 and thus completed the proof of the proposition. □

Remark. The condition $x\mathbf{P}(|X_1| > x) \to 0$ for $x \to \infty$ is indeed necessary for the existence of non-random μ_n such that $n^{-1}S_n - \mu_n \xrightarrow{p} 0$ (c.f. [**Fel71**, Page 234-236] for a proof).

Exercise 2.1.13. *Let $\{X_i\}$ be i.i.d. with $\mathbf{P}(X_1 = (-1)^k k) = 1/(ck^2 \log k)$ for integers $k \geq 2$ and a normalization constant $c = \sum_k 1/(k^2 \log k)$. Show that $\mathbf{E}|X_1| = \infty$, but there is a non-random $\mu < \infty$ such that $n^{-1}S_n \xrightarrow{p} \mu$.*

As a corollary to Proposition 2.1.12 we next show that $n^{-1}S_n \xrightarrow{p} m_X$ as soon as the i.i.d. random variables X_i are in L^1.

Corollary 2.1.14. *Consider $S_n = \sum\limits_{k=1}^{n} X_k$ for i.i.d. random variables $\{X_i\}$ such that $\mathbf{E}|X_1| < \infty$. Then, $n^{-1}S_n \xrightarrow{p} \mathbf{E}X_1$ as $n \to \infty$.*

Proof. In view of Proposition 2.1.12, it suffices to show that if $\mathbf{E}|X_1| < \infty$, then both $n\mathbf{P}(|X_1| > n) \to 0$ and $\mathbf{E}X_1 - \mu_n = \mathbf{E}[X_1 I_{\{|X_1|>n\}}] \to 0$ as $n \to \infty$. To this end, recall that $\mathbf{E}|X_1| < \infty$ implies that $\mathbf{P}(|X_1| < \infty) = 1$ and hence the sequence $X_1 I_{\{|X_1|>n\}}$ converges to zero a.s. and is bounded by the integrable $|X_1|$. Thus, by dominated convergence $\mathbf{E}[X_1 I_{\{|X_1|>n\}}] \to 0$ as $n \to \infty$. Applying dominated convergence for the sequence $nI_{\{|X_1|>n\}}$ (which also converges a.s. to zero and is bounded by the integrable $|X_1|$), we deduce that $n\mathbf{P}(|X_1| > n) = \mathbf{E}[nI_{\{|X_1|>n\}}] \to 0$ when $n \to \infty$, thus completing the proof of the corollary. □

We conclude this section by considering an example for which $\mathbf{E}|X_1| = \infty$ and Proposition 2.1.12 does not apply, but nevertheless, Theorem 2.1.11 allows us to deduce that $c_n^{-1}S_n \xrightarrow{p} 1$ for some c_n such that $c_n/n \to \infty$.

Example 2.1.15. *Let $\{X_i\}$ be i.i.d. random variables such that $\mathbf{P}(X_1 = 2^j) = 2^{-j}$ for $j = 1, 2, \ldots$. This has the interpretation of a game, where in each of its independent rounds you win 2^j dollars if it takes exactly j tosses of a fair coin to get the first Head. This example is called the* St. Petersburg paradox*, since though $\mathbf{E}X_1 = \infty$, you clearly would not pay an infinite amount just in order to play this game. Applying Theorem 2.1.11 we find that one should be willing to pay roughly $n \log_2 n$ dollars for playing n rounds of this game, since $S_n/(n \log_2 n) \xrightarrow{p} 1$ as $n \to \infty$. Indeed, the conditions of Theorem 2.1.11 apply for $b_n = 2^{m_n}$ provided* *the integers m_n are such that $m_n - \log_2 n \to \infty$. Taking $m_n \leq \log_2 n + \log_2(\log_2 n)$ implies that $b_n \leq n \log_2 n$ and $a_n/(n \log_2 n) = m_n / \log_2 n \to 1$ as $n \to \infty$, with the consequence of $S_n/(n \log_2 n) \xrightarrow{p} 1$ (for details see* [**Dur03**, Example 1.5.7]*).*

2.2. The Borel-Cantelli lemmas

When dealing with asymptotic theory, we often wish to understand the relation between countably many events A_n in the same probability space. The two Borel-Cantelli lemmas of Subsection 2.2.1 provide information on the probability of the set of outcomes that are in infinitely many of these events, based only on $\mathbf{P}(A_n)$. There are numerous applications to these lemmas, few of which are given in Subsection 2.2.2 while many more appear in later sections of these notes.

2.2.1. Limit superior and the Borel-Cantelli lemmas.

We are often interested in the *limits superior* and *limits inferior* of a sequence of events A_n on the same measurable space $(\Omega, \mathcal{F})$.

Definition 2.2.1. *For a sequence of subsets $A_n \subseteq \Omega$, define*

$$\begin{aligned} A^\infty := \limsup A_n &= \bigcap_{m=1}^{\infty} \bigcup_{\ell=m}^{\infty} A_\ell \\ &= \{\omega : \omega \in A_n \text{ for infinitely many } n\text{'s } \} \\ &= \{\omega : \omega \in A_n \text{ infinitely often } \} = \{A_n \text{ i.o. } \} \end{aligned}$$

Similarly,

$$\begin{aligned} \liminf A_n &= \bigcup_{m=1}^{\infty} \bigcap_{\ell=m}^{\infty} A_\ell \\ &= \{\omega : \omega \in A_n \text{ for all but finitely many } n\text{'s } \} \\ &= \{\omega : \omega \in A_n \text{ eventually } \} = \{A_n \text{ ev. } \} \end{aligned}$$

Remark. Note that if $A_n \in \mathcal{F}$ are measurable, then so are $\limsup A_n$ and $\liminf A_n$. By DeMorgan's law, we have that $\{A_n \textit{ ev. }\} = \{A_n^c \textit{ i.o. }\}^c$, that is, $\omega \in A_n$ for all n large enough if and only if $\omega \in A_n^c$ for finitely many n's.

Also, if $\omega \in A_n$ eventually, then certainly $\omega \in A_n$ infinitely often, that is

$$\liminf A_n \subseteq \limsup A_n \,.$$

The notations $\limsup A_n$ and $\liminf A_n$ are due to the intimate connection of these sets to the lim sup and lim inf of the indicator functions on the sets A_n. For example,

$$\limsup_{n\to\infty} I_{A_n}(\omega) = I_{\limsup A_n}(\omega),$$

since for a given $\omega \in \Omega$, the lim sup on the left side equals 1 if and only if the sequence $n \mapsto I_{A_n}(\omega)$ contains an infinite subsequence of ones. In other words, if and only if the given ω is in infinitely many of the sets A_n. Similarly,

$$\liminf_{n\to\infty} I_{A_n}(\omega) = I_{\liminf A_n}(\omega),$$

since for a given $\omega \in \Omega$, the lim inf on the left side equals 1 if and only if there are only finitely many zeros in the sequence $n \mapsto I_{A_n}(\omega)$ (for otherwise, their limit inferior is zero). In other words, if and only if the given ω is in A_n for all n large enough.

In view of the preceding remark, Fatou's lemma yields the following relations.

EXERCISE 2.2.2. *Prove that for any sequence $A_n \in \mathcal{F}$,*

$$\mathbf{P}(\limsup A_n) \geq \limsup_{n\to\infty} \mathbf{P}(A_n) \geq \liminf_{n\to\infty} \mathbf{P}(A_n) \geq \mathbf{P}(\liminf A_n)\,.$$

Show that the right most inequality holds even when the probability measure is replaced by an arbitrary measure $\mu(\cdot)$, but the left most inequality may then fail unless $\mu(\bigcup_{k\geq n} A_k) < \infty$ for some n.

Practice your understanding of the concepts of lim sup and lim inf of sets by solving the following exercise.

EXERCISE 2.2.3. *Assume that $\mathbf{P}(\limsup A_n) = 1$ and $\mathbf{P}(\liminf B_n) = 1$. Prove that $\mathbf{P}(\limsup(A_n \cap B_n)) = 1$. What happens if the condition on $\{B_n\}$ is weakened to $\mathbf{P}(\limsup B_n) = 1$?*

Our next result, called the first Borel-Cantelli lemma, states that if the probabilities $\mathbf{P}(A_n)$ of the individual events A_n converge to zero fast enough, then almost surely, A_n occurs for only finitely many values of n, that is, $\mathbf{P}(A_n \text{ i.o.}) = 0$. This lemma is extremely useful, as the possibly complex relation between the different events A_n is irrelevant for its conclusion.

LEMMA 2.2.4 (BOREL-CANTELLI I). *Suppose $A_n \in \mathcal{F}$ and $\sum\limits_{n=1}^{\infty} \mathbf{P}(A_n) < \infty$. Then, $\mathbf{P}(A_n \text{ i.o.}) = 0$.*

PROOF. Define $N(\omega) = \sum_{k=1}^{\infty} I_{A_k}(\omega)$. By the monotone convergence theorem and our assumption,

$$\mathbf{E}[N(\omega)] = \mathbf{E}\Big[\sum_{k=1}^{\infty} I_{A_k}(\omega)\Big] = \sum_{k=1}^{\infty} \mathbf{P}(A_k) < \infty.$$

Since the expectation of N is finite, certainly $\mathbf{P}(\{\omega : N(\omega) = \infty\}) = 0$. Noting that the set $\{\omega : N(\omega) = \infty\}$ is merely $\{\omega : A_n \text{ i.o.}\}$, the conclusion $\mathbf{P}(A_n \text{ i.o.}) = 0$ of the lemma follows. □

Our next result, left for the reader to prove, relaxes somewhat the conditions of Lemma 2.2.4.

EXERCISE 2.2.5. *Suppose $A_n \in \mathcal{F}$ are such that $\sum\limits_{n=1}^{\infty} \mathbf{P}(A_n \cap A_{n+1}^c) < \infty$ and $\mathbf{P}(A_n) \to 0$. Show that then $\mathbf{P}(A_n \text{ i.o.}) = 0$.*

The first Borel-Cantelli lemma states that if the series $\sum_n \mathbf{P}(A_n)$ converges then almost every ω is in finitely many sets A_n. If $\mathbf{P}(A_n) \to 0$, but the series $\sum_n \mathbf{P}(A_n)$ diverges, then the event $\{A_n \text{ i.o.}\}$ might or might not have positive probability. In this sense, the Borel-Cantelli I is not tight, as the following example demonstrates.

EXAMPLE 2.2.6. *Consider the uniform probability measure U on $((0,1], \mathcal{B}_{(0,1]})$, and the events $A_n = (0, 1/n]$. Then $A_n \downarrow \emptyset$, so $\{A_n \text{ i.o.}\} = \emptyset$, but $U(A_n) = 1/n$, so $\sum_n U(A_n) = \infty$ and the Borel-Cantelli I does not apply.*

Recall also Example 1.3.25 showing the existence of $A_n = (t_n, t_n + 1/n]$ such that $U(A_n) = 1/n$ while $\{A_n \text{ i.o.}\} = (0,1]$. Thus, in general the probability of $\{A_n \text{ i.o.}\}$ depends on the relation between the different events A_n.

As seen in the preceding example, the divergence of the series $\sum_n \mathbf{P}(A_n)$ is not sufficient for the occurrence of a set of positive probability of ω values, each of which is in infinitely many events A_n. However, upon adding the assumption that the events A_n are mutually independent (flagrantly not the case in Example 2.2.6), we conclude that *almost all* ω must be in infinitely many of the events A_n:

LEMMA 2.2.7 (BOREL-CANTELLI II). *Suppose $A_n \in \mathcal{F}$ are mutually independent and $\sum\limits_{n=1}^{\infty} \mathbf{P}(A_n) = \infty$. Then, necessarily $\mathbf{P}(A_n \text{ i.o.}) = 1$.*

PROOF. Fix $0 < m < n < \infty$. Use the mutual independence of the events A_ℓ and the inequality $1 - x \le e^{-x}$ for $x \ge 0$, to deduce that

$$\begin{aligned}\mathbf{P}\Big(\bigcap_{\ell=m}^{n} A_\ell^c\Big) &= \prod_{\ell=m}^{n} \mathbf{P}(A_\ell^c) = \prod_{\ell=m}^{n} (1 - \mathbf{P}(A_\ell)) \\ &\le \prod_{\ell=m}^{n} e^{-\mathbf{P}(A_\ell)} = \exp(-\sum_{\ell=m}^{n} \mathbf{P}(A_\ell)) \,.\end{aligned}$$

As $n \to \infty$, the set $\bigcap\limits_{\ell=m}^{n} A_\ell^c$ shrinks. With the series in the exponent diverging, by continuity from above of the probability measure $\mathbf{P}(\cdot)$ we see that for any m,

$$\mathbf{P}\Big(\bigcap_{\ell=m}^{\infty} A_\ell^c\Big) \le \exp(-\sum_{\ell=m}^{\infty} \mathbf{P}(A_\ell)) = 0 \,.$$

Take the complement to see that $\mathbf{P}(B_m) = 1$ for $B_m = \bigcup_{\ell=m}^{\infty} A_\ell$ and all m. Since $B_m \downarrow \{A_n \text{ i.o. }\}$ when $m \uparrow \infty$, it follows by continuity from above of $\mathbf{P}(\cdot)$ that

$$\mathbf{P}(A_n \text{ i.o.}) = \lim_{m\to\infty} \mathbf{P}(B_m) = 1 \,,$$

as stated. □

As an immediate corollary of the two Borel-Cantelli lemmas, we observe yet another 0-1 law.

COROLLARY 2.2.8. *If $A_n \in \mathcal{F}$ are $\mathbf{P}$-mutually independent then $\mathbf{P}(A_n \text{ i.o.})$ is either 0 or 1. In other words, for any given sequence of mutually independent events, either almost all outcomes are in infinitely many of these events, or almost all outcomes are in finitely many of them.*

The *Kochen-Stone lemma*, left as an exercise, generalizes Borel-Cantelli II to situations lacking independence.

EXERCISE 2.2.9. *Suppose A_k are events on the same probability space such that $\sum_k \mathbf{P}(A_k) = \infty$ and*

$$\limsup_{n\to\infty} \Big(\sum_{k=1}^{n} \mathbf{P}(A_k)\Big)^2 \Big/ \Big(\sum_{1\le j,k\le n} \mathbf{P}(A_j \cap A_k)\Big) = \alpha > 0 \,.$$

Prove that then $\mathbf{P}(A_n \text{ i.o. }) \ge \alpha$.
Hint: *Consider part (a) of Exercise 1.3.21 for $Y_n = \sum_{k\le n} I_{A_k}$ and $a_n = \lambda \mathbf{E} Y_n$.*

2.2.2. Applications. In the sequel we explore various applications of the two Borel-Cantelli lemmas. In doing so, unless explicitly stated otherwise, all events and random variables are defined on the same probability space.

We know that the convergence a.s. of X_n to X_∞ implies the convergence in probability of X_n to X_∞, but not vice versa (see Exercise 1.3.23 and Example 1.3.25). As our first application of Borel-Cantelli I, we refine the relation between these two modes of convergence, showing that convergence in probability is equivalent to convergence almost surely along sub-sequences.

THEOREM 2.2.10. *$X_n \xrightarrow{p} X_\infty$ if and only if for every subsequence $m \mapsto X_{n(m)}$ there exists a further sub-subsequence $X_{n(m_k)}$ such that $X_{n(m_k)} \overset{a.s.}{\to} X_\infty$ as $k \to \infty$.*

We start the proof of this theorem with a simple analysis lemma.

LEMMA 2.2.11. *Let y_n be a sequence in a topological space. If every subsequence $y_{n(m)}$ has a further sub-subsequence $y_{n(m_k)}$ that converges to y, then $y_n \to y$.*

PROOF. If y_n does not converge to y, then there exists an open set G containing y and a subsequence $y_{n(m)}$ such that $y_{n(m)} \notin G$ for all m. But clearly, then we cannot find a further subsequence of $y_{n(m)}$ that converges to y. $\square$

REMARK. Applying Lemma 2.2.11 to $y_n = \mathbf{E}|X_n - X_\infty|$ we deduce that $X_n \xrightarrow{L^1} X_\infty$ if and only if any subsequence $n(m)$ has a further sub-subsequence $n(m_k)$ such that $X_{n(m_k)} \xrightarrow{L^1} X_\infty$ as $k \to \infty$.

PROOF OF THEOREM 2.2.10. First, we show sufficiency, assuming $X_n \xrightarrow{p} X_\infty$. Fix a subsequence $n(m)$ and $\varepsilon_k \downarrow 0$. By the definition of convergence in probability, there exists a sub-subsequence $n(m_k) \uparrow \infty$ such that $\mathbf{P}\left(|X_{n(m_k)} - X_\infty| > \varepsilon_k\right) \leq 2^{-k}$. Call this sequence of events $A_k = \left\{\omega : |X_{n(m_k)}(\omega) - X_\infty(\omega)| > \varepsilon_k\right\}$. Then the series $\sum_k \mathbf{P}(A_k)$ converges. Therefore, by Borel-Cantelli I, $\mathbf{P}(\limsup A_k) = 0$. For any $\omega \notin \limsup A_k$ there are only finitely many values of k such that $|X_{n(m_k)} - X_\infty| > \varepsilon_k$, or alternatively, $|X_{n(m_k)} - X_\infty| \leq \varepsilon_k$ for all k large enough. Since $\varepsilon_k \downarrow 0$, it follows that $X_{n(m_k)}(\omega) \to X_\infty(\omega)$ when $\omega \notin \limsup A_k$, that is, with probability one.

Conversely, fix $\delta > 0$. Let $y_n = \mathbf{P}(|X_n - X_\infty| > \delta)$. By assumption, for every subsequence $n(m)$ there exists a further subsequence $n(m_k)$ so that $X_{n(m_k)}$ converges to X_∞ almost surely, hence in probability, and in particular, $y_{n(m_k)} \to 0$. Applying Lemma 2.2.11 we deduce that $y_n \to 0$, and since $\delta > 0$ is arbitrary it follows that $X_n \xrightarrow{p} X_\infty$. $\square$

It is not hard to check that convergence almost surely is invariant under application of an a.s. continuous mapping.

EXERCISE 2.2.12. *Let $g : \mathbb{R} \mapsto \mathbb{R}$ be a Borel function and denote by $\mathbf{D}_g$ its set of discontinuities. Show that if $X_n \overset{a.s.}{\to} X_\infty$ finite valued, and $\mathbf{P}(X_\infty \in \mathbf{D}_g) = 0$, then $g(X_n) \overset{a.s.}{\to} g(X_\infty)$ as well (recall Exercise 1.2.28 that $\mathbf{D}_g \in \mathcal{B}$). This applies for a continuous function g in which case $\mathbf{D}_g = \emptyset$.*

A direct consequence of Theorem 2.2.10 is that convergence in probability is also preserved under an a.s. *continuous mapping* (and if the mapping is also bounded, we even get L^1 convergence).

Corollary 2.2.13. *Suppose $X_n \xrightarrow{p} X_\infty$, g is a Borel function and $\mathbf{P}(X_\infty \in \mathbf{D}_g) = 0$. Then, $g(X_n) \xrightarrow{p} g(X_\infty)$. If in addition g is bounded, then $g(X_n) \xrightarrow{L^1} g(X_\infty)$ (and $\mathbf{E}g(X_n) \to \mathbf{E}g(X_\infty)$).*

Proof. Fix a subsequence $X_{n(m)}$. By Theorem 2.2.10 there exists a subsequence $X_{n(m_k)}$ such that $\mathbf{P}(A) = 1$ for $A = \{\omega : X_{n(m_k)}(\omega) \to X_\infty(\omega) \text{ as } k \to \infty\}$. Let $B = \{\omega : X_\infty(\omega) \notin \mathbf{D}_g\}$, noting that by assumption $\mathbf{P}(B) = 1$. For any $\omega \in A \cap B$ we have $g(X_{n(m_k)}(\omega)) \to g(X_\infty(\omega))$ by the continuity of g outside $\mathbf{D}_g$. Therefore, $g(X_{n(m_k)}) \xrightarrow{a.s.} g(X_\infty)$. Now apply Theorem 2.2.10 in the reverse direction: For any subsequence, we have just constructed a further subsequence with convergence a.s., hence $g(X_n) \xrightarrow{p} g(X_\infty)$.

Finally, if g is bounded, then the collection $\{g(X_n)\}$ is U.I. yielding, by Vitali's convergence theorem, its convergence in L^1 (and hence that $\mathbf{E}g(X_n) \to \mathbf{E}g(X_\infty)$). □

You are next to extend the scope of Theorem 2.2.10 and the continuous mapping of Corollary 2.2.13 to random variables taking values in a separable metric space.

Exercise 2.2.14. *Recall the definition of convergence in probability in a separable metric space $(\mathbb{S}, \rho)$ as in Remark 1.3.24.*

(a) *Extend the proof of Theorem 2.2.10 to apply for any $(\mathbb{S}, \mathcal{B}_{\mathbb{S}})$-valued random variables $\{X_n, n \le \infty\}$ (and in particular for $\overline{\mathbb{R}}$-valued variables).*

(b) *Denote by $\mathbf{D}_g$ the set of discontinuities of a Borel measurable $g : \mathbb{S} \mapsto \overline{\mathbb{R}}$ (defined similarly to Exercise 1.2.28, where real-valued functions are considered). Suppose $X_n \xrightarrow{p} X_\infty$ and $\mathbf{P}(X_\infty \in \mathbf{D}_g) = 0$. Show that then $g(X_n) \xrightarrow{p} g(X_\infty)$ and if in addition g is bounded, then also $g(X_n) \xrightarrow{L^1} g(X_\infty)$.*

The following result in analysis is obtained by combining the continuous mapping of Corollary 2.2.13 with the weak law of large numbers.

Exercise 2.2.15 (Inverting Laplace transforms). *The* Laplace transform *of a bounded, continuous function $h(x)$ on $[0, \infty)$ is the function $L_h(s) = \int_0^\infty e^{-sx} h(x)dx$ on $(0, \infty)$.*

(a) *Show that for any $s > 0$ and positive integer k,*

$$(-1)^{k-1} \frac{s^k L_h^{(k-1)}(s)}{(k-1)!} = \int_0^\infty e^{-sx} \frac{s^k x^{k-1}}{(k-1)!} h(x)dx = \mathbf{E}[h(W_k)]\,,$$

where $L_h^{(k-1)}(\cdot)$ denotes the $(k-1)$-th derivative of the function $L_h(\cdot)$ and W_k has the gamma density *with parameters k and s.*

(b) *Recall Exercise 1.4.46 that for $s = n/y$ the law of W_n coincides with the law of $n^{-1}\sum_{i=1}^n T_i$ where $T_i \ge 0$ are i.i.d. random variables, each having the exponential distribution of parameter $1/y$ (with $\mathbf{E}T_1 = y$ and finite moments of all order, c.f. Example 1.3.68). Deduce that the inversion formula*

$$h(y) = \lim_{n\to\infty} (-1)^{n-1} \frac{(n/y)^n}{(n-1)!} L_h^{(n-1)}(n/y)\,,$$

holds for any $y > 0$.

The next application of Borel-Cantelli I provides our first strong law of large numbers.

Proposition 2.2.16. *Suppose* $\mathbf{E}[Z_n^2] \le C$ *for some* $C < \infty$ *and all* n*. Then,* $n^{-1}Z_n \overset{a.s.}{\to} 0$ *as* $n \to \infty$*.*

Proof. Fixing $\delta > 0$ let $A_k = \{\omega : |k^{-1}Z_k(\omega)| > \delta\}$ for $k = 1, 2, \ldots$. Then, by Chebyshev's inequality and our assumption,

$$\mathbf{P}(A_k) = \mathbf{P}(\{\omega : |Z_k(\omega)| \ge k\delta\}) \le \frac{\mathbf{E}(Z_k^2)}{(k\delta)^2} \le \frac{C}{\delta^2}k^{-2}\,.$$

Since $\sum_k k^{-2} < \infty$, it follows by Borel Cantelli I that $\mathbf{P}(A^\infty) = 0$, where $A^\infty = \{\omega : |k^{-1}Z_k(\omega)| > \delta$ for infinitely many values of $k\}$. Hence, for any fixed $\delta > 0$, with probability one $k^{-1}|Z_k(\omega)| \le \delta$ for all large enough k, that is, $\limsup_{n\to\infty} n^{-1}|Z_n(\omega)| \le \delta$ a.s. Considering a sequence $\delta_m \downarrow 0$ we conclude that $n^{-1}Z_n \to 0$ for $n \to \infty$ and a.e. ω. □

Exercise 2.2.17. *Let* $S_n = \sum\limits_{l=1}^{n} X_l$*, where* $\{X_i\}$ *are i.i.d. random variables with* $\mathbf{E}X_1 = 0$ *and* $\mathbf{E}X_1^4 < \infty$*.*

(a) *Show that* $n^{-1}S_n \overset{a.s.}{\to} 0$*.*
Hint: *Verify that Proposition 2.2.16 applies for* $Z_n = n^{-1}S_n^2$*.*

(b) *Show that* $n^{-1}D_n \overset{a.s.}{\to} 0$ *where* D_n *denotes the number of distinct integers among* $\{\xi_k, k \le n\}$ *and* $\{\xi_k\}$ *are i.i.d. integer valued random variables.*
Hint: $D_n \le 2M + \sum_{k=1}^n I_{|\xi_k|\ge M}$.

In contrast, here is an example where the empirical averages of integrable, zero mean independent variables do not converge to zero. Of course, the trick is to have non-identical distributions, with the bulk of the probability drifting to negative one.

Exercise 2.2.18. *Suppose* X_i *are mutually independent random variables such that* $\mathbf{P}(X_n = n^2 - 1) = 1 - \mathbf{P}(X_n = -1) = n^{-2}$ *for* $n = 1, 2, \ldots$. *Show that* $\mathbf{E}X_n = 0$*, for all* n*, while* $n^{-1}\sum_{i=1}^n X_i \overset{a.s.}{\to} -1$ *for* $n \to \infty$*.*

Next we have few other applications of Borel-Cantelli I, starting with some additional properties of convergence a.s.

Exercise 2.2.19. *Show that for any R.V.* X_n

(a) $X_n \overset{a.s.}{\to} 0$ *if and only if* $\mathbf{P}(|X_n| > \varepsilon \text{ i.o. }) = 0$ *for each* $\varepsilon > 0$*.*

(b) *There exist non-random constants* $b_n \uparrow \infty$ *such that* $X_n/b_n \overset{a.s.}{\to} 0$*.*

Exercise 2.2.20. *Show that if* $W_n > 0$ *and* $\mathbf{E}W_n \le 1$ *for every* n*, then almost surely,*

$$\limsup_{n\to\infty} n^{-1}\log W_n \le 0\,.$$

Our next example demonstrates how Borel-Cantelli I is typically applied in the study of the asymptotic growth of running maxima of random variables.

Example 2.2.21 (Head runs). *Let* $\{X_k, k \in \mathbf{Z}\}$ *be a two-sided sequence of i.i.d.* $\{0,1\}$*-valued random variables, with* $\mathbf{P}(X_1 = 1) = \mathbf{P}(X_1 = 0) = 1/2$*. With* $\ell_m = \max\{i : X_{m-i+1} = \cdots = X_m = 1\}$ *denoting the length of the run of* 1*'s going* *backwards from time m, we are interested in the asymptotics of the longest such run during $1, 2, \dots, n$, that is,*

$$\begin{aligned} L_n &= \max\{\ell_m : m = 1, \dots, n\} \\ &= \max\{m - k : X_{k+1} = \cdots = X_m = 1 \text{ for some } m = 1, \dots, n\} \, . \end{aligned}$$

Noting that $\ell_m + 1$ has a geometric distribution of success probability $p = 1/2$, we deduce by an application of Borel-Cantelli I that for each $\varepsilon > 0$, with probability one, $\ell_n \le (1+\varepsilon)\log_2 n$ for all n large enough. Hence, on the same set of probability one, we have $N = N(\omega)$ finite such that $L_n \le \max(L_N, (1+\varepsilon)\log_2 n)$ for all $n \ge N$. Dividing by $\log_2 n$ and considering $n \to \infty$ followed by $\varepsilon_k \downarrow 0$, this implies that

$$\limsup_{n\to\infty} \frac{L_n}{\log_2 n} \overset{a.s.}{\le} 1 \, .$$

For each fixed $\varepsilon > 0$ let $A_n = \{L_n < k_n\}$ for $k_n = [(1-\varepsilon)\log_2 n]$. Noting that

$$A_n \subseteq \bigcap_{i=1}^{m_n} B_i^c \, ,$$

for $m_n = [n/k_n]$ and the independent events $B_i = \{X_{(i-1)k_n+1} = \cdots = X_{ik_n} = 1\}$, yields a bound of the form $\mathbf{P}(A_n) \le \exp(-n^\varepsilon/(2\log_2 n))$ for all n large enough (c.f. [**Dur03**, Example 1.6.3] *for details). Since $\sum_n \mathbf{P}(A_n) < \infty$, we have that*

$$\liminf_{n\to\infty} \frac{L_n}{\log_2 n} \overset{a.s.}{\ge} 1$$

by yet another application of Borel-Cantelli I, followed by $\varepsilon_k \downarrow 0$. We thus conclude that

$$\frac{L_n}{\log_2 n} \overset{a.s.}{\to} 1 \, .$$

The next exercise combines both Borel-Cantelli lemmas to provide the 0-1 law for another problem about head runs.

Exercise 2.2.22. *Let $\{X_k\}$ be a sequence of i.i.d. $\{0,1\}$-valued random variables, with $\mathbf{P}(X_1 = 1) = p$ and $\mathbf{P}(X_1 = 0) = 1 - p$. Let A_k be the event that $X_m = \cdots = X_{m+k-1} = 1$ for some $2^k \le m \le 2^{k+1} - k$. Show that $\mathbf{P}(A_k \text{ i.o. }) = 1$ if $p \ge 1/2$ and $\mathbf{P}(A_k \text{ i.o. }) = 0$ if $p < 1/2$.*
Hint: *When $p \ge 1/2$ consider only $m = 2^k + (i-1)k$ for $i = 0, \dots, [2^k/k]$.*

Here are a few direct applications of the second Borel-Cantelli lemma.

Exercise 2.2.23. *Suppose that $\{Z_k\}$ are i.i.d. random variables such that $\mathbf{P}(Z_1 = z) < 1$ for any $z \in \mathbb{R}$.*

(a) *Show that $\mathbf{P}(Z_k$ converges for $k \to \infty) = 0$.*
(b) *Determine the values of* $\limsup_{n\to\infty}(Z_n/\log n)$ *and* $\liminf_{n\to\infty}(Z_n/\log n)$ *in case Z_k has the* exponential distribution *(of parameter $\lambda = 1$).*

After deriving the classical bounds on the tail of the normal distribution, you use both Borel-Cantelli lemmas in bounding the fluctuations of the sums of i.i.d. standard normal variables.

Exercise 2.2.24. *Let $\{G_i\}$ be i.i.d. standard normal random variables.*

(a) *Show that for any $x > 0$,*

$$(x^{-1} - x^{-3})e^{-x^2/2} \leq \int_x^\infty e^{-y^2/2}dy \leq x^{-1}e^{-x^2/2}\,.$$

Many texts prove these estimates, for example see [**Dur03**, Theorem 1.1.4].

(b) *Show that, with probability one,*

$$\limsup_{n\to\infty} \frac{G_n}{\sqrt{2\log n}} = 1\,.$$

(c) *Let $S_n = G_1 + \cdots + G_n$. Recall that $n^{-1/2}S_n$ has the standard normal distribution. Show that*

$$\mathbf{P}(|S_n| < 2\sqrt{n\log n},\ \textit{ev.}\ ) = 1\,.$$

Remark. Ignoring the dependence between the elements of the sequence S_k, the bound in part (c) of the preceding exercise is not tight. The definite result here is the *law of the iterated logarithm* (in short LIL), which states that when the i.i.d. summands are of zero mean and variance one,

$$\mathbf{P}(\limsup_{n\to\infty} \frac{S_n}{\sqrt{2n\log\log n}} = 1) = 1\,. \tag{2.2.1}$$

We defer the derivation of (2.2.1) to Theorem 9.2.28, building on a similar LIL for the Brownian motion (but, see [**Bil95**, Theorem 9.5] for a direct proof of (2.2.1), using both Borel-Cantelli lemmas).

The next exercise relates explicit integrability conditions for i.i.d. random variables to the asymptotics of their running maxima.

Exercise 2.2.25. *Consider possibly $\overline{\mathbb{R}}$-valued, i.i.d. random variables $\{Y_i\}$ and their running maxima $M_n = \max_{k\leq n} Y_k$.*

(a) *Using (2.3.4) if needed, show that $\mathbf{P}(|Y_n| > n\ \ i.o.\ ) = 0$ if and only if $\mathbf{E}[|Y_1|] < \infty$.*

(b) *Show that $n^{-1}Y_n \overset{a.s.}{\to} 0$ if and only if $\mathbf{E}[|Y_1|] < \infty$.*

(c) *Show that $n^{-1}M_n \overset{a.s.}{\to} 0$ if and only if $\mathbf{E}[(Y_1)_+] < \infty$ and $\mathbf{P}(Y_1 > -\infty) > 0$.*

(d) *Show that $n^{-1}M_n \overset{p}{\to} 0$ if and only if $n\mathbf{P}(Y_1 > n) \to 0$ and $\mathbf{P}(Y_1 > -\infty) > 0$.*

(e) *Show that $n^{-1}Y_n \overset{p}{\to} 0$ if and only if $\mathbf{P}(|Y_1| < \infty) = 1$.*

In the following exercise, you combine Borel Cantelli I and the variance computation of Lemma 2.1.2 to improve upon Borel Cantelli II.

Exercise 2.2.26. *Suppose $\sum_{n=1}^\infty \mathbf{P}(A_n) = \infty$ for pairwise independent events $\{A_i\}$. Let $S_n = \sum_{i=1}^n I_{A_i}$ be the number of events occurring among the first n.*

(a) *Prove that $\mathsf{Var}(S_n) \leq \mathbf{E}(S_n)$ and deduce from it that $S_n/\mathbf{E}(S_n) \overset{p}{\to} 1$.*

(b) *Applying Borel-Cantelli I show that $S_{n_k}/\mathbf{E}(S_{n_k}) \overset{a.s.}{\to} 1$ as $k \to \infty$, where $n_k = \inf\{n : \mathbf{E}(S_n) \geq k^2\}$.*

(c) *Show that $\mathbf{E}(S_{n_{k+1}})/\mathbf{E}(S_{n_k}) \to 1$ and since $n \mapsto S_n$ is non-decreasing, deduce that $S_n/\mathbf{E}(S_n) \overset{a.s.}{\to} 1$.*

Remark. Borel-Cantelli II is the a.s. convergence $S_n \to \infty$ for $n \to \infty$, which is a consequence of part (c) of the preceding exercise (since $\mathbf{E}S_n \to \infty$).

We conclude this section with an example in which the asymptotic rate of growth of random variables of interest is obtained by an application of Exercise 2.2.26.

Example 2.2.27 (Record values). *Let $\{X_i\}$ be a sequence of i.i.d. random variables with a continuous distribution function $F_X(x)$. The event $A_k = \{X_k > X_j, j = 1, \ldots, k-1\}$ represents the occurrence of a record at the k instance (for example, think of X_k as an athlete's kth distance jump). We are interested in the asymptotics of the count $R_n = \sum\limits_{i=1}^{n} I_{A_i}$ of record events during the first n instances. Because of the continuity of F_X we know that a.s. the values of X_i, $i = 1, 2, \ldots$ are distinct. Further, rearranging the random variables $X_1, X_2, \ldots, X_n$ in a decreasing order induces a random permutation π_n on $\{1, 2, \ldots, n\}$, where all $n!$ possible permutations are equally likely. From this it follows that $\mathbf{P}(A_k) = \mathbf{P}(\pi_k(k) = 1) = 1/k$, and though definitely not obvious at first sight, the events A_k are mutually independent (see* [**Dur03**, Example 1.6.2] *for details). So, $\mathbf{E}R_n = \log n + \gamma_n$ where γ_n is between zero and one, and from Exercise 2.2.26 we deduce that $(\log n)^{-1} R_n \overset{a.s.}{\to} 1$ as $n \to \infty$. Note that this result is independent of the law of X, as long as the distribution function F_X is continuous.*

2.3. Strong law of large numbers

In Corollary 2.1.14 we got the classical weak law of large numbers, namely, the convergence in probability of the empirical averages $n^{-1}\sum_{i=1}^{n} X_i$ of i.i.d. integrable random variables X_i to the mean $\mathbf{E}X_1$. Assuming in addition that $\mathbf{E}X_1^4 < \infty$, you used Borel-Cantelli I in Exercise 2.2.17 en-route to the corresponding strong law of large numbers, that is, replacing the convergence in probability with the stronger notion of convergence almost surely.

We provide here two approaches to the strong law of large numbers, both of which get rid of the unnecessary finite moment assumptions. Subsection 2.3.1 follows Etemadi's (1981) direct proof of this result via the subsequence method. Subsection 2.3.2 deals in a more systematic way with the convergence of random series, yielding the strong law of large numbers as one of its consequences.

2.3.1. The subsequence method.

Etemadi's key observation is that it essentially suffices to consider non-negative X_i, for which upon proving the a.s. convergence along a not too sparse subsequence n_l, the interpolation to the whole sequence can be done by the monotonicity of $n \mapsto \sum^n X_i$. This is an example of a general approach to a.s. convergence, called the *subsequence method*, which you have already encountered in Exercise 2.2.26.

We thus start with the strong law for integrable, non-negative variables.

Proposition 2.3.1. *Let $S_n = \sum_{i=1}^{n} X_i$ for non-negative, pairwise independent and identically distributed, integrable random variables $\{X_i\}$. Then, $n^{-1}S_n \overset{a.s.}{\to} \mathbf{E}X_1$ as $n \to \infty$.*

Proof. The proof progresses along the themes of Section 2.1, starting with the truncation $\overline{X}_k = X_k I_{|X_k| \le k}$ and its corresponding sums $\overline{S}_n = \sum_{i=1}^{n} \overline{X}_i$.

Since $\{X_i\}$ are identically distributed and $x \mapsto \mathbf{P}(|X_1| > x)$ is non-increasing, we have that

$$\sum_{k=1}^{\infty} \mathbf{P}(X_k \neq \overline{X}_k) = \sum_{k=1}^{\infty} \mathbf{P}(|X_1| > k) \leq \int_0^{\infty} \mathbf{P}(|X_1| > x) dx = \mathbf{E}|X_1| < \infty$$

(see part (a) of Lemma 1.4.31 for the rightmost identity and recall our assumption that X_1 is integrable). Thus, by Borel-Cantelli I, with probability one, $X_k(\omega) = \overline{X}_k(\omega)$ for all but finitely many k's, in which case necessarily $\sup_n |S_n(\omega) - \overline{S}_n(\omega)|$ is finite. This shows that $n^{-1}(S_n - \overline{S}_n) \overset{a.s.}{\to} 0$, whereby it suffices to prove that $n^{-1}\overline{S}_n \overset{a.s.}{\to} \mathbf{E}X_1$.

To this end, we next show that it suffices to prove the following lemma about almost sure convergence of $\overline{S}_n$ along suitably chosen subsequences.

Lemma 2.3.2. *Fixing $\alpha > 1$ let $n_l = [\alpha^l]$. Under the conditions of the proposition, $n_l^{-1}(\overline{S}_{n_l} - \mathbf{E}\overline{S}_{n_l}) \overset{a.s.}{\to} 0$ as $l \to \infty$.*

By dominated convergence, $\mathbf{E}[X_1 I_{|X_1| \leq k}] \to \mathbf{E}X_1$ as $k \to \infty$, and consequently, as $n \to \infty$,

$$\frac{1}{n}\mathbf{E}\overline{S}_n = \frac{1}{n}\sum_{k=1}^{n} \mathbf{E}\overline{X}_k = \frac{1}{n}\sum_{k=1}^{n} \mathbf{E}[X_1 I_{|X_1|\leq k}] \to \mathbf{E}X_1$$

(we have used here the consistency of Cesáro averages, c.f. Exercise 1.3.52 for an integral version). Thus, assuming that Lemma 2.3.2 holds, we have that $n_l^{-1}\overline{S}_{n_l} \overset{a.s.}{\to} \mathbf{E}X_1$ when $l \to \infty$, for each $\alpha > 1$.

We complete the proof of the proposition by interpolating from the subsequences $n_l = [\alpha^l]$ to the whole sequence. To this end, fix $\alpha > 1$. Since $n \mapsto \overline{S}_n$ is non-decreasing, we have for all $\omega \in \Omega$ and any $n \in [n_l, n_{l+1}]$,

$$\frac{n_l}{n_{l+1}} \frac{\overline{S}_{n_l}(\omega)}{n_l} \leq \frac{\overline{S}_n(\omega)}{n} \leq \frac{n_{l+1}}{n_l} \frac{\overline{S}_{n_{l+1}}(\omega)}{n_{l+1}}$$

With $n_l/n_{l+1} \to 1/\alpha$ for $l \to \infty$, the a.s. convergence of $m^{-1}\overline{S}_m$ along the subsequence $m = n_l$ implies that the event

$$A_\alpha := \{\omega : \frac{1}{\alpha}\mathbf{E}X_1 \leq \liminf_{n\to\infty} \frac{\overline{S}_n(\omega)}{n} \leq \limsup_{n\to\infty} \frac{\overline{S}_n(\omega)}{n} \leq \alpha\mathbf{E}X_1\},$$

has probability one. Consequently, taking $\alpha_m \downarrow 1$, we deduce that the event $B := \bigcap_m A_{\alpha_m}$ also has probability one, and further, $n^{-1}\overline{S}_n(\omega) \to \mathbf{E}X_1$ for each $\omega \in B$. We thus deduce that $n^{-1}\overline{S}_n \overset{a.s.}{\to} \mathbf{E}X_1$, as needed to complete the proof of the proposition. □

Remark. The monotonicity of certain random variables (here $n \mapsto \overline{S}_n$), is crucial to the successful application of the subsequence method. The subsequence n_l for which we need a direct proof of convergence is completely determined by the scaling function b_n^{-1} applied to this monotone sequence (here $b_n = n$); we need $b_{n_{l+1}}/b_{n_l} \to \alpha$, which should be arbitrarily close to 1. For example, same subsequences $n_l = [\alpha^l]$ are to be used whenever b_n is roughly of a polynomial growth in n, while even $n_l = (l!)^c$ would work in case $b_n = \log n$.

Likewise, the truncation level is determined by the highest moment of the basic variables which is assumed to be finite. For example, we can take $\overline{X}_k = X_k I_{|X_k| \leq k^p}$ for any $p > 0$ such that $\mathbf{E}|X_1|^{1/p} < \infty$.

Proof of Lemma 2.3.2. Note that $\mathbf{E}[\overline{X}_k^2]$ is non-decreasing in k. Further, $\overline{X}_k$ are pairwise independent, hence uncorrelated, so by Lemma 2.1.2,

$$\mathsf{Var}(\overline{S}_n) = \sum_{k=1}^{n} \mathsf{Var}(\overline{X}_k) \leq \sum_{k=1}^{n} \mathbf{E}[\overline{X}_k^2] \leq n\mathbf{E}[\overline{X}_n^2] = n\mathbf{E}[X_1^2 I_{|X_1|\leq n}] \,.$$

Combining this with Chebychev's inequality yield the bound

$$\mathbf{P}(|\overline{S}_n - \mathbf{E}\overline{S}_n| \geq \varepsilon n) \leq (\varepsilon n)^{-2} \mathsf{Var}(\overline{S}_n) \leq \varepsilon^{-2} n^{-1} \mathbf{E}[X_1^2 I_{|X_1|\leq n}] \,,$$

for any $\varepsilon > 0$. Applying Borel-Cantelli I for the events $A_l = \{|\overline{S}_{n_l} - \mathbf{E}\overline{S}_{n_l}| \geq \varepsilon n_l\}$, followed by $\varepsilon_m \downarrow 0$, we get the a.s. convergence to zero of $n^{-1}|\overline{S}_n - \mathbf{E}\overline{S}_n|$ along any subsequence n_l for which

$$\sum_{l=1}^{\infty} n_l^{-1} \mathbf{E}[X_1^2 I_{|X_1|\leq n_l}] = \mathbf{E}[X_1^2 \sum_{l=1}^{\infty} n_l^{-1} I_{|X_1|\leq n_l}] < \infty$$

(the latter identity is a special case of Exercise 1.3.40). Since $\mathbf{E}|X_1| < \infty$, it thus suffices to show that for $n_l = [\alpha^l]$ and any $x > 0$,

$$u(x) := \sum_{l=1}^{\infty} n_l^{-1} I_{x\leq n_l} \leq c x^{-1} \,, \tag{2.3.1}$$

where $c = 2\alpha/(\alpha - 1) < \infty$. To establish (2.3.1) fix $\alpha > 1$ and $x > 0$, setting $L = \min\{l \geq 1 : n_l \geq x\}$. Then, $\alpha^L \geq x$, and since $[y] \geq y/2$ for all $y \geq 1$,

$$u(x) = \sum_{l=L}^{\infty} n_l^{-1} \leq 2\sum_{l=L}^{\infty} \alpha^{-l} = c\alpha^{-L} \leq c x^{-1} \,.$$

So, we have established (2.3.1) and hence completed the proof of the lemma. □

As already promised, it is not hard to extend the scope of the strong law of large numbers beyond integrable and non-negative random variables.

Theorem 2.3.3 (Strong law of large numbers). *Let $S_n = \sum_{i=1}^n X_i$ for pairwise independent and identically distributed random variables $\{X_i\}$, such that either $\mathbf{E}[(X_1)_+]$ is finite or $\mathbf{E}[(X_1)_-]$ is finite. Then, $n^{-1}S_n \overset{a.s.}{\to} \mathbf{E}X_1$ as $n \to \infty$.*

Proof. First consider non-negative X_i. The case of $\mathbf{E}X_1 < \infty$ has already been dealt with in Proposition 2.3.1. In case $\mathbf{E}X_1 = \infty$, consider $S_n^{(m)} = \sum_{i=1}^n X_i^{(m)}$ for the bounded, non-negative, pairwise independent and identically distributed random variables $X_i^{(m)} = \min(X_i, m) \leq X_i$. Since Proposition 2.3.1 applies for $\{X_i^{(m)}\}$, it follows that a.s. for any fixed $m < \infty$,

$$\liminf_{n\to\infty} n^{-1}S_n \geq \liminf_{n\to\infty} n^{-1}S_n^{(m)} = \mathbf{E}X_1^{(m)} = \mathbf{E}\min(X_1, m) \,. \tag{2.3.2}$$

Taking $m \uparrow \infty$, by monotone convergence $\mathbf{E}\min(X_1, m) \uparrow \mathbf{E}X_1 = \infty$, so (2.3.2) results with $n^{-1}S_n \to \infty$ a.s.

Turning to the general case, we have the decomposition $X_i = (X_i)_+ - (X_i)_-$ of each random variable to its positive and negative parts, with

$$n^{-1}S_n = n^{-1}\sum_{i=1}^{n}(X_i)_+ - n^{-1}\sum_{i=1}^{n}(X_i)_- \tag{2.3.3}$$

Since $(X_i)_+$ are non-negative, pairwise independent and identically distributed, it follows that $n^{-1}\sum_{i=1}^n (X_i)_+ \overset{a.s.}{\to} \mathbf{E}[(X_1)_+]$ as $n \to \infty$. For the same reason, also $n^{-1}\sum_{i=1}^n (X_i)_- \overset{a.s.}{\to} \mathbf{E}[(X_1)_-]$. Our assumption that either $\mathbf{E}[(X_1)_+] < \infty$ or $\mathbf{E}[(X_1)_-] < \infty$ implies that $\mathbf{E}X_1 = \mathbf{E}[(X_1)_+] - \mathbf{E}[(X_1)_-]$ is well defined, and in view of (2.3.3) we have the stated a.s. convergence of $n^{-1}S_n$ to $\mathbf{E}X_1$. □

Exercise 2.3.4. *You are to prove now a converse to the strong law of large numbers (for a more general result, due to Feller (1946), see* [**Dur03**, Theorem 1.8.9]*).*

(a) *Let Y denote the integer part of a random variable $Z \geq 0$. Show that $Y = \sum_{n=1}^\infty I_{\{Z\geq n\}}$, and deduce that*

$$\sum_{n=1}^\infty \mathbf{P}(Z \geq n) \leq \mathbf{E}Z \leq 1 + \sum_{n=1}^\infty \mathbf{P}(Z \geq n)\,. \tag{2.3.4}$$

(b) *Suppose $\{X_i\}$ are i.i.d R.V.s with $\mathbf{E}[|X_1|^\alpha] = \infty$ for some $\alpha > 0$. Show that for any $k > 0$,*

$$\sum_{n=1}^\infty \mathbf{P}(|X_n| > kn^{1/\alpha}) = \infty\,,$$

and deduce that a.s. $\limsup_{n\to\infty} n^{-1/\alpha}|X_n| = \infty$.

(c) *Conclude that if $S_n = X_1 + X_2 + \cdots + X_n$, then*

$$\limsup_{n\to\infty} n^{-1/\alpha}|S_n| = \infty, \qquad \text{a.s.}$$

We provide next two classical applications of the strong law of large numbers, the first of which deals with the large sample asymptotics of the empirical distribution function.

Example 2.3.5 (Empirical distribution function). *Let*

$$F_n(x) = n^{-1}\sum_{i=1}^n I_{(-\infty,x]}(X_i)\,,$$

denote the observed fraction of values among the first n variables of the sequence $\{X_i\}$ which do not exceed x. The functions $F_n(\cdot)$ are thus called the empirical distribution functions of this sequence.

For i.i.d. $\{X_i\}$ with distribution function F_X our next result improves the strong law of large numbers by showing that F_n converges uniformly to F_X as $n \to \infty$.

Theorem 2.3.6 (Glivenko-Cantelli). *For i.i.d. $\{X_i\}$ with arbitrary distribution function F_X, as $n \to \infty$,*

$$D_n = \sup_{x\in\mathbb{R}} |F_n(x) - F_X(x)| \overset{a.s.}{\to} 0\,.$$

Remark. While outside our scope, we note in passing the Dvoretzky-Kiefer-Wolfowitz inequality that $\mathbf{P}(D_n > \varepsilon) \leq 2\exp(-2n\varepsilon^2)$ for any n and all $\varepsilon > 0$, quantifying the rate of convergence of D_n to zero (see [**DKW56**], or [**Mas90**] for the optimal pre-exponential constant).

Proof. By the right continuity of both $x \mapsto F_n(x)$ and $x \mapsto F_X(x)$ (c.f. Theorem 1.2.36), the value of D_n is unchanged when the supremum over $x \in \mathbb{R}$ is replaced by the one over $x \in \mathbb{Q}$ (the rational numbers). In particular, this shows that each D_n is a random variable (c.f. Theorem 1.2.22).

Applying the strong law of large numbers for the i.i.d. non-negative $I_{(-\infty,x]}(X_i)$ whose expectation is $F_X(x)$, we deduce that $F_n(x) \overset{a.s.}{\to} F_X(x)$ for each fixed non-random $x \in \mathbb{R}$. Similarly, considering the strong law of large numbers for the i.i.d. non-negative $I_{(-\infty,x)}(X_i)$ whose expectation is $F_X(x^-)$, we have that $F_n(x^-) \overset{a.s.}{\to} F_X(x^-)$ for each fixed non-random $x \in \mathbb{R}$. Consequently, for any fixed $l < \infty$ and $x_{1,l}, \dots, x_{l,l}$ we have that

$$D_{n,l} = \max(\max_{k=1}^{l} |F_n(x_{k,l}) - F_X(x_{k,l})|, \max_{k=1}^{l} |F_n(x_{k,l}^-) - F_X(x_{k,l}^-)|) \overset{a.s.}{\to} 0 \,,$$

as $n \to \infty$. Choosing $x_{k,l} = \inf\{x : F_X(x) \geq k/(l+1)\}$, we get out of the monotonicity of $x \mapsto F_n(x)$ and $x \mapsto F_X(x)$ that $D_n \leq D_{n,l} + l^{-1}$ (c.f. [**Bil95**, Proof of Theorem 20.6] or [**Dur03**, Proof of Theorem 1.7.4]). Therefore, taking $n \to \infty$ followed by $l \to \infty$ completes the proof of the theorem. □

We turn to our second example, which is about counting processes.

Example 2.3.7 (Renewal theory). *Let $\{\tau_i\}$ be i.i.d. positive, finite random variables and $T_n = \sum_{k=1}^n \tau_k$. Here T_n is interpreted as the time of the n-th occurrence of a given event, with τ_k representing the length of the time interval between the $(k-1)$ occurrence and that of the k-th such occurrence. Associated with T_n is the dual process $N_t = \sup\{n : T_n \leq t\}$ counting the number of occurrences during the time interval $[0, t]$. In the next exercise you are to derive the strong law for the large t asymptotics of $t^{-1}N_t$.*

Exercise 2.3.8. *Consider the setting of Example 2.3.7.*

(a) *By the strong law of large numbers argue that $n^{-1}T_n \overset{a.s.}{\to} \mathbf{E}\tau_1$. Then, adopting the convention $\frac{1}{\infty} = 0$, deduce that $t^{-1}N_t \overset{a.s.}{\to} 1/\mathbf{E}\tau_1$ for $t \to \infty$.*
Hint: *From the definition of N_t it follows that $T_{N_t} \leq t < T_{N_t+1}$ for all $t \geq 0$.*

(b) *Show that $t^{-1}N_t \overset{a.s.}{\to} 1/\mathbf{E}\tau_2$ as $t \to \infty$, even if the law of τ_1 is different from that of the i.i.d. $\{\tau_i, i \geq 2\}$.*

Here is a strengthening of the preceding result to convergence in L^1.

Exercise 2.3.9. *In the context of Example 2.3.7 fix $\delta > 0$ such that $\mathbf{P}(\tau_1 > \delta) > \delta$ and let $\widetilde{T}_n = \sum_{k=1}^n \widetilde{\tau}_k$ for the i.i.d. random variables $\widetilde{\tau}_i = \delta I_{\{\tau_i > \delta\}}$. Note that $\widetilde{T}_n \leq T_n$ and consequently $N_t \leq \widetilde{N}_t = \sup\{n : \widetilde{T}_n \leq t\}$.*

(a) *Show that* $\limsup_{t\to\infty} t^{-2}\mathbf{E}\widetilde{N}_t^2 < \infty$.

(b) *Deduce that $\{t^{-1}N_t : t \geq 1\}$ is uniformly integrable (see Exercise 1.3.54), and conclude that $t^{-1}\mathbf{E}N_t \to 1/\mathbf{E}\tau_1$ when $t \to \infty$.*

The next exercise deals with an elaboration over Example 2.3.7.

Exercise 2.3.10. *For $i = 1, 2, \dots$ the ith light bulb burns for an amount of time τ_i and then remains burned out for time s_i before being replaced by the $(i+1)$th bulb. Let R_t denote the fraction of time during $[0, t]$ in which we have a working light. Assuming that the two sequences $\{\tau_i\}$ and $\{s_i\}$ are independent, each consisting of i.i.d. positive and integrable random variables, show that $R_t \overset{a.s.}{\to} \mathbf{E}\tau_1/(\mathbf{E}\tau_1 + \mathbf{E}s_1)$.*

Here is another exercise, dealing with sampling "at times of heads" in independent fair coin tosses, from a non-random bounded sequence of weights $v(l)$, the averages of which converge.

Exercise 2.3.11. *For a sequence $\{B_i\}$ of i.i.d. Bernoulli random variables of parameter $p = 1/2$, let T_n be the time that the corresponding partial sums reach level n. That is, $T_n = \inf\{k : \sum_{i=1}^k B_i \geq n\}$, for $n = 1, 2, \ldots$.*

(a) *Show that $n^{-1} T_n \overset{a.s.}{\to} 2$ as $n \to \infty$.*
(b) *Given non-negative, non-random $\{v(k)\}$ show that $k^{-1} \sum_{i=1}^k v(T_i) \overset{a.s.}{\to} s$ as $k \to \infty$, for some non-random s, if and only if $n^{-1} \sum_{l=1}^n v(l) B_l \overset{a.s.}{\to} s/2$ as $n \to \infty$.*
(c) *Deduce that if $n^{-1} \sum_{l=1}^n v(l)^2$ is bounded and $n^{-1} \sum_{l=1}^n v(l) \to s$ as $n \to \infty$, then $k^{-1} \sum_{i=1}^k v(T_i) \overset{a.s.}{\to} s$ as $k \to \infty$.*

Hint: *For part (c) consider first the limit of $n^{-1} \sum_{l=1}^n v(l)(B_l - 0.5)$ as $n \to \infty$.*

We conclude this subsection with few additional applications of the strong law of large numbers, first to a problem of *universal hypothesis testing*, then an application involving stochastic geometry, and finally one motivated by investment science.

Exercise 2.3.12. *Consider i.i.d. $[0,1]$-valued random variables $\{X_k\}$.*

(a) *Find Borel measurable functions $f_n : [0,1]^n \mapsto \{0,1\}$, which are independent of the law of X_k, such that $f_n(X_1, X_2, \ldots, X_n) \overset{a.s.}{\to} 0$ whenever $\mathbf{E}X_1 < 1/2$ and $f_n(X_1, X_2, \ldots, X_n) \overset{a.s.}{\to} 1$ whenever $\mathbf{E}X_1 > 1/2$.*
(b) *Modify your answer to assure that $f_n(X_1, X_2, \ldots, X_n) \overset{a.s.}{\to} 1$ also in case $\mathbf{E}X_1 = 1/2$.*

Exercise 2.3.13. *Let $\{U_n\}$ be i.i.d. random vectors, each uniformly distributed on the unit ball $\{u \in \mathbb{R}^2 : |u| \leq 1\}$. Consider the $\mathbb{R}^2$-valued random vectors $X_n = |X_{n-1}| U_n$, $n = 1, 2, \ldots$ starting at a non-random, non-zero vector X_0 (that is, each point is uniformly chosen in a ball centered at the origin and whose radius is the distance from the origin to the previously chosen point). Show that $n^{-1} \log |X_n| \overset{a.s.}{\to} -1/2$ as $n \to \infty$.*

Exercise 2.3.14. *Let $\{V_n\}$ be i.i.d. non-negative random variables. Fixing $r > 0$ and $q \in (0,1]$, consider the sequence $W_0 = 1$ and $W_n = (qr + (1-q)V_n) W_{n-1}$, $n = 1, 2, \ldots$. A motivating example is of W_n recording the relative growth of a portfolio where a constant fraction q of one's wealth is re-invested each year in a risk-less asset that grows by r per year, with the remainder re-invested in a risky asset whose annual growth factors are the random V_n.*

(a) *Show that $n^{-1} \log W_n \overset{a.s.}{\to} w(q)$, for $w(q) = \mathbf{E} \log(qr + (1-q)V_1)$.*
(b) *Show that $q \mapsto w(q)$ is concave on $(0,1]$.*
(c) *Using Jensen's inequality show that $w(q) \leq w(1)$ in case $\mathbf{E}V_1 \leq r$. Further, show that if $\mathbf{E}V_1^{-1} \leq r^{-1}$, then the almost sure convergence applies also for $q = 0$ and that $w(q) \leq w(0)$.*
(d) *Assuming that $\mathbf{E}V_1^2 < \infty$ and $\mathbf{E}V_1^{-2} < \infty$ show that $\sup\{w(q) : q \in [0,1]\}$ is finite, and further that the maximum of $w(q)$ is obtained at some $q^* \in (0,1)$ when $\mathbf{E}V_1 > r > 1/\mathbf{E}V_1^{-1}$. Interpret your results in terms of the preceding investment example.*

Hint: *Consider small $q > 0$ and small $1 - q > 0$ and recall that $\log(1+x) \geq x - x^2/2$ for any $x \geq 0$.*

2.3.2. Convergence of random series. A second approach to the strong law of large numbers is based on studying the convergence of random series. The key tool in this approach is Kolmogorov's maximal inequality, which we prove next.

Proposition 2.3.15 (Kolmogorov's maximal inequality). *The random variables $Y_1, \ldots, Y_n$ are mutually independent, with $\mathbf{E}Y_l^2 < \infty$ and $\mathbf{E}Y_l = 0$ for $l = 1, \ldots, n$. Then, for $Z_k = Y_1 + \cdots + Y_k$ and any $z > 0$,*

$$z^2 \mathbf{P}(\max_{1 \le k \le n} |Z_k| \ge z) \le \mathsf{Var}(Z_n) \,. \tag{2.3.5}$$

Remark. Chebyshev's inequality gives only $z^2 \mathbf{P}(|Z_n| \ge z) \le \mathsf{Var}(Z_n)$ which is significantly weaker and insufficient for our current goals.

Proof. Fixing $z > 0$ we decompose the event $A = \{\max_{1 \le k \le n} |Z_k| \ge z\}$ according to the minimal index k for which $|Z_k| \ge z$. That is, A is the union of the disjoint events $A_k = \{|Z_k| \ge z > |Z_j|, \; j = 1, \ldots, k-1\}$ over $1 \le k \le n$. Obviously,

$$z^2 \mathbf{P}(A) = \sum_{k=1}^n z^2 \mathbf{P}(A_k) \le \sum_{k=1}^n \mathbf{E}[Z_k^2; A_k] \,, \tag{2.3.6}$$

since $Z_k^2 \ge z^2$ on A_k. Further, $\mathbf{E}Z_n = 0$ and A_k are disjoint, so

$$\mathsf{Var}(Z_n) = \mathbf{E}Z_n^2 \ge \sum_{k=1}^n \mathbf{E}[Z_n^2; A_k] \,. \tag{2.3.7}$$

It suffices to show that $\mathbf{E}[(Z_n - Z_k)Z_k; A_k] = 0$ for any $1 \le k \le n$, since then

$$\begin{aligned} \mathbf{E}[Z_n^2; A_k] - \mathbf{E}[Z_k^2; A_k] &= \mathbf{E}[(Z_n - Z_k)^2; A_k] + 2\mathbf{E}[(Z_n - Z_k)Z_k; A_k] \\ &= \mathbf{E}[(Z_n - Z_k)^2; A_k] \ge 0 \,, \end{aligned}$$

and (2.3.5) follows by comparing (2.3.6) and (2.3.7). Since $Z_k I_{A_k}$ can be represented as a non-random Borel function of $(Y_1, \ldots, Y_k)$, it follows that $Z_k I_{A_k}$ is measurable on $\sigma(Y_1, \ldots, Y_k)$. Consequently, for fixed k and $l > k$ the variables Y_l and $Z_k I_{A_k}$ are independent, hence uncorrelated. Further $\mathbf{E}Y_l = 0$, so

$$\mathbf{E}[(Z_n - Z_k)Z_k; A_k] = \sum_{l=k+1}^n \mathbf{E}[Y_l Z_k I_{A_k}] = \sum_{l=k+1}^n \mathbf{E}(Y_l)\mathbf{E}(Z_k I_{A_k}) = 0 \,,$$

completing the proof of Kolmogorov's inequality. □

Equipped with Kolmogorov's inequality, we provide an easy to check sufficient condition for the convergence of random series of independent R.V.

Theorem 2.3.16. *Suppose $\{X_i\}$ are independent random variables with $\mathsf{Var}(X_i) < \infty$ and $\mathbf{E}X_i = 0$. If $\sum_n \mathsf{Var}(X_n) < \infty$ then w.p.1. the random series $\sum_n X_n(\omega)$ converges (that is, the sequence $S_n(\omega) = \sum_{k=1}^n X_k(\omega)$ has a finite limit $S_\infty(\omega)$).*

Proof. Applying Kolmogorov's maximal inequality for the independent variables $Y_l = X_{l+r}$, we have that for any $\varepsilon > 0$ and positive integers r and n,

$$\mathbf{P}(\max_{r \le k \le r+n} |S_k - S_r| \ge \varepsilon) \le \varepsilon^{-2} \mathsf{Var}(S_{r+n} - S_r) = \varepsilon^{-2} \sum_{l=r+1}^{r+n} \mathsf{Var}(X_l) \,.$$

Taking $n \to \infty$, we get by continuity from below of $\mathbf{P}$ that

$$\mathbf{P}(\sup_{k \ge r} |S_k - S_r| \ge \varepsilon) \le \varepsilon^{-2} \sum_{l=r+1}^{\infty} \mathsf{Var}(X_l)$$

By our assumption that $\sum_n \mathsf{Var}(X_n)$ is finite, it follows that $\sum_{l>r} \mathsf{Var}(X_l) \to 0$ as $r \to \infty$. Hence, if we let $T_r = \sup_{n,m \geq r} |S_n - S_m|$, then for any $\varepsilon > 0$,

$$\mathbf{P}(T_r \geq 2\varepsilon) \leq \mathbf{P}(\sup_{k \geq r} |S_k - S_r| \geq \varepsilon) \to 0$$

as $r \to \infty$. Further, $r \mapsto T_r(\omega)$ is non-increasing, hence,

$$\mathbf{P}(\limsup_{M\to\infty} T_M \geq 2\varepsilon) = \mathbf{P}(\inf_M T_M \geq 2\varepsilon) \leq \mathbf{P}(T_r \geq 2\varepsilon) \to 0 \,.$$

That is, $T_M(\omega) \overset{a.s.}{\to} 0$ for $M \to \infty$. By definition, the convergence to zero of $T_M(\omega)$ is the statement that $S_n(\omega)$ is a Cauchy sequence. Since every Cauchy sequence in $\mathbb{R}$ converges to a finite limit, we have the stated a.s. convergence of $S_n(\omega)$. □

We next provide some applications of Theorem 2.3.16.

Example 2.3.17. *Considering non-random a_n such that $\sum_n a_n^2 < \infty$ and independent Bernoulli variables B_n of parameter $p = 1/2$, Theorem 2.3.16 tells us that $\sum_n (-1)^{B_n} a_n$ converges with probability one. That is, when the signs in $\sum_n \pm a_n$ are chosen on the toss of a fair coin, the series almost always converges (though quite possibly $\sum_n |a_n| = \infty$).*

Exercise 2.3.18. *Consider the* record events *A_k of Example 2.2.27.*

(a) *Verify that the events A_k are mutually independent with $\mathbf{P}(A_k) = 1/k$.*
(b) *Show that the random series $\sum_{n\geq 2}(I_{A_n} - 1/n)/\log n$ converges almost surely and deduce that $(\log n)^{-1} R_n \overset{a.s.}{\to} 1$ as $n \to \infty$.*
(c) *Provide a counterexample to the preceding in case the distribution function $F_X(x)$ is not continuous.*

The link between convergence of random series and the strong law of large numbers is the following classical analysis lemma.

Lemma 2.3.19 (Kronecker's lemma). *Consider two sequences of real numbers $\{x_n\}$ and $\{b_n\}$ where $b_n > 0$ and $b_n \uparrow \infty$. If $\sum_n x_n/b_n$ converges, then $s_n/b_n \to 0$ for $s_n = x_1 + \cdots + x_n$.*

Proof. Let $u_n = \sum_{k=1}^n (x_k/b_k)$ which by assumption converges to a finite limit denoted u_∞. Setting $u_0 = b_0 = 0$, "summation by parts" yields the identity,

$$s_n = \sum_{k=1}^n b_k(u_k - u_{k-1}) = b_n u_n - \sum_{k=1}^n (b_k - b_{k-1}) u_{k-1} \,.$$

Since $u_n \to u_\infty$ and $b_n \uparrow \infty$, the Cesáro averages $b_n^{-1} \sum_{k=1}^n (b_k - b_{k-1}) u_{k-1}$ also converge to u_∞. Consequently, $s_n/b_n \to u_\infty - u_\infty = 0$. □

Theorem 2.3.16 provides an alternative proof for the strong law of large numbers of Theorem 2.3.3 in case $\{X_i\}$ are i.i.d. (that is, replacing pairwise independence by mutual independence). Indeed, applying the same truncation scheme as in the proof of Proposition 2.3.1, it suffices to prove the following alternative to Lemma 2.3.2.

Lemma 2.3.20. *For integrable i.i.d. random variables $\{X_k\}$, let $\overline{S}_m = \sum_{k=1}^m \overline{X}_k$ and $\overline{X}_k = X_k I_{|X_k| \leq k}$. Then, $n^{-1}(\overline{S}_n - \mathbf{E}\overline{S}_n) \overset{a.s.}{\to} 0$ as $n \to \infty$.*

Lemma 2.3.20, in contrast to Lemma 2.3.2, does not require the restriction to a subsequence n_l. Consequently, in this proof of the strong law there is no need for an interpolation argument so it is carried directly for X_k, with no need to split each variable to its positive and negative parts.

PROOF OF LEMMA 2.3.20. We will shortly show that

$$\sum_{k=1}^{\infty} k^{-2} \operatorname{\mathsf{Var}}(\overline{X}_k) \leq 2\mathbf{E}|X_1| \,. \tag{2.3.8}$$

With X_1 integrable, applying Theorem 2.3.16 for the independent variables $Y_k = k^{-1}(\overline{X}_k - \mathbf{E}\overline{X}_k)$ this implies that for some A with $\mathbf{P}(A) = 1$, the random series $\sum_n Y_n(\omega)$ converges for all $\omega \in A$. Using Kronecker's lemma for $b_n = n$ and $x_n = \overline{X}_n(\omega) - \mathbf{E}\overline{X}_n$ we get that $n^{-1}\sum_{k=1}^n (\overline{X}_k - \mathbf{E}\overline{X}_k) \to 0$ as $n \to \infty$, for every $\omega \in A$, as stated.

The proof of (2.3.8) is similar to the computation employed in the proof of Lemma 2.3.2. That is, $\operatorname{\mathsf{Var}}(\overline{X}_k) \leq \mathbf{E}\overline{X}_k^2 = \mathbf{E}X_1^2 I_{|X_1|\leq k}$ and $k^{-2} \leq 2/(k(k+1))$, yielding that

$$\sum_{k=1}^{\infty} k^{-2} \operatorname{\mathsf{Var}}(\overline{X}_k) \leq \sum_{k=1}^{\infty} \frac{2}{k(k+1)} \mathbf{E}X_1^2 I_{|X_1|\leq k} = \mathbf{E}X_1^2 v(|X_1|) \,,$$

where for any $x > 0$,

$$v(x) = 2\sum_{k=\lceil x\rceil}^{\infty} \frac{1}{k(k+1)} = 2\sum_{k=\lceil x\rceil}^{\infty} \Big[\frac{1}{k} - \frac{1}{k+1}\Big] = \frac{2}{\lceil x\rceil} \leq 2x^{-1} \,.$$

Consequently, $\mathbf{E}X_1^2 v(|X_1|) \leq 2\mathbf{E}|X_1|$, and (2.3.8) follows. □

Many of the ingredients of this proof of the strong law of large numbers are also relevant for solving the following exercise.

EXERCISE 2.3.21. *Let c_n be a bounded sequence of non-random constants, and $\{X_i\}$ i.i.d. integrable R.V.-s of zero mean. Show that $n^{-1}\sum_{k=1}^n c_k X_k \overset{a.s.}{\to} 0$ for $n \to \infty$.*

Next you find few exercises that illustrate how useful Kronecker's lemma is when proving the strong law of large numbers in case of independent but not identically distributed summands.

EXERCISE 2.3.22. *Let $S_n = \sum_{k=1}^n Y_k$ for independent random variables $\{Y_i\}$ such that $\operatorname{\mathsf{Var}}(Y_k) < B < \infty$ and $\mathbf{E}Y_k = 0$ for all k. Show that $[n(\log n)^{1+\epsilon}]^{-1/2} S_n \overset{a.s.}{\to} 0$ as $n \to \infty$ and $\epsilon > 0$ is fixed (this falls short of the law of the iterated logarithm of (2.2.1), but each Y_k is allowed here to have a different distribution).*

EXERCISE 2.3.23. *Suppose the independent random variables $\{X_i\}$ are such that $\operatorname{\mathsf{Var}}(X_k) \leq p_k < \infty$ and $\mathbf{E}X_k = 0$ for $k = 1, 2, \ldots$.*

(a) *Show that if $\sum_k p_k < \infty$ then $n^{-1}\sum_{k=1}^n kX_k \overset{a.s.}{\to} 0$.*

(b) *Conversely, assuming $\sum_k p_k = \infty$, give an example of independent random variables $\{X_i\}$, such that $\operatorname{\mathsf{Var}}(X_k) \leq p_k$, $\mathbf{E}X_k = 0$, for which almost surely $\limsup_n X_n(\omega) = 1$.*

(c) *Show that the example you just gave is such that with probability one, the sequence $n^{-1}\sum_{k=1}^n kX_k(\omega)$ does not converge to a finite limit.*

EXERCISE 2.3.24. *Consider independent, non-negative random variables X_n.*

(a) *Show that if*

$$\sum_{n=1}^{\infty}[\mathbf{P}(X_n \geq 1)+\mathbf{E}(X_n I_{X_n<1})] < \infty \qquad (2.3.9)$$

then the random series $\sum_n X_n(\omega)$ *converges w.p.1.*

(b) *Prove the converse, namely, that if* $\sum_n X_n(\omega)$ *converges w.p.1. then (2.3.9) holds.*

(c) *Suppose* G_n *are mutually independent random variables, with* G_n *having the normal distribution* $\mathcal{N}(\mu_n, v_n)$*. Show that w.p.1. the random series* $\sum_n G_n^2(\omega)$ *converges if and only if* $e = \sum_n(\mu_n^2 + v_n)$ *is finite.*

(d) *Suppose* τ_n *are mutually independent random variables, with* τ_n *having the exponential distribution of parameter* $\lambda_n > 0$*. Show that w.p.1. the random series* $\sum_n \tau_n(\omega)$ *converges if and only if* $\sum_n 1/\lambda_n$ *is finite.*

Hint: *For part (b) recall that for any* $a_n \in [0,1)$*, the series* $\sum_n a_n$ *is finite if and only if* $\prod_n(1-a_n) > 0$*. For part (c) let* $f(y) = \sum_n \min((\mu_n + \sqrt{v_n}y)^2, 1)$ *and observe that if* $e = \infty$ *then* $f(y)+f(-y) = \infty$ *for all* $y \neq 0$*.*

You can now also show that for such strong law of large numbers (that is, with independent but not identically distributed summands), it suffices to strengthen the corresponding weak law (only) along the subsequence $n_r = 2^r$.

EXERCISE 2.3.25. *Let* $Z_k = \sum_{j=1}^{k} Y_j$ *where* Y_j *are mutually independent R.V.-s.*

(a) *Fixing* $\varepsilon > 0$ *show that if* $2^{-r}Z_{2^r} \overset{a.s.}{\to} 0$ *then* $\sum_r \mathbf{P}(|Z_{2^{r+1}} - Z_{2^r}| > 2^r\varepsilon)$ *is finite and if* $m^{-1}Z_m \overset{p}{\to} 0$ *then* $\max_{m<k\leq 2m}\mathbf{P}(|Z_{2m} - Z_k| \geq \varepsilon m) \to 0$.

(b) *Adapting the proof of Kolmogorov's maximal inequality show that for any* n *and* $z > 0$*,*

$$\mathbf{P}(\max_{1\leq k\leq n}|Z_k| \geq 2z)\min_{1\leq k\leq n}\mathbf{P}(|Z_n - Z_k| < z) \leq \mathbf{P}(|Z_n| > z)\,.$$

(c) *Deduce that if both* $m^{-1}Z_m \overset{p}{\to} 0$ *and* $2^{-r}Z_{2^r} \overset{a.s.}{\to} 0$ *then also* $n^{-1}Z_n \overset{a.s.}{\to} 0$.

Hint: *For part (c) combine parts (a) and (b) with* $z = n\varepsilon$*,* $n = 2^r$ *and the mutually independent* Y_{j+n}*,* $1 \leq j \leq n$*, to show that* $\sum_r \mathbf{P}(2^{-r}D_r \geq 2\varepsilon)$ *is finite for* $D_r = \max_{2^r<k\leq 2^{r+1}}|Z_k - Z_{2^r}|$ *and any fixed* $\varepsilon > 0$*.*

CHAPTER 3

Weak convergence, CLT and Poisson approximation

After dealing in Chapter 2 with examples in which random variables converge to non-random constants, we focus here on the more general theory of weak convergence, that is situations in which the laws of random variables converge to a limiting law, typically of a non-constant random variable. To motivate this theory, we start with Section 3.1 where we derive the celebrated Central Limit Theorem (in short CLT), the most widely used example of weak convergence. This is followed by the exposition of the theory, to which Section 3.2 is devoted. Section 3.3 is about the key tool of characteristic functions and their role in establishing convergence results such as the CLT. This tool is used in Section 3.4 to derive the Poisson approximation and provide an introduction to the Poisson process. In Section 3.5 we generalize the characteristic function to the setting of random vectors and study their properties while deriving the multivariate CLT.

3.1. The Central Limit Theorem

We start this section with the property of the normal distribution that makes it the likely limit for properly scaled sums of independent random variables. This is followed by a bare-hands proof of the CLT for triangular arrays in Subsection 3.1.1. We then present in Subsection 3.1.2 some of the many examples and applications of the CLT.

Recall the *normal distribution* of mean $\mu \in \mathbb{R}$ and variance $v > 0$, denoted hereafter $\mathcal{N}(\mu, v)$, the density of which is

$$f(y) = \frac{1}{\sqrt{2\pi v}} \exp(-\frac{(y-\mu)^2}{2v}) \,. \tag{3.1.1}$$

As we show next, the normal distribution is preserved when the sum of independent variables is considered (which is the main reason for its role as the limiting law for the CLT).

LEMMA 3.1.1. *Let $Y_{n,k}$ be mutually independent random variables, each having the normal distribution $\mathcal{N}(\mu_{n,k}, v_{n,k})$. Then, $G_n = \sum_{k=1}^n Y_{n,k}$ has the normal distribution $\mathcal{N}(\mu_n, v_n)$, with $\mu_n = \sum_{k=1}^n \mu_{n,k}$ and $v_n = \sum_{k=1}^n v_{n,k}$.*

PROOF. Recall that Y has a $\mathcal{N}(\mu, v)$ distribution if and only if $Y - \mu$ has the $\mathcal{N}(0, v)$ distribution. Therefore, we may and shall assume without loss of generality that $\mu_{n,k} = 0$ for all k and n. Further, it suffices to prove the lemma for $n = 2$, as the general case immediately follows by an induction argument. With $n = 2$ fixed, we simplify our notations by omitting it everywhere. Next recall the formula of Corollary 1.4.33 for the probability density function of $G = Y_1 + Y_2$, which for Y_i

of $\mathcal{N}(0,v_i)$ distribution, $i=1,2$, is

$$f_G(z)=\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi v_1}}\exp(-\frac{(z-y)^2}{2v_1})\frac{1}{\sqrt{2\pi v_2}}\exp(-\frac{y^2}{2v_2})dy\,.$$

Comparing this with the formula of (3.1.1) for $v=v_1+v_2$, it just remains to show that for any $z\in\mathbb{R}$,

$$1=\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi u}}\exp(\frac{z^2}{2v}-\frac{(z-y)^2}{2v_1}-\frac{y^2}{2v_2})dy\,, \tag{3.1.2}$$

where $u=v_1v_2/(v_1+v_2)$. It is not hard to check that the argument of the exponential function in (3.1.2) is $-(y-cz)^2/(2u)$ for $c=v_2/(v_1+v_2)$. Consequently, (3.1.2) is merely the obvious fact that the $\mathcal{N}(cz,u)$ density function integrates to one (as any density function should), no matter what the value of z is. □

Considering Lemma 3.1.1 for $Y_{n,k}=(nv)^{-1/2}(Y_k-\mu)$ and i.i.d. random variables Y_k, each having a normal distribution of mean μ and variance v, we see that $\mu_{n,k}=0$ and $v_{n,k}=1/n$, so $G_n=(nv)^{-1/2}(\sum_{k=1}^n Y_k-n\mu)$ has the standard $\mathcal{N}(0,1)$ distribution, regardless of n.

3.1.1. Lindeberg's CLT for triangular arrays.

Our next proposition, the celebrated CLT, states that the distribution of $\widehat{S}_n=(nv)^{-1/2}(\sum_{k=1}^n X_k-n\mu)$ approaches the standard normal distribution in the limit $n\to\infty$, even though X_k may well be non-normal random variables.

PROPOSITION 3.1.2 (CENTRAL LIMIT THEOREM). *Let*

$$\widehat{S}_n=\frac{1}{\sqrt{nv}}(\sum_{k=1}^n X_k-n\mu)\,,$$

where $\{X_k\}$ are i.i.d with $v=\mathsf{Var}(X_1)\in(0,\infty)$ and $\mu=\mathbf{E}(X_1)$. Then,

$$\lim_{n\to\infty}\mathbf{P}(\widehat{S}_n\le b)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{b}\exp(-\frac{y^2}{2})dy\qquad\text{for every }\ b\in\mathbb{R}\,. \tag{3.1.3}$$

As we have seen in the context of the weak law of large numbers, it pays to extend the scope of consideration to triangular arrays in which the random variables $X_{n,k}$ are independent within each row, but not necessarily of identical distribution. This is the context of Lindeberg's CLT, which we state next.

THEOREM 3.1.3 (LINDEBERG'S CLT). *Let $\widehat{S}_n=\sum_{k=1}^n X_{n,k}$ for $\mathbf{P}$-mutually independent random variables $X_{n,k}$, $k=1,\ldots,n$, such that $\mathbf{E}X_{n,k}=0$ for all k and*

$$v_n=\sum_{k=1}^n\mathbf{E}X_{n,k}^2\to 1\qquad\text{as } n\to\infty\,.$$

Then, the conclusion (3.1.3) applies if for each $\varepsilon>0$,

$$g_n(\varepsilon)=\sum_{k=1}^n\mathbf{E}[X_{n,k}^2;|X_{n,k}|\ge\varepsilon]\to 0\quad\text{as } n\to\infty\,. \tag{3.1.4}$$

Note that the variables in different rows need not be independent of each other and could even be defined on different probability spaces.

Remark 3.1.4. Under the assumptions of Proposition 3.1.2 the variables $X_{n,k} = (nv)^{-1/2}(X_k - \mu)$ are mutually independent and such that

$$\mathbf{E}X_{n,k} = (nv)^{-1/2}(\mathbf{E}X_k - \mu) = 0, \qquad v_n = \sum_{k=1}^n \mathbf{E}X_{n,k}^2 = \frac{1}{nv}\sum_{k=1}^n \mathsf{Var}(X_k) = 1\,.$$

Further, per fixed n these $X_{n,k}$ are identically distributed, so

$$g_n(\varepsilon) = n\mathbf{E}[X_{n,1}^2\,;\,|X_{n,1}| \geq \varepsilon] = v^{-1}\mathbf{E}[(X_1-\mu)^2 I_{|X_1-\mu|\geq\sqrt{nv}\varepsilon}]\,.$$

For each $\varepsilon > 0$ the sequence $(X_1-\mu)^2\mathbf{I}_{|X_1-\mu|\geq\sqrt{nv}\varepsilon}$ converges a.s. to zero for $n \to \infty$ and is dominated by the integrable random variable $(X_1-\mu)^2$. Thus, by dominated convergence, $g_n(\varepsilon) \to 0$ as $n \to \infty$. We conclude that all assumptions of Theorem 3.1.3 are satisfied for this choice of $X_{n,k}$, hence Proposition 3.1.2 is a special instance of Lindeberg's CLT, to which we turn our attention next.

Let $r_n = \max\{\sqrt{v_{n,k}} : k = 1, \ldots, n\}$ for $v_{n,k} = \mathbf{E}X_{n,k}^2$. Since for every n, k and $\varepsilon > 0$,

$$v_{n,k} = \mathbf{E}X_{n,k}^2 = \mathbf{E}[X_{n,k}^2; |X_{n,k}| < \varepsilon] + \mathbf{E}[X_{n,k}^2; |X_{n,k}| \geq \varepsilon] \leq \varepsilon^2 + g_n(\varepsilon)\,,$$

it follows that

$$r_n^2 \leq \varepsilon^2 + g_n(\varepsilon) \qquad \forall n, \varepsilon > 0\,,$$

hence Lindeberg's condition (3.1.4) implies that $r_n \to 0$ as $n \to \infty$.

Remark. Lindeberg proved Theorem 3.1.3, introducing the condition (3.1.4). Later, Feller proved that (3.1.3) plus $r_n \to 0$ implies that Lindeberg's condition holds. Together, these two results are known as the Feller-Lindeberg Theorem.

We see that the variables $X_{n,k}$ are of uniformly small variance for large n. So, considering independent random variables $Y_{n,k}$ that are also independent of the $X_{n,k}$ and such that each $Y_{n,k}$ has a $\mathcal{N}(0, v_{n,k})$ distribution, for a smooth function $h(\cdot)$ one may control $|\mathbf{E}h(\widehat{S}_n) - \mathbf{E}h(G_n)|$ by a Taylor expansion upon successively replacing the $X_{n,k}$ by $Y_{n,k}$. This indeed is the outline of Lindeberg's proof, whose core is the following lemma.

Lemma 3.1.5. *For $h : \mathbb{R} \mapsto \mathbb{R}$ of continuous and uniformly bounded second and third derivatives, G_n having the $\mathcal{N}(0, v_n)$ law, every n and $\varepsilon > 0$, we have that*

$$|\mathbf{E}h(\widehat{S}_n) - \mathbf{E}h(G_n)| \leq \Big(\frac{\varepsilon}{6} + \frac{r_n}{2}\Big)\, v_n\|h'''\|_\infty + g_n(\varepsilon)\|h''\|_\infty\,,$$

with $\|f\|_\infty = \sup_{x\in\mathbb{R}} |f(x)|$ denoting the supremum norm.

Remark. Recall that $G_n \overset{\mathcal{D}}{=} \sigma_n G$ for $\sigma_n = \sqrt{v_n}$. So, assuming $v_n \to 1$ and Lindeberg's condition which implies that $r_n \to 0$ for $n \to \infty$, it follows from the lemma that $|\mathbf{E}h(\widehat{S}_n) - \mathbf{E}h(\sigma_n G)| \to 0$ as $n \to \infty$. Further, $|h(\sigma_n x) - h(x)| \leq |\sigma_n - 1||x|\|h'\|_\infty$, so taking the expectation with respect to the standard normal law we see that $|\mathbf{E}h(\sigma_n G) - \mathbf{E}h(G)| \to 0$ if the first derivative of h is also uniformly bounded. Hence,

$$\lim_{n\to\infty} \mathbf{E}h(\widehat{S}_n) = \mathbf{E}h(G)\,, \tag{3.1.5}$$

for any continuous function $h(\cdot)$ of continuous and uniformly bounded first three derivatives. This is actually all we need from Lemma 3.1.5 in order to prove Lindeberg's CLT. Further, as we show in Section 3.2, convergence in distribution as in (3.1.3) is *equivalent* to (3.1.5) holding for all continuous, bounded functions $h(\cdot)$.

PROOF OF LEMMA 3.1.5. Let $G_n = \sum_{k=1}^n Y_{n,k}$ for mutually independent $Y_{n,k}$, distributed according to $\mathcal{N}(0, v_{n,k})$, that are independent of $\{X_{n,k}\}$. Fixing n and h, we simplify the notations by eliminating n, that is, we write Y_k for $Y_{n,k}$, and X_k for $X_{n,k}$. To facilitate the proof define the mixed sums

$$U_l = \sum_{k=1}^{l-1} X_k + \sum_{k=l+1}^{n} Y_k \,, \qquad l = 1, \dots, n$$

Note the following identities

$$G_n = U_1 + Y_1, \quad U_l + X_l = U_{l+1} + Y_{l+1}, \;\; l = 1, \dots, n-1, \quad U_n + X_n = \widehat{S}_n \,,$$

which imply that,

$$|\mathbf{E}h(G_n) - \mathbf{E}h(\widehat{S}_n)| = |\mathbf{E}h(U_1 + Y_1) - \mathbf{E}h(U_n + X_n)| \le \sum_{l=1}^{n} \Delta_l \,, \tag{3.1.6}$$

where $\Delta_l = |\mathbf{E}[h(U_l + Y_l) - h(U_l + X_l)]|$, for $l = 1, \dots, n$. For any l and $\xi \in \mathbb{R}$, consider the remainder term

$$R_l(\xi) = h(U_l + \xi) - h(U_l) - \xi h'(U_l) - \frac{\xi^2}{2} h''(U_l)$$

in second order Taylor's expansion of $h(\cdot)$ at U_l. By Taylor's theorem, we have that

$$\begin{aligned} |R_l(\xi)| &\le \|h'''\|_\infty \frac{|\xi|^3}{6}, && \text{(from third order expansion)} \\ |R_l(\xi)| &\le \|h''\|_\infty |\xi|^2, && \text{(from second order expansion)} \end{aligned}$$

whence,

$$|R_l(\xi)| \le \min\left\{ \|h'''\|_\infty \frac{|\xi|^3}{6}, \|h''\|_\infty |\xi|^2 \right\} . \tag{3.1.7}$$

Considering the expectation of the difference between the two identities,

$$\begin{aligned} h(U_l + X_l) &= h(U_l) + X_l h'(U_l) + \frac{X_l^2}{2} h''(U_l) + R_l(X_l) \,, \\ h(U_l + Y_l) &= h(U_l) + Y_l h'(U_l) + \frac{Y_l^2}{2} h''(U_l) + R_l(Y_l) \,, \end{aligned}$$

we get that

$$\Delta_l \le \Big| \mathbf{E}[(X_l - Y_l) h'(U_l)] \Big| + \Big| \mathbf{E}\big[(\frac{X_l^2}{2} - \frac{Y_l^2}{2}) h''(U_l) \big] \Big| + |\mathbf{E}[R_l(X_l) - R_l(Y_l)]| \,.$$

Recall that X_l and Y_l are independent of U_l and chosen such that $\mathbf{E}X_l = \mathbf{E}Y_l$ and $\mathbf{E}X_l^2 = \mathbf{E}Y_l^2$. As the first two terms in the bound on Δ_l vanish we have that

$$\Delta_l \le \mathbf{E}|R_l(X_l)| + \mathbf{E}|R_l(Y_l)| \,. \tag{3.1.8}$$

Further, utilizing (3.1.7),

$$\mathbf{E}|R_l(X_l)| \le \|h'''\|_\infty \mathbf{E}\big[\frac{|X_l|^3}{6}; |X_l| \le \varepsilon\big] + \|h''\|_\infty \mathbf{E}[|X_l|^2; |X_l| \ge \varepsilon]$$

$$\le \frac{\varepsilon}{6}\|h'''\|_\infty \mathbf{E}[|X_l|^2] + \|h''\|_\infty \mathbf{E}[X_l^2; |X_l| \ge \varepsilon]\,.$$

Summing these bounds over $l = 1, \dots, n$, by our assumption that $\sum_{l=1}^n \mathbf{E}X_l^2 = v_n$ and the definition of $g_n(\varepsilon)$, we get that

$$\sum_{l=1}^n \mathbf{E}|R_l(X_l)| \le \frac{\varepsilon}{6} v_n \|h'''\|_\infty + g_n(\varepsilon)\|h''\|_\infty\,. \tag{3.1.9}$$

Recall that $Y_l/\sqrt{v_{n,l}}$ is a standard normal random variable, whose fourth moment is 3 (see (1.3.18)). By monotonicity in q of the L^q-norms (c.f. Lemma 1.3.16), it follows that $\mathbf{E}[|Y_l/\sqrt{v_{n,l}}|^3] \le 3$, hence $\mathbf{E}|Y_l|^3 \le 3v_{n,l}^{3/2} \le 3r_n v_{n,l}$. Utilizing once more (3.1.7) and the fact that $v_n = \sum_{l=1}^n v_{n,l}$, we arrive at

$$\sum_{l=1}^n \mathbf{E}|R_l(Y_l)| \le \frac{\|h'''\|_\infty}{6} \sum_{l=1}^n \mathbf{E}|Y_l|^3 \le \frac{r_n}{2} v_n \|h'''\|_\infty\,. \tag{3.1.10}$$

Plugging (3.1.8)–(3.1.10) into (3.1.6) completes the proof of the lemma. □

In view of (3.1.5), Lindeberg's CLT builds on the following elementary lemma, whereby we approximate the indicator function on $(-\infty, b]$ by continuous, bounded functions $h_k : \mathbb{R} \mapsto \mathbb{R}$ for each of which Lemma 3.1.5 applies.

Lemma 3.1.6. *There exist $h_k^\pm(x)$ of continuous and uniformly bounded first three derivatives, such that $0 \le h_k^-(x) \uparrow I_{(-\infty,b)}(x)$ and $1 \ge h_k^+(x) \downarrow I_{(-\infty,b]}(x)$ as $k \to \infty$.*

Proof. There are many ways to prove this. Here is one which is from first principles, hence requires no analysis knowledge. The function $\psi : [0,1] \mapsto [0,1]$ given by $\psi(x) = 140 \int_x^1 u^3(1-u)^3 du$ is monotone decreasing, with continuous derivatives of all order, such that $\psi(0) = 1$, $\psi(1) = 0$ and whose first three derivatives at 0 and at 1 are all zero. Its extension $\phi(x) = \psi(\min(x, 1)_+)$ to a function on $\mathbb{R}$ that is one for $x \le 0$ and zero for $x \ge 1$ is thus non-increasing, with continuous and uniformly bounded first three derivatives. It is easy to check that the translated and scaled functions $h_k^+(x) = \phi(k(x - b))$ and $h_k^-(x) = \phi(k(x - b) + 1)$ have all the claimed properties. □

Proof of Theorem 3.1.3. Applying (3.1.5) for $h_k^-(\cdot)$, then taking $k \to \infty$ we have by monotone convergence that

$$\liminf_{n\to\infty} \mathbf{P}(\widehat{S}_n < b) \ge \lim_{n\to\infty} \mathbf{E}[h_k^-(\widehat{S}_n)] = \mathbf{E}[h_k^-(G)] \uparrow F_G(b^-)\,.$$

Similarly, considering $h_k^+(\cdot)$, then taking $k \to \infty$ we have by bounded convergence that

$$\limsup_{n\to\infty} \mathbf{P}(\widehat{S}_n \le b) \le \lim_{n\to\infty} \mathbf{E}[h_k^+(\widehat{S}_n)] = \mathbf{E}[h_k^+(G)] \downarrow F_G(b)\,.$$

Since $F_G(\cdot)$ is a continuous function we conclude that $\mathbf{P}(\widehat{S}_n \le b)$ converges to $F_G(b) = F_G(b^-)$, as $n \to \infty$. This holds for every $b \in \mathbb{R}$ as claimed. □

3.1.2. Applications of the CLT. We start with the simpler, i.i.d. case. In doing so, we use the notation $Z_n \xrightarrow{\mathcal{D}} G$ when the analog of (3.1.3) holds for the sequence $\{Z_n\}$, that is $\mathbf{P}(Z_n \le b) \to \mathbf{P}(G \le b)$ as $n \to \infty$ for all $b \in \mathbb{R}$ (where G is a standard normal variable).

EXAMPLE 3.1.7 (NORMAL APPROXIMATION OF THE BINOMIAL). *Consider i.i.d. random variables $\{B_i\}$, each of whom is Bernoulli of parameter $0 < p < 1$ (i.e. $P(B_1 = 1) = 1 - P(B_1 = 0) = p$). The sum $S_n = B_1 + \cdots + B_n$ has the* Binomial distribution of parameters *(n, p), that is,*

$$\mathbf{P}(S_n = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, \dots, n.$$

For example, if B_i indicates that the ith independent toss of the same coin lands on a Head then S_n counts the total numbers of Heads in the first n tosses of the coin. Recall that $\mathbf{E}B = p$ and $\mathsf{Var}(B) = p(1-p)$ (see Example 1.3.69), so the CLT *states that $(S_n - np)/\sqrt{np(1-p)} \xrightarrow{\mathcal{D}} G$. It allows us to approximate, for all large enough n, the typically non-computable weighted sums of binomial terms by integrals with respect to the standard normal density.*

Here is another example that is similar and almost as widely used.

EXAMPLE 3.1.8 (NORMAL APPROXIMATION OF THE POISSON DISTRIBUTION). *It is not hard to verify that the sum of two independent Poisson random variables has the Poisson distribution, with a parameter which is the sum of the parameters of the summands. Thus, by induction, if $\{X_i\}$ are i.i.d. each of Poisson distribution of parameter 1, then $N_n = X_1 + \ldots + X_n$ has a Poisson distribution of parameter n. Since $\mathbf{E}(N_1) = \mathsf{Var}(N_1) = 1$ (see Example 1.3.69), the* CLT *applies for $(N_n - n)/n^{1/2}$. This provides an approximation for the distribution function of the Poisson variable N_λ of parameter λ that is a large integer. To deal with non-integer values $\lambda = n + \eta$ for some $\eta \in (0, 1)$, consider the mutually independent Poisson variables N_n, N_η and $N_{1-\eta}$. Since $N_\lambda \overset{\mathcal{D}}{=} N_n + N_\eta$ and $N_{n+1} \overset{\mathcal{D}}{=} N_n + N_\eta + N_{1-\eta}$, this provides a* monotone coupling *, that is, a construction of the random variables N_n, N_λ and N_{n+1} on the same probability space, such that $N_n \le N_\lambda \le N_{n+1}$. Because of this monotonicity, for any $\varepsilon > 0$ and all $n \ge n_0(b, \varepsilon)$ the event $\{(N_\lambda - \lambda)/\sqrt{\lambda} \le b\}$ is between $\{(N_{n+1} - (n+1))/\sqrt{n+1} \le b - \varepsilon\}$ and $\{(N_n - n)/\sqrt{n} \le b + \varepsilon\}$. Considering the limit as $n \to \infty$ followed by $\varepsilon \to 0$, it thus follows that the convergence $(N_n - n)/n^{1/2} \xrightarrow{\mathcal{D}} G$ implies also that $(N_\lambda - \lambda)/\lambda^{1/2} \xrightarrow{\mathcal{D}} G$ as $\lambda \to \infty$. In words, the normal distribution is a good approximation of a Poisson with large parameter.*

In Theorem 2.3.3 we established the strong law of large numbers when the summands X_i are only *pairwise independent.* Unfortunately, as the next example shows, pairwise independence is not good enough for the CLT.

EXAMPLE 3.1.9. *Consider i.i.d. $\{\xi_i\}$ such that $\mathbf{P}(\xi_i = 1) = \mathbf{P}(\xi_i = -1) = 1/2$ for all i. Set $X_1 = \xi_1$ and successively let $X_{2^k+j} = X_j \xi_{k+2}$ for $j = 1, \dots, 2^k$ and $k = 0, 1, \dots$. Note that each X_l is a $\{-1, 1\}$-valued variable, specifically, a product of a different finite subset of ξ_i-s that corresponds to the positions of ones in the binary representation of $2l - 1$ (with ξ_1 for its least significant digit, ξ_2 for the next digit, etc.). Consequently, each X_l is of zero mean and if $l \neq r$ then in $\mathbf{E}X_l X_r$ at least one of the ξ_i-s will appear exactly once, resulting with $\mathbf{E}X_l X_r = 0$, hence with $\{X_l\}$ being uncorrelated variables. Recall part (b) of Exercise 1.4.42, that such* *variables are pairwise independent. Further,* $\mathbf{E}X_l = 0$ *and* $X_l \in \{-1, 1\}$ *mean that* $\mathbf{P}(X_l = -1) = \mathbf{P}(X_l = 1) = 1/2$ *are identically distributed. As for the zero mean variables* $S_n = \sum_{j=1}^n X_j$*, we have arranged things such that* $S_1 = \xi_1$ *and for any* $k \geq 0$

$$S_{2^{k+1}} = \sum_{j=1}^{2^k} (X_j + X_{2^k+j}) = \sum_{j=1}^{2^k} X_j(1 + \xi_{k+2}) = S_{2^k}(1 + \xi_{k+2}) \, ,$$

hence $S_{2^k} = \xi_1 \prod_{i=2}^{k+1}(1 + \xi_i)$ *for all* $k \geq 1$*. In particular,* $S_{2^k} = 0$ *unless* $\xi_2 = \xi_3 = \ldots = \xi_{k+1} = 1$*, an event of probability* 2^{-k}*. Thus,* $\mathbf{P}(S_{2^k} \neq 0) = 2^{-k}$ *and certainly the* CLT *result (3.1.3) does not hold along the subsequence* $n = 2^k$*.*

We turn next to applications of Lindeberg's triangular array CLT, starting with the asymptotic of the count of record events till time $n \gg 1$.

Exercise 3.1.10. *Consider the count* R_n *of record events during the first* n *instances of i.i.d. R.V. with a continuous distribution function, as in Example 2.2.27. Recall that* $R_n = B_1 + \cdots + B_n$ *for mutually independent Bernoulli random variables* $\{B_k\}$ *such that* $\mathbf{P}(B_k = 1) = 1 - \mathbf{P}(B_k = 0) = k^{-1}$*.*

(a) *Check that* $b_n / \log n \to 1$ *where* $b_n = \mathsf{Var}(R_n)$*.*
(b) *Show that Lindeberg's* CLT *applies for* $X_{n,k} = (\log n)^{-1/2}(B_k - k^{-1})$*.*
(c) *Recall that* $|\mathbf{E}R_n - \log n| \leq 1$*, and conclude that* $(R_n - \log n)/\sqrt{\log n} \xrightarrow{\mathcal{D}} G$*.*

Remark. Let $\mathcal{S}_n$ denote the symmetric group of permutations on $\{1, \ldots, n\}$. For $s \in \mathcal{S}_n$ and $i \in \{1, \ldots, n\}$, denoting by $L_i(s)$ the smallest $j \leq n$ such that $s^j(i) = i$, we call $\{s^j(i) : 1 \leq j \leq L_i(s)\}$ the cycle of s containing i. If each $s \in \mathcal{S}_n$ is equally likely, then the law of the number $T_n(s)$ of different cycles in s is the same as that of R_n of Example 2.2.27 (for a proof see [**Dur03**, Example 1.5.4]). Consequently, Exercise 3.1.10 also shows that in this setting $(T_n - \log n)/\sqrt{\log n} \xrightarrow{\mathcal{D}} G$.

Part (a) of the following exercise is a special case of Lindeberg's CLT, known also as *Lyapunov's theorem.*

Exercise 3.1.11 (Lyapunov's theorem). *Let* $S_n = \sum_{k=1}^n X_k$ *for* $\{X_k\}$ *mutually independent such that* $v_n = \mathsf{Var}(S_n) < \infty$*.*

(a) *Show that if there exists* $q > 2$ *such that*

$$\lim_{n \to \infty} v_n^{-q/2} \sum_{k=1}^n \mathbf{E}(|X_k - \mathbf{E}X_k|^q) = 0 \, ,$$

then $v_n^{-1/2}(S_n - \mathbf{E}S_n) \xrightarrow{\mathcal{D}} G$*.*
(b) *Use the preceding result to show that* $n^{-1/2} S_n \xrightarrow{\mathcal{D}} G$ *when also* $\mathbf{E}X_k = 0$*,* $\mathbf{E}X_k^2 = 1$ *and* $\mathbf{E}|X_k|^q \leq C$ *for some* $q > 2$*,* $C < \infty$ *and* $k = 1, 2, \ldots$*.*
(c) *Show that* $(\log n)^{-1/2} S_n \xrightarrow{\mathcal{D}} G$ *when the mutually independent* X_k *are such that* $\mathbf{P}(X_k = 0) = 1 - k^{-1}$ *and* $\mathbf{P}(X_k = -1) = \mathbf{P}(X_k = 1) = 1/(2k)$*.*

The next application of Lindeberg's CLT involves the use of truncation (which we have already introduced in the context of the weak law of large numbers), to derive the CLT for normalized sums of certain i.i.d. random variables of *infinite variance.*

Proposition 3.1.12. *Suppose $\{X_k\}$ are i.i.d of symmetric distribution, that is $X_1 \overset{\mathcal{D}}{=} -X_1$ (or $\mathbf{P}(X_1 > x) = \mathbf{P}(X_1 < -x)$ for all x) such that $\mathbf{P}(|X_1| > x) = x^{-2}$ for $x \geq 1$. Then, $\frac{1}{\sqrt{n \log n}} \sum_{k=1}^n X_k \xrightarrow{\mathcal{D}} G$ as $n \to \infty$.*

Remark 3.1.13. Note that $\mathsf{Var}(X_1) = \mathbf{E}X_1^2 = \int_0^\infty 2x\mathbf{P}(|X_1| > x)dx = \infty$ (c.f. part (a) of Lemma 1.4.31), so the usual CLT of Proposition 3.1.2 does not apply here. Indeed, the infinite variance of the summands results in a different normalization of the sums $S_n = \sum_{k=1}^n X_k$ that is tailored to the specific tail behavior of $x \mapsto \mathbf{P}(|X_1| > x)$.

Caution should be exercised here, since when $\mathbf{P}(|X_1| > x) = x^{-\alpha}$ for $x > 1$ and some $0 < \alpha < 2$, there is no way to approximate the distribution of $(S_n - a_n)/b_n$ by the standard normal distribution. Indeed, in this case $b_n = n^{1/\alpha}$ and the approximation is by an α-stable law (c.f. Definition 3.3.31 and Exercise 3.3.33).

Proof. We plan to apply Lindeberg's CLT for the truncated random variables $X_{n,k} = b_n^{-1} X_k I_{|X_k| \leq c_n}$ where $b_n = \sqrt{n \log n}$ and $c_n \geq 1$ are such that both $c_n/b_n \to 0$ and $c_n/\sqrt{n} \to \infty$. Indeed, for each n the variables $X_{n,k}$, $k = 1, \ldots, n$, are i.i.d. of bounded and symmetric distribution (since both the distribution of X_k and the truncation function are symmetric). Consequently, $\mathbf{E}X_{n,k} = 0$ for all n and k. Further, we have chosen b_n such that

$$v_n = n\mathbf{E}X_{n,1}^2 = \frac{n}{b_n^2}\mathbf{E}X_1^2 I_{|X_1| \leq c_n} = \frac{n}{b_n^2}\int_0^{c_n} 2x[\mathbf{P}(|X_1| > x) - \mathbf{P}(|X_1| > c_n)]dx$$
$$= \frac{n}{b_n^2}\Big[\int_0^1 2x dx + \int_1^{c_n} \frac{2}{x} dx - \int_0^{c_n} \frac{2x}{c_n^2} dx\Big] = \frac{2n \log c_n}{b_n^2} \to 1$$

as $n \to \infty$. Finally, note that $|X_{n,k}| \leq c_n/b_n \to 0$ as $n \to \infty$, implying that $g_n(\varepsilon) = 0$ for any $\varepsilon > 0$ and all n large enough, hence Lindeberg's condition trivially holds. We thus deduce from Lindeberg's CLT that $\frac{1}{\sqrt{n \log n}}\overline{S}_n \xrightarrow{\mathcal{D}} G$ as $n \to \infty$, where $\overline{S}_n = \sum_{k=1}^n X_k I_{|X_k| \leq c_n}$ is the sum of the truncated variables. We have chosen the truncation level c_n large enough to assure that

$$\mathbf{P}(S_n \neq \overline{S}_n) \leq \sum_{k=1}^n \mathbf{P}(|X_k| > c_n) = n\mathbf{P}(|X_1| > c_n) = nc_n^{-2} \to 0$$

for $n \to \infty$, hence we may now conclude that $\frac{1}{\sqrt{n \log n}} S_n \xrightarrow{\mathcal{D}} G$ as claimed. □

We conclude this section with Kolmogorov's three series theorem, the most definitive result on the convergence of random series.

Theorem 3.1.14 (Kolmogorov's three series theorem). *Suppose $\{X_k\}$ are independent random variables. For non-random $c > 0$ let $X_n^{(c)} = X_n I_{|X_n| \leq c}$ be the corresponding truncated variables and consider the three series*

$$\sum_n \mathbf{P}(|X_n| > c), \qquad \sum_n \mathbf{E}X_n^{(c)}, \qquad \sum_n \mathsf{Var}(X_n^{(c)}). \tag{3.1.11}$$

Then, the random series $\sum_n X_n$ converges a.s. if and only if for some $c > 0$ all three series of (3.1.11) converge.

Remark. By convergence of a series we mean the existence of a finite limit to the sum of its first m entries when $m \to \infty$. Note that the theorem implies that if all three series of (3.1.11) converge for some $c > 0$, then they necessarily converge for every $c > 0$.

Proof. We prove the sufficiency first, that is, assume that for some $c > 0$ all three series of (3.1.11) converge. By Theorem 2.3.16 and the finiteness of $\sum_n \mathsf{Var}(X_n^{(c)})$ it follows that the random series $\sum_n (X_n^{(c)} - \mathbf{E}X_n^{(c)})$ converges a.s. Then, by our assumption that $\sum_n \mathbf{E}X_n^{(c)}$ converges, also $\sum_n X_n^{(c)}$ converges a.s. Further, by assumption the sequence of probabilities $\mathbf{P}(X_n \neq X_n^{(c)}) = \mathbf{P}(|X_n| > c)$ is summable, hence by Borel-Cantelli I, we have that a.s. $X_n \neq X_n^{(c)}$ for at most finitely many n's. The convergence a.s. of $\sum_n X_n^{(c)}$ thus results with the convergence a.s. of $\sum_n X_n$, as claimed.

We turn to prove the necessity of convergence of the three series in (3.1.11) to the convergence of $\sum_n X_n$, which is where we use the CLT. To this end, assume the random series $\sum_n X_n$ converges a.s. (to a finite limit) and fix an arbitrary constant $c > 0$. The convergence of $\sum_n X_n$ implies that $|X_n| \to 0$, hence a.s. $|X_n| > c$ for only finitely many n's. In view of the independence of these events and Borel-Cantelli II, necessarily the sequence $\mathbf{P}(|X_n| > c)$ is summable, that is, the series $\sum_n \mathbf{P}(|X_n| > c)$ converges. Further, the convergence a.s. of $\sum_n X_n$ then results with the a.s. convergence of $\sum_n X_n^{(c)}$.

Suppose now that the non-decreasing sequence $v_n = \sum_{k=1}^n \mathsf{Var}(X_k^{(c)})$ is unbounded, in which case the latter convergence implies that a.s. $T_n = v_n^{-1/2} \sum_{k=1}^n X_k^{(c)} \to 0$ when $n \to \infty$. We further claim that in this case Lindeberg's CLT applies for $\widehat{S}_n = \sum_{k=1}^n X_{n,k}$, where

$$X_{n,k} = v_n^{-1/2}(X_k^{(c)} - m_k^{(c)}), \quad \text{and} \quad m_k^{(c)} = \mathbf{E}X_k^{(c)}.$$

Indeed, per fixed n the variables $X_{n,k}$ are mutually independent of zero mean and such that $\sum_{k=1}^n \mathbf{E}X_{n,k}^2 = 1$. Further, since $|X_k^{(c)}| \le c$ and we assumed that $v_n \uparrow \infty$ it follows that $|X_{n,k}| \le 2c/\sqrt{v_n} \to 0$ as $n \to \infty$, resulting with Lindeberg's condition holding (as $g_n(\varepsilon) = 0$ when $\varepsilon > 2c/\sqrt{v_n}$, i.e. for all n large enough). Combining Lindeberg's CLT conclusion that $\widehat{S}_n \xrightarrow{\mathcal{D}} G$ and $T_n \overset{a.s.}{\to} 0$, we deduce that $(\widehat{S}_n - T_n) \xrightarrow{\mathcal{D}} G$ (c.f. Exercise 3.2.8). However, since $\widehat{S}_n - T_n = -v_n^{-1/2} \sum_{k=1}^n m_k^{(c)}$ are *non-random*, the sequence $\mathbf{P}(\widehat{S}_n - T_n \le 0)$ is composed of zeros and ones, hence cannot converge to $\mathbf{P}(G \le 0) = 1/2$. We arrive at a contradiction to our assumption that $v_n \uparrow \infty$, and so conclude that the sequence $\mathsf{Var}(X_n^{(c)})$ is summable, that is, the series $\sum_n \mathsf{Var}(X_n^{(c)})$ converges.

By Theorem 2.3.16, the summability of $\mathsf{Var}(X_n^{(c)})$ implies that the series $\sum_n (X_n^{(c)} - m_n^{(c)})$ converges a.s. We have already seen that $\sum_n X_n^{(c)}$ converges a.s. so it follows that their difference $\sum_n m_n^{(c)}$, which is the middle term of (3.1.11), converges as well. □

3.2. Weak convergence

Focusing here on the theory of *weak convergence*, we first consider in Subsection 3.2.1 the *convergence in distribution* in a more general setting than that of the CLT. This is followed by the study in Subsection 3.2.2 of weak convergence of probability measures and the theory associated with it. Most notably its relation to other modes of convergence, such as convergence in *total variation* or point-wise convergence of probability density functions. We conclude by introducing in Subsection 3.2.3 the key concept of *uniform tightness* which is instrumental to the derivation of weak convergence statements, as demonstrated in later sections of this chapter.

3.2.1. Convergence in distribution. Motivated by the CLT, we explore here the convergence in distribution, its relation to convergence in probability, some additional properties and examples in which the limiting law is not the normal law. To start off, here is the definition of convergence in distribution.

Definition 3.2.1. *We say that R.V.-s X_n* converge in distribution *to a R.V. X_∞, denoted by $X_n \xrightarrow{\mathcal{D}} X_\infty$, if $F_{X_n}(\alpha) \to F_{X_\infty}(\alpha)$ as $n \to \infty$ for each fixed α which is a continuity point of F_{X_∞}.*

Similarly, we say that distribution functions F_n converge weakly *to F_∞, denoted by $F_n \overset{w}{\to} F_\infty$, if $F_n(\alpha) \to F_\infty(\alpha)$ as $n \to \infty$ for each fixed α which is a continuity point of F_∞.*

Remark. If the limit R.V. X_∞ has a probability density function, or more generally whenever F_{X_∞} is a continuous function, the convergence in distribution of X_n to X_∞ is equivalent to the point-wise convergence of the corresponding distribution functions. Such is the case of the CLT, since the normal R.V. G has a density. Further,

Exercise 3.2.2. *Show that if $F_n \overset{w}{\to} F_\infty$ and $F_\infty(\cdot)$ is a continuous function then also* $\sup_x |F_n(x) - F_\infty(x)| \to 0$.

The CLT is not the only example of convergence in distribution we have already met. Recall the Glivenko-Cantelli theorem (see Theorem 2.3.6), whereby a.s. the empirical distribution functions F_n of an i.i.d. sequence of variables $\{X_i\}$ converge uniformly, hence point-wise to the true distribution function F_X.

Here is an explicit necessary and sufficient condition for the convergence in distribution of integer valued random variables

Exercise 3.2.3. *Let $X_n, 1 \le n \le \infty$ be integer valued R.V.-s. Show that $X_n \xrightarrow{\mathcal{D}} X_\infty$ if and only if $\mathbf{P}(X_n = k) \to_{n\to\infty} \mathbf{P}(X_\infty = k)$ for each $k \in \mathbf{Z}$.*

In contrast with all of the preceding examples, we demonstrate next why the convergence $X_n \xrightarrow{\mathcal{D}} X_\infty$ has been chosen to be strictly weaker than the point-wise convergence of the corresponding distribution functions. We also see that $\mathbf{E}h(X_n) \to \mathbf{E}h(X_\infty)$ or not, depending upon the choice of $h(\cdot)$, and even within the collection of continuous functions with image in $[-1, 1]$, the rate of this convergence is not uniform in h.

Example 3.2.4. *The random variables $X_n = 1/n$ converge in distribution to $X_\infty = 0$. Indeed, it is easy to check that $F_{X_n}(\alpha) = I_{[1/n,\infty)}(\alpha)$ converge to $F_{X_\infty}(\alpha) = I_{[0,\infty)}(\alpha)$ at each $\alpha \neq 0$. However, there is no convergence at the discontinuity point $\alpha = 0$ of F_{X_∞} as $F_{X_\infty}(0) = 1$ while $F_{X_n}(0) = 0$ for all n.*

Further, $\mathbf{E}h(X_n) = h(\frac{1}{n}) \to h(0) = \mathbf{E}h(X_\infty)$ if and only if $h(x)$ is continuous at $x = 0$, and the rate of convergence varies with the modulus of continuity of $h(x)$ at $x = 0$.

More generally, if $X_n = X + 1/n$ then $F_{X_n}(\alpha) = F_X(\alpha - 1/n) \to F_X(\alpha^-)$ as $n \to \infty$. So, in order for $X + 1/n$ to converge in distribution to X as $n \to \infty$, we *have to restrict such convergence to the continuity points of the limiting distribution function F_X, as done in Definition 3.2.1.*

We have seen in Examples 3.1.7 and 3.1.8 that the normal distribution is a good approximation for the Binomial and the Poisson distributions (when the corresponding parameter is large). Our next example is of the same type, now with the approximation of the Geometric distribution by the Exponential one.

Example 3.2.5 (Exponential approximation of the Geometric). *Let Z_p be a random variable with a Geometric distribution of parameter $p \in (0,1)$, that is, $\mathbf{P}(Z_p \geq k) = (1-p)^{k-1}$ for any positive integer k. As $p \to 0$, we see that*

$$\mathbf{P}(pZ_p > t) = (1-p)^{\lfloor t/p \rfloor} \to e^{-t} \qquad \text{for all} \quad t \geq 0$$

That is, $pZ_p \xrightarrow{\mathcal{D}} T$, with T having a standard exponential distribution. *As Z_p corresponds to the number of independent trials till the first occurrence of a specific event whose probability is p, this approximation corresponds to waiting for the occurrence of rare events.*

At this point, you are to check that convergence in probability implies the convergence in distribution, which is hence weaker than all notions of convergence explored in Section 1.3.3 (and is perhaps a reason for naming it weak convergence). The converse cannot hold, for example because convergence in distribution does not require X_n and X_∞ to be even defined on the same probability space. However, convergence in distribution is equivalent to convergence in probability when the limiting random variable is a non-random constant.

Exercise 3.2.6. *Show that if $X_n \overset{p}{\to} X_\infty$, then $X_n \xrightarrow{\mathcal{D}} X_\infty$. Conversely, if $X_n \xrightarrow{\mathcal{D}} X_\infty$ and X_∞ is almost surely a non-random constant, then $X_n \overset{p}{\to} X_\infty$.*

Further, as the next theorem shows, given $F_n \overset{w}{\to} F_\infty$, it is possible to construct random variables Y_n, $n \leq \infty$ such that $F_{Y_n} = F_n$ and $Y_n \overset{a.s.}{\to} Y_\infty$. The catch of course is to construct the appropriate *coupling*, that is, to specify the relation between the different Y_n's.

Theorem 3.2.7. *Let F_n be a sequence of distribution functions that converges weakly to F_∞. Then there exist random variables Y_n, $1 \leq n \leq \infty$ on the probability space $((0,1], \mathcal{B}_{(0,1]}, U)$ such that $F_{Y_n} = F_n$ for $1 \leq n \leq \infty$ and $Y_n \xrightarrow{a.s.} Y_\infty$.*

Proof. We use Skorokhod's representation as in the proof of Theorem 1.2.36. That is, for $\omega \in (0,1]$ and $1 \leq n \leq \infty$ let $Y_n^+(\omega) \geq Y_n^-(\omega)$ be

$$Y_n^+(\omega) = \sup\{y : F_n(y) \leq \omega\}, \qquad Y_n^-(\omega) = \sup\{y : F_n(y) < \omega\}\,.$$

While proving Theorem 1.2.36 we saw that $F_{Y_n^-} = F_n$ for any $n \leq \infty$, and as remarked there $Y_n^-(\omega) = Y_n^+(\omega)$ for all but at most countably many values of ω, hence $\mathbf{P}(Y_n^- = Y_n^+) = 1$. It thus suffices to show that for all $\omega \in (0,1)$,

$$\begin{aligned} Y_\infty^+(\omega) &\geq \limsup_{n\to\infty} Y_n^+(\omega) \geq \limsup_{n\to\infty} Y_n^-(\omega) \\ &\geq \liminf_{n\to\infty} Y_n^-(\omega) \geq Y_\infty^-(\omega)\,. \end{aligned} \tag{3.2.1}$$

Indeed, then $Y_n^-(\omega) \to Y_\infty^-(\omega)$ for any $\omega \in A = \{\omega : Y_\infty^+(\omega) = Y_\infty^-(\omega)\}$ where $\mathbf{P}(A) = 1$. Hence, setting $Y_n = Y_n^+$ for $1 \leq n \leq \infty$ would complete the proof of the theorem.

Turning to prove (3.2.1) note that the two middle inequalities are trivial. Fixing $\omega \in (0,1)$ we proceed to show that

$$Y_\infty^+(\omega) \geq \limsup_{n\to\infty} Y_n^+(\omega) \ . \tag{3.2.2}$$

Since the continuity points of F_∞ form a dense subset of $\mathbb{R}$ (see Exercise 1.2.38), it suffices for (3.2.2) to show that if $z > Y_\infty^+(\omega)$ is a continuity point of F_∞, then necessarily $z \geq Y_n^+(\omega)$ for all n large enough. To this end, note that $z > Y_\infty^+(\omega)$ implies by definition that $F_\infty(z) > \omega$. Since z is a continuity point of F_∞ and $F_n \xrightarrow{w} F_\infty$ we know that $F_n(z) \to F_\infty(z)$. Hence, $F_n(z) > \omega$ for all sufficiently large n. By definition of Y_n^+ and monotonicity of F_n, this implies that $z \geq Y_n^+(\omega)$, as needed. The proof of

$$\liminf_{n\to\infty} Y_n^-(\omega) \geq Y_\infty^-(\omega) \, , \tag{3.2.3}$$

is analogous. For $y < Y_\infty^-(\omega)$ we know by monotonicity of F_∞ that $F_\infty(y) < \omega$. Assuming further that y is a continuity point of F_∞, this implies that $F_n(y) < \omega$ for all sufficiently large n, which in turn results with $y \leq Y_n^-(\omega)$. Taking continuity points y_k of F_∞ such that $y_k \uparrow Y_\infty^-(\omega)$ will yield (3.2.3) and complete the proof. □

The next exercise provides useful ways to get convergence in distribution for one sequence out of that of another sequence. Its result is also called *the converging together lemma* or *Slutsky's lemma.*

Exercise 3.2.8. *Suppose that $X_n \xrightarrow{\mathcal{D}} X_\infty$ and $Y_n \xrightarrow{\mathcal{D}} Y_\infty$, where Y_∞ is non-random and for each n the variables X_n and Y_n are defined on the same probability space.*

(a) *Show that then $X_n + Y_n \xrightarrow{\mathcal{D}} X_\infty + Y_\infty$.*
Hint: *Recall that the collection of continuity points of F_{X_∞} is dense.*
(b) *Deduce that if $Z_n - X_n \xrightarrow{\mathcal{D}} 0$ then $X_n \xrightarrow{\mathcal{D}} X$ if and only if $Z_n \xrightarrow{\mathcal{D}} X$.*
(c) *Show that $Y_n X_n \xrightarrow{\mathcal{D}} Y_\infty X_\infty$.*

For example, here is an application of Exercise 3.2.8 en-route to a CLT connected to *renewal theory.*

Exercise 3.2.9.

(a) *Suppose $\{N_m\}$ are non-negative integer-valued random variables and $b_m \to \infty$ are non-random integers such that $N_m/b_m \xrightarrow{p} 1$. Show that if $S_n = \sum_{k=1}^n X_k$ for i.i.d. random variables $\{X_k\}$ with $v = \mathsf{Var}(X_1) \in (0,\infty)$ and $\mathbf{E}(X_1) = 0$, then $S_{N_m}/\sqrt{vb_m} \xrightarrow{\mathcal{D}} G$ as $m \to \infty$.*
Hint: *Use Kolmogorov's inequality to show that $S_{N_m}/\sqrt{vb_m} - S_{b_m}/\sqrt{vb_m} \xrightarrow{p} 0$.*
(b) *Let $N_t = \sup\{n : S_n \leq t\}$ for $S_n = \sum_{k=1}^n Y_k$ and i.i.d. random variables $Y_k > 0$ such that $v = \mathsf{Var}(Y_1) \in (0,\infty)$ and $\mathbf{E}(Y_1) = 1$. Show that $(N_t - t)/\sqrt{vt} \xrightarrow{\mathcal{D}} G$ as $t \to \infty$.*

Theorem 3.2.7 is key to solving the following:

Exercise 3.2.10. *Suppose that $Z_n \xrightarrow{\mathcal{D}} Z_\infty$. Show that then $b_n(f(c + Z_n/b_n) - f(c))/f'(c) \xrightarrow{\mathcal{D}} Z_\infty$ for every positive constants $b_n \to \infty$ and every Borel function*

$f : \mathbb{R} \to \mathbb{R}$ (not necessarily continuous) that is differentiable at $c \in \mathbb{R}$, with a derivative $f'(c) \neq 0$.

Consider the following exercise as a cautionary note about your interpretation of Theorem 3.2.7.

Exercise 3.2.11. *Let $M_n = \sum_{k=1}^n \prod_{i=1}^k U_i$ and $W_n = \sum_{k=1}^n \prod_{i=k}^n U_i$, where $\{U_i\}$ are i.i.d. uniformly on $[0, c]$ and $c > 0$.*

(a) *Show that $M_n \xrightarrow{a.s.} M_\infty$ as $n \to \infty$, with M_∞ taking values in $[0, \infty]$.*
(b) *Prove that M_∞ is a.s. finite if and only if $c < e$ (but $\mathbf{E}M_\infty$ is finite only for $c < 2$).*
(c) *In case $c < e$ prove that $W_n \xrightarrow{\mathcal{D}} M_\infty$ as $n \to \infty$ while W_n can not have an almost sure limit. Explain why this does not contradict Theorem 3.2.7.*

The next exercise relates the decay (in n) of $\sup_s |F_{X_\infty}(s) - F_{X_n}(s)|$ to that of $\sup |\mathbf{E}h(X_n) - \mathbf{E}h(X_\infty)|$ over all functions $h : \mathbb{R} \mapsto [-M, M]$ with $\sup_x |h'(x)| \le L$.

Exercise 3.2.12. *Let $\Delta_n = \sup_s |F_{X_\infty}(s) - F_{X_n}(s)|$.*

(a) *Show that if $\sup_x |h(x)| \le M$ and $\sup_x |h'(x)| \le L$, then for any $b > a$, $C = 4M + L(b-a)$ and all n*
$$|\mathbf{E}h(X_n) - \mathbf{E}h(X_\infty)| \le C\Delta_n + 4M\mathbf{P}(X_\infty \notin [a, b]) .$$
(b) *Show that if $X_\infty \in [a, b]$ and $f_{X_\infty}(x) \ge \eta > 0$ for all $x \in [a, b]$, then $|Q_n(\alpha) - Q_\infty(\alpha)| \le \eta^{-1}\Delta_n$ for any $\alpha \in (\Delta_n, 1 - \Delta_n)$, where $Q_n(\alpha) = \sup\{x : F_{X_n}(x) < \alpha\}$ denotes α-quantile for the law of X_n. Using this, construct $Y_n \overset{\mathcal{D}}{=} X_n$ such that $\mathbf{P}(|Y_n - Y_\infty| > \eta^{-1}\Delta_n) \le 2\Delta_n$ and deduce the bound of part (a), albeit the larger value $4M + L/\eta$ of C.*

Here is another example of convergence in distribution, this time in the context of extreme value theory.

Exercise 3.2.13. *Let $M_n = \max_{1 \le i \le n}\{T_i\}$, where T_i, $i = 1, 2, \ldots$ are i.i.d. random variables of distribution function $F_T(t)$. Noting that $F_{M_n}(x) = F_T(x)^n$, show that $b_n^{-1}(M_n - a_n) \xrightarrow{\mathcal{D}} M_\infty$ when:*

(a) *$F_T(t) = 1 - e^{-t}$ for $t \ge 0$ (i.e. T_i are Exponential of parameter one). Here, $a_n = \log n$, $b_n = 1$ and $F_{M_\infty}(y) = \exp(-e^{-y})$ for $y \in \mathbb{R}$.*
(b) *$F_T(t) = 1 - t^{-\alpha}$ for $t \ge 1$ and $\alpha > 0$. Here, $a_n = 0$, $b_n = n^{1/\alpha}$ and $F_{M_\infty}(y) = \exp(-y^{-\alpha})$ for $y > 0$.*
(c) *$F_T(t) = 1 - |t|^\alpha$ for $-1 \le t \le 0$ and $\alpha > 0$. Here, $a_n = 0$, $b_n = n^{-1/\alpha}$ and $F_{M_\infty}(y) = \exp(-|y|^\alpha)$ for $y \le 0$.*

Remark. Up to the linear transformation $y \mapsto (y - \mu)/\sigma$, the three distributions of M_∞ provided in Exercise 3.2.13 are the only possible limits of maxima of i.i.d. random variables. They are thus called the *extreme value distributions* of Type 1 (or Gumbel-type), in case (a), Type 2 (or Fréchet-type), in case (b), and Type 3 (or Weibull-type), in case (c). Indeed,

Exercise 3.2.14.

(a) *Building upon part* (a) *of Exercise 2.2.24, show that if G has the standard normal distribution, then for any $y \in \mathbb{R}$*
$$\lim_{t \to \infty} \frac{1 - F_G(t + y/t)}{1 - F_G(t)} = e^{-y} .$$

(b) *Let $M_n = \max_{1 \le i \le n} \{G_i\}$ for i.i.d. standard normal random variables G_i. Show that $b_n(M_n - b_n) \xrightarrow{\mathcal{D}} M_\infty$ where $F_{M_\infty}(y) = \exp(-e^{-y})$ and b_n is such that $1 - F_G(b_n) = n^{-1}$.*

(c) *Show that $b_n/\sqrt{2\log n} \to 1$ as $n \to \infty$ and deduce that $M_n/\sqrt{2\log n} \xrightarrow{p} 1$.*

(d) *More generally, suppose $T_t = \inf\{x \ge 0 : M_x \ge t\}$, where $x \mapsto M_x$ is some monotone non-decreasing family of random variables such that $M_0 = 0$. Show that if $e^{-t}T_t \xrightarrow{\mathcal{D}} T_\infty$ as $t \to \infty$ with T_∞ having the standard* exponential distribution *then $(M_x - \log x) \xrightarrow{\mathcal{D}} M_\infty$ as $x \to \infty$, where $F_{M_\infty}(y) = \exp(-e^{-y})$.*

Our next example is of a more combinatorial flavor.

EXERCISE 3.2.15 (THE BIRTHDAY PROBLEM). *Suppose $\{X_i\}$ are i.i.d. with each X_i uniformly distributed on $\{1, \dots, n\}$. Let $T_n = \min\{k : X_k = X_l$, for some $l < k\}$ mark the first coincidence among the entries of the sequence $X_1, X_2, \dots$, so*

$$\mathbf{P}(T_n > r) = \prod_{k=2}^{r} (1 - \frac{k-1}{n}),$$

is the probability that among r items chosen uniformly and independently from a set of n different objects, no two are the same (the name "birthday problem" corresponds to $n = 365$ with the items interpreted as the birthdays for a group of size r). Show that $\mathbf{P}(n^{-1/2}T_n > s) \to \exp(-s^2/2)$ as $n \to \infty$, for any fixed $s \ge 0$.
Hint: *Recall that $-x - x^2 \le \log(1-x) \le -x$ for $x \in [0, 1/2]$.*

The *symmetric, simple random walk* on the integers is the sequence of random variables $S_n = \sum_{k=1}^n \xi_k$ where ξ_k are i.i.d. such that $\mathbf{P}(\xi_k = 1) = \mathbf{P}(\xi_k = -1) = \frac{1}{2}$. From the CLT we already know that $n^{-1/2}S_n \xrightarrow{\mathcal{D}} G$. The next exercise provides the asymptotics of the first and last visits to zero by this random sequence, namely $R = \inf\{\ell \ge 1 : S_\ell = 0\}$ and $L_n = \sup\{\ell \le n : S_\ell = 0\}$. Much more is known about this random sequence (c.f. [**Dur03**, Section 3.3] or [**Fel68**, Chapter 3]).

EXERCISE 3.2.16. *Let $q_{n,r} = \mathbf{P}(S_1 > 0, \dots, S_{n-1} > 0, S_n = r)$ and*

$$p_{n,r} = \mathbf{P}(S_n = r) = 2^{-n}\binom{n}{k} \qquad k = (n+r)/2\,.$$

(a) *Counting paths of the walk, prove the* discrete reflection principle *that $\mathbf{P}_x(R < n, S_n = y) = \mathbf{P}_{-x}(S_n = y) = p_{n,x+y}$ for any positive integers x, y, where $\mathbf{P}_x(\cdot)$ denote probabilities for the walk starting at $S_0 = x$.*

(b) *Verify that $q_{n,r} = \frac{1}{2}(p_{n-1,r-1} - p_{n-1,r+1})$ for any $n, r \ge 1$.*
Hint: *Paths of the walk contributing to $q_{n,r}$ must have $S_1 = 1$. Hence, use part (a) with $x = 1$ and $y = r$.*

(c) *Deduce that $\mathbf{P}(R > n) = p_{n-1,0} + p_{n-1,1}$ and that $\mathbf{P}(L_{2n} = 2k) = p_{2k,0}p_{2n-2k,0}$ for $k = 0, 1, \dots, n$.*

(d) *Using* Stirling's formula *(that $\sqrt{2\pi n}(n/e)^n/n! \to 1$ as $n \to \infty$), show that $\sqrt{\pi n}\mathbf{P}(R > 2n) \to 1$ and that $(2n)^{-1}L_{2n} \xrightarrow{\mathcal{D}} X$, where X has the* arc-sine *probability density function $f_X(x) = \frac{1}{\pi\sqrt{x(1-x)}}$ on $[0, 1]$.*

(e) *Let H_{2n} count the number of $1 \le k \le 2n$ such that $S_k \ge 0$ and $S_{k-1} \ge 0$. Show that $H_{2n} \overset{\mathcal{D}}{=} L_{2n}$, hence $(2n)^{-1}H_{2n} \xrightarrow{\mathcal{D}} X$.*

3.2.2. Weak convergence of probability measures. We first extend the definition of weak convergence from distribution functions to measures on Borel σ-algebras.

DEFINITION 3.2.17. *For a topological space $\mathbb{S}$, let $C_b(\mathbb{S})$ denote the collection of all* continuous bounded functions *on $\mathbb{S}$. We say that a sequence of probability measures ν_n on a topological space $\mathbb{S}$ equipped with its Borel σ-algebra (see Example 1.1.15),* converges weakly *to a probability measure ν_∞, denoted $\nu_n \overset{w}{\Rightarrow} \nu_\infty$, if $\nu_n(h) \to \nu_\infty(h)$ for each $h \in C_b(\mathbb{S})$.*

As we show next, Definition 3.2.17 is an alternative definition of convergence in distribution, which, in contrast to Definition 3.2.1, applies to more general R.V. (for example to the $\mathbb{R}^d$-valued random variables we consider in Section 3.5).

PROPOSITION 3.2.18. *The weak convergence of distribution functions is equivalent to the weak convergence of the corresponding laws as probability measures on $(\mathbb{R}, \mathcal{B})$. Consequently, $X_n \xrightarrow{\mathcal{D}} X_\infty$ if and only if for each $h \in C_b(\mathbb{R})$, we have $\mathbf{E}h(X_n) \to \mathbf{E}h(X_\infty)$ as $n \to \infty$.*

PROOF. Suppose first that $F_n \overset{w}{\to} F_\infty$ and let Y_n, $1 \le n \le \infty$ be the random variables given by Theorem 3.2.7 such that $Y_n \overset{a.s.}{\to} Y_\infty$. For $h \in C_b(\mathbb{R})$ we have by continuity of h that $h(Y_n) \overset{a.s.}{\to} h(Y_\infty)$, and by bounded convergence also

$$\mathcal{P}_n(h) = \mathbf{E}(h(Y_n)) \to \mathbf{E}(h(Y_\infty)) = \mathcal{P}_\infty(h)\,.$$

Conversely, suppose that $\mathcal{P}_n \overset{w}{\Rightarrow} \mathcal{P}_\infty$ per Definition 3.2.17. Fixing $\alpha \in \mathbb{R}$, let the non-negative $h_k^\pm \in C_b(\mathbb{R})$ be such that $h_k^-(x) \uparrow I_{(-\infty,\alpha)}(x)$ and $h_k^+(x) \downarrow I_{(-\infty,\alpha]}(x)$ as $k \to \infty$ (c.f. Lemma 3.1.6 for a construction of such functions). We have by the weak convergence of the laws when $n \to \infty$, followed by monotone convergence as $k \to \infty$, that

$$\liminf_{n\to\infty} \mathcal{P}_n((-\infty,\alpha)) \ge \lim_{n\to\infty} \mathcal{P}_n(h_k^-) = \mathcal{P}_\infty(h_k^-) \uparrow \mathcal{P}_\infty((-\infty,\alpha)) = F_\infty(\alpha^-)\,.$$

Similarly, considering $h_k^+(\cdot)$ and then $k \to \infty$, we have by bounded convergence that

$$\limsup_{n\to\infty} \mathcal{P}_n((-\infty,\alpha]) \le \lim_{n\to\infty} \mathcal{P}_n(h_k^+) = \mathcal{P}_\infty(h_k^+) \downarrow \mathcal{P}_\infty((-\infty,\alpha]) = F_\infty(\alpha)\,.$$

For any continuity point α of F_∞ we conclude that $F_n(\alpha) = \mathcal{P}_n((-\infty,\alpha])$ converges as $n \to \infty$ to $F_\infty(\alpha) = F_\infty(\alpha^-)$, thus completing the proof. □

By yet another application of Theorem 3.2.7 we find that convergence in distribution is preserved under a.s. continuous mappings (see Corollary 2.2.13 for the analogous statement for convergence in probability).

PROPOSITION 3.2.19 (CONTINUOUS MAPPING). *For a Borel function g let $\mathbf{D}_g$ denote its set of points of discontinuity. If $X_n \xrightarrow{\mathcal{D}} X_\infty$ and $\mathbf{P}(X_\infty \in \mathbf{D}_g) = 0$, then $g(X_n) \xrightarrow{\mathcal{D}} g(X_\infty)$. If in addition g is bounded then $\mathbf{E}g(X_n) \to \mathbf{E}g(X_\infty)$.*

PROOF. Given $X_n \xrightarrow{\mathcal{D}} X_\infty$, by Theorem 3.2.7 there exists $Y_n \overset{\mathcal{D}}{=} X_n$, such that $Y_n \xrightarrow{a.s.} Y_\infty$. Fixing $h \in C_b(\mathbb{R})$, clearly $\mathbf{D}_{h\circ g} \subseteq \mathbf{D}_g$, so

$$\mathbf{P}(Y_\infty \in \mathbf{D}_{h\circ g}) \le \mathbf{P}(Y_\infty \in \mathbf{D}_g) = 0.$$

Therefore, by Exercise 2.2.12, it follows that $h(g(Y_n)) \xrightarrow{a.s.} h(g(Y_\infty))$. Since $h \circ g$ is bounded and $Y_n \overset{\mathcal{D}}{=} X_n$ for all n, it follows by bounded convergence that

$$\mathbf{E}h(g(X_n)) = \mathbf{E}h(g(Y_n)) \to \mathbf{E}(h(g(Y_\infty)) = \mathbf{E}h(g(X_\infty)) \,.$$

This holds for any $h \in C_b(\mathbb{R})$, so by Proposition 3.2.18, we conclude that $g(X_n) \xrightarrow{\mathcal{D}} g(X_\infty)$. □

Our next theorem collects several equivalent characterizations of weak convergence of probability measures on $(\mathbb{R}, \mathcal{B})$. To this end we need the following definition.

Definition 3.2.20. *For a subset A of a topological space $\mathbb{S}$, we denote by ∂A the* boundary *of A, that is $\partial A = \overline{A} \setminus A^o$ is the closed set of points in the closure of A but not in the interior of A. For a measure μ on $(\mathbb{S}, \mathcal{B}_\mathbb{S})$ we say that $A \in \mathcal{B}_\mathbb{S}$ is a* μ-continuity set *if $\mu(\partial A) = 0$.*

Theorem 3.2.21 (portmanteau theorem). *The following four statements are equivalent for any probability measures ν_n, $1 \le n \le \infty$ on $(\mathbb{R}, \mathcal{B})$.*

(a) $\nu_n \overset{w}{\Rightarrow} \nu_\infty$
(b) *For every closed set F, one has* $\limsup\limits_{n\to\infty} \nu_n(F) \le \nu_\infty(F)$
(c) *For every open set G, one has* $\liminf\limits_{n\to\infty} \nu_n(G) \ge \nu_\infty(G)$
(d) *For every ν_∞-continuity set A, one has* $\lim\limits_{n\to\infty} \nu_n(A) = \nu_\infty(A)$

Remark. As shown in Subsection 3.5.1, this theorem holds with $(\mathbb{R}, \mathcal{B})$ replaced by any metric space $\mathbb{S}$ and its Borel σ-algebra $\mathcal{B}_\mathbb{S}$.

For $\nu_n = \mathcal{P}_{X_n}$ we get the formulation of the Portmanteau theorem for random variables X_n, $1 \le n \le \infty$, where the following four statements are then equivalent to $X_n \xrightarrow{\mathcal{D}} X_\infty$:

(a) $\mathbf{E}h(X_n) \to \mathbf{E}h(X_\infty)$ for each bounded continuous h
(b) For every closed set F one has $\limsup\limits_{n\to\infty} \mathbf{P}(X_n \in F) \le \mathbf{P}(X_\infty \in F)$
(c) For every open set G one has $\liminf\limits_{n\to\infty} \mathbf{P}(X_n \in G) \ge \mathbf{P}(X_\infty \in G)$
(d) For every Borel set A such that $\mathbf{P}(X_\infty \in \partial A) = 0$, one has $\lim\limits_{n\to\infty} \mathbf{P}(X_n \in A) = \mathbf{P}(X_\infty \in A)$

Proof. It suffices to show that $(a) \Rightarrow (b) \Rightarrow (c) \Rightarrow (d) \Rightarrow (a)$, which we shall establish in that order. To this end, with $F_n(x) = \nu_n((-\infty, x])$ denoting the corresponding distribution functions, we replace $\nu_n \overset{w}{\Rightarrow} \nu_\infty$ of (a) by the equivalent condition $F_n \overset{w}{\to} F_\infty$ (see Proposition 3.2.18).
$(a) \Rightarrow (b)$. Assuming $F_n \overset{w}{\to} F_\infty$, we have the random variables Y_n, $1 \le n \le \infty$ of Theorem 3.2.7, such that $\mathcal{P}_{Y_n} = \nu_n$ and $Y_n \overset{a.s.}{\to} Y_\infty$. Since F is closed, the function I_F is upper semi-continuous bounded by one, so it follows that a.s.

$$\limsup_{n\to\infty} I_F(Y_n) \le I_F(Y_\infty) \,,$$

and by Fatou's lemma,

$$\limsup_{n\to\infty} \nu_n(F) = \limsup_{n\to\infty} \mathbf{E}I_F(Y_n) \le \mathbf{E} \limsup_{n\to\infty} I_F(Y_n) \le \mathbf{E}I_F(Y_\infty) = \nu_\infty(F) \,,$$

as stated in (b).

$(b) \Rightarrow (c)$. The complement $F = G^c$ of an open set G is a closed set, so by (b) we have that

$$1 - \liminf_{n\to\infty} \nu_n(G) = \limsup_{n\to\infty} \nu_n(G^c) \leq \nu_\infty(G^c) = 1 - \nu_\infty(G)\,,$$

implying that (c) holds. In an analogous manner we can show that $(c) \Rightarrow (b)$, so (b) and (c) are equivalent.
$(c) \Rightarrow (d)$. Since (b) and (c) are equivalent, we assume now that both (b) and (c) hold. Then, applying (c) for the open set $G = A^o$ and (b) for the closed set $F = \overline{A}$ we have that

$$\begin{aligned} \nu_\infty(\overline{A}) &\geq \limsup_{n\to\infty} \nu_n(\overline{A}) \geq \limsup_{n\to\infty} \nu_n(A) \\ &\geq \liminf_{n\to\infty} \nu_n(A) \geq \liminf_{n\to\infty} \nu_n(A^o) \geq \nu_\infty(A^o)\,. \end{aligned} \tag{3.2.4}$$

Further, $\overline{A} = A^o \cup \partial A$ so $\nu_\infty(\partial A) = 0$ implies that $\nu_\infty(\overline{A}) = \nu_\infty(A^o) = \nu_\infty(A)$ (with the last equality due to the fact that $A^o \subseteq A \subseteq \overline{A}$). Consequently, for such a set A all the inequalities in (3.2.4) are equalities, yielding (d).
$(d) \Rightarrow (a)$. Consider the set $A = (-\infty, \alpha]$ where α is a continuity point of F_∞. Then, $\partial A = \{\alpha\}$ and $\nu_\infty(\{\alpha\}) = F_\infty(\alpha) - F_\infty(\alpha^-) = 0$. Applying (d) for this choice of A, we have that

$$\lim_{n\to\infty} F_n(\alpha) = \lim_{n\to\infty} \nu_n((-\infty,\alpha]) = \nu_\infty((-\infty,\alpha]) = F_\infty(\alpha)\,,$$

which is our version of (a). □

We turn to relate the weak convergence to the convergence point-wise of probability density functions. To this end, we first define a new concept of convergence of measures, the *convergence in total-variation*.

Definition 3.2.22. *The total variation norm of a finite signed measure ν on the measurable space $(\mathbb{S}, \mathcal{F})$ is*

$$\|\nu\|_{tv} = \sup\{\nu(h) : h \in m\mathcal{F}, \sup_{s\in\mathbb{S}} |h(s)| \leq 1\}.$$

We say that a sequence of probability measures ν_n converges in total variation to a probability measure ν_∞, denoted $\nu_n \xrightarrow{t.v.} \nu_\infty$, if $\|\nu_n - \nu_\infty\|_{tv} \to 0$.

Remark. Note that $\|\nu\|_{tv} = 1$ for any probability measure ν (since $\nu(h) \leq \nu(|h|) \leq \|h\|_\infty \nu(1) \leq 1$ for the functions h considered, with equality for $h = 1$). By a similar reasoning, $\|\nu - \nu'\|_{tv} \leq 2$ for any two probability measures ν, ν' on $(\mathbb{S}, \mathcal{F})$.

Convergence in total-variation obviously implies weak convergence of the same probability measures, but the converse fails, as demonstrated for example by $\nu_n = \delta_{1/n}$, the probability measure on $(\mathbb{R}, \mathcal{B})$ assigning probability one to the point $1/n$, which converge weakly to $\nu_\infty = \delta_0$ (see Example 3.2.4), whereas $\|\nu_n - \nu_\infty\| = 2$ for all n. The difference of course has to do with the non-uniformity of the weak convergence with respect to the continuous function h.
To gain a better understanding of the convergence in total-variation, we consider an important special case.

Proposition 3.2.23. *Suppose $\mathbf{P} = f\mu$ and $\mathbf{Q} = g\mu$ for some measure μ on $(\mathbb{S}, \mathcal{F})$ and $f, g \in m\mathcal{F}_+$ such that $\mu(f) = \mu(g) = 1$. Then,*

$$\|\mathbf{P} - \mathbf{Q}\|_{tv} = \int_{\mathbb{S}} |f(s) - g(s)| d\mu(s)\,. \tag{3.2.5}$$

Further, suppose $\nu_n = f_n\mu$ with $f_n \in m\mathcal{F}_+$ such that $\mu(f_n) = 1$ for all $n \leq \infty$. Then, $\nu_n \xrightarrow{t.v.} \nu_\infty$ if $f_n(s) \to f_\infty(s)$ for μ-almost-every $s \in \mathbb{S}$.

Proof. For any measurable function $h : \mathbb{S} \mapsto [-1, 1]$ we have that

$$(f\mu)(h) - (g\mu)(h) = \mu(fh) - \mu(gh) = \mu((f - g)h) \leq \mu(|f - g|)\,,$$

with equality when $h(s) = \text{sgn}((f(s)-g(s))$ (see Proposition 1.3.56 for the left-most identity and note that fh and gh are in $L^1(\mathbb{S}, \mathcal{F}, \mu)$). Consequently, $\|\mathbf{P} - \mathbf{Q}\|_{tv} = \sup\{(f\mu)(h) - (g\mu)(h) : h \text{ as above }\} = \mu(|f - g|)$, as claimed.

For $\nu_n = f_n\mu$, we thus have that $\|\nu_n - \nu_\infty\|_{tv} = \mu(|f_n - f_\infty|)$, so the convergence in total-variation is equivalent to $f_n \to f_\infty$ in $L^1(\mathbb{S}, \mathcal{F}, \mu)$. Since $f_n \geq 0$ and $\mu(f_n) = 1$ for any $n \leq \infty$, it follows from Scheffé's lemma (see Lemma 1.3.35) that the latter convergence is a consequence of $f_n(s) \to f_\infty(s)$ for μ a.e. $s \in \mathbb{S}$. □

Two specific instances of Proposition 3.2.23 are of particular value in applications.

Example 3.2.24. *Let $\nu_n = \mathcal{P}_{X_n}$ denote the laws of random variables X_n that have probability density functions f_n, $n = 1, 2, \ldots, \infty$. Recall Exercise 1.3.66 that then $\nu_n = f_n\lambda$ for Lebesgue's measure λ on $(\mathbb{R}, \mathcal{B})$. Hence, by the preceding proposition, the convergence point-wise of $f_n(x)$ to $f_\infty(x)$ implies the convergence in total-variation of $\mathcal{P}_{X_n}$ to $\mathcal{P}_{X_\infty}$, and in particular implies that $X_n \xrightarrow{\mathcal{D}} X_\infty$.*

Example 3.2.25. *Similarly, if X_n are integer valued for $n = 1, 2\ldots$, then $\nu_n = f_n\widetilde{\lambda}$ for $f_n(k) = \mathbf{P}(X_n = k)$ and the* counting measure *$\widetilde{\lambda}$ on $(\mathbf{Z}, 2^{\mathbf{Z}})$ such that $\widetilde{\lambda}(\{k\}) = 1$ for each $k \in \mathbf{Z}$. So, by the preceding proposition, the point-wise convergence of Exercise 3.2.3 is not only necessary and sufficient for weak convergence but also for convergence in total-variation of the laws of X_n to that of X_∞.*

In the next exercise, you are to rephrase Example 3.2.25 in terms of the topological space of all probability measures on $\mathbf{Z}$.

Exercise 3.2.26. *Show that $d(\mu, \nu) = \|\mu - \nu\|_{tv}$ is a metric on the collection of all probability measures on $\mathbf{Z}$, and that in this space the convergence in total variation is equivalent to the weak convergence which in turn is equivalent to the point-wise convergence at each $x \in \mathbf{Z}$.*

Hence, under the framework of Example 3.2.25, the Glivenko-Cantelli theorem tells us that the empirical measures of integer valued i.i.d. R.V.-s $\{X_i\}$ converge in total-variation to the true law of X_1.

Here is an example from statistics that corresponds to the framework of Example 3.2.24.

Exercise 3.2.27. *Let V_{n+1} denote the central value on a list of $2n+1$ values (that is, the $(n + 1)$th largest value on the list). Suppose the list consists of mutually independent R.V., each chosen uniformly in $[0, 1)$.*

(a) *Show that V_{n+1} has probability density function $(2n + 1)\binom{2n}{n}v^n(1 - v)^n$ at each $v \in [0, 1)$.*

(b) *Verify that the density $f_n(v)$ of $\widehat{V}_n = \sqrt{2n}(2V_{n+1} - 1)$ is of the form $f_n(v) = c_n(1 - v^2/(2n))^n$ for some normalization constant c_n that is independent of $|v| \leq \sqrt{2n}$.*

(c) *Deduce that for $n \to \infty$ the densities $f_n(v)$ converge point-wise to the standard normal density, and conclude that $\widehat{V}_n \xrightarrow{\mathcal{D}} G$.*

Here is an interesting interpretation of the CLT in terms of weak convergence of probability measures.

EXERCISE 3.2.28. *Let $\mathcal{M}$ denote the set of probability measures ν on $(\mathbb{R},\mathcal{B})$ for which $\int x^2 d\nu(x) = 1$ and $\int x d\nu(x) = 0$, and $\gamma \in \mathcal{M}$ denote the standard normal distribution. Consider the mapping $T : \mathcal{M} \mapsto \mathcal{M}$ where $T\nu$ is the law of $(X_1 + X_2)/\sqrt{2}$ for X_1 and X_2 i.i.d. of law ν each. Explain why the* CLT *implies that $T^m \nu \overset{w}{\Rightarrow} \gamma$ as $m \to \infty$, for any $\nu \in \mathcal{M}$. Show that $T\gamma = \gamma$ (see Lemma 3.1.1), and explain why γ is the unique, globally attracting fixed point of T in $\mathcal{M}$.*

Your next exercise is the basis behind the celebrated *method of moments* for weak convergence.

EXERCISE 3.2.29. *Suppose that X and Y are $[0,1]$-valued random variables such that $\mathbf{E}(X^n) = \mathbf{E}(Y^n)$ for $n = 0, 1, 2, \ldots$.*

(a) *Show that $\mathbf{E}p(X) = \mathbf{E}p(Y)$ for any polynomial $p(\cdot)$.*
(b) *Show that $\mathbf{E}h(X) = \mathbf{E}h(Y)$ for any continuous function $h : [0,1] \mapsto \mathbb{R}$ and deduce that $X \overset{\mathcal{D}}{=} Y$.*

Hint: *Recall Weierstrass approximation theorem, that if h is continuous on $[0,1]$ then there exist polynomials p_n such that $\sup_{x\in[0,1]} |h(x) - p_n(x)| \to 0$ as $n \to \infty$.*

We conclude with the following example about weak convergence of measures in the space of infinite binary sequences.

EXERCISE 3.2.30. *Consider the topology of coordinate wise convergence on $\mathbb{S} = \{0,1\}^{\mathbb{N}}$ and the Borel probability measures $\{\nu_n\}$ on $\mathbb{S}$, where ν_n is the uniform measure over the $\binom{2n}{n}$ binary sequences of precisely n ones among the first $2n$ coordinates, followed by zeros from position $2n+1$ onwards. Show that $\nu_n \overset{w}{\Rightarrow} \nu_\infty$ where ν_∞ denotes the law of i.i.d. Bernoulli random variables of parameter $p = 1/2$.*
Hint: *Any open subset of $\mathbb{S}$ is a countable union of disjoint sets of the form $A_{\theta,k} = \{\omega \in \mathbb{S} : \omega_i = \theta_i, i = 1,\ldots,k\}$ for some $\theta = (\theta_1,\ldots,\theta_k) \in \{0,1\}^k$ and $k \in \mathbb{N}$.*

3.2.3. Uniform tightness and vague convergence.

So far we have studied the properties of weak convergence. We turn to deal with general ways to establish such convergence, a subject to which we return in Subsection 3.3.2. To this end, the most important concept is that of *uniform tightness*, which we now define.

DEFINITION 3.2.31. *We say that a probability measure μ on $(\mathbb{S},\mathcal{B}_{\mathbb{S}})$ is* tight *if for each $\varepsilon > 0$ there exists a compact set $K_\varepsilon \subseteq \mathbb{S}$ such that $\mu(K_\varepsilon^c) < \varepsilon$. A collection $\{\mu_\beta\}$ of probability measures on $(\mathbb{S},\mathcal{B}_{\mathbb{S}})$ is called* uniformly tight *if for each $\varepsilon > 0$ there exists* one *compact set K_ε such that $\mu_\beta(K_\varepsilon^c) < \varepsilon$ for all β.*

Since bounded closed intervals are compact and $[-M,M]^c \downarrow \emptyset$ as $M \uparrow \infty$, by continuity from above we deduce that each probability measure μ on $(\mathbb{R},\mathcal{B})$ is tight. The same argument applies for a finite collection of probability measures on $(\mathbb{R},\mathcal{B})$ (just choose the maximal value among the finitely many values of $M = M_\varepsilon$ that are needed for the different measures). Further, in the case of $\mathbb{S} = \mathbb{R}$ which we study here one can take without loss of generality the compact K_ε as a symmetric bounded interval $[-M_\varepsilon, M_\varepsilon]$, or even consider instead $(-M_\varepsilon, M_\varepsilon]$ (whose closure is compact) in order to simplify notations. Thus, expressing uniform tightness in terms of the corresponding distribution functions leads in this setting to the following alternative definition.

Definition 3.2.32. *A sequence of distribution functions F_n is called* uniformly tight*, if for every $\varepsilon > 0$ there exists $M = M_\varepsilon$ such that*

$$\limsup_{n\to\infty} [1 - F_n(M) + F_n(-M)] < \varepsilon \,.$$

Remark. As most texts use in the context of Definition 3.2.32 "tight" (or "tight sequence") instead of uniformly tight, we shall adopt the same convention here.

Uniform tightness of distribution functions has some structural resemblance to the U.I. condition (1.3.11). As such we have the following simple sufficient condition for uniform tightness (which is the analog of Exercise 1.3.54).

Exercise 3.2.33. *A sequence of probability measures ν_n on $(\mathbb{R}, \mathcal{B})$ is uniformly tight if $\sup_n \nu_n(f(|x|))$ is finite for some non-negative Borel function such that $f(r) \to \infty$ as $r \to \infty$. Alternatively, if $\sup_n Ef(|X_n|) < \infty$ then the distribution functions F_{X_n} form a tight sequence.*

The importance of uniform tightness is that it guarantees the existence of limit points for weak convergence.

Theorem 3.2.34 (Prohorov theorem). *A collection Γ of probability measures on a complete, separable metric space $\mathbb{S}$ equipped with its Borel σ-algebra $\mathcal{B}_{\mathbb{S}}$, is uniformly tight if and only if for any sequence $\nu_m \in \Gamma$ there exists a subsequence ν_{m_k} that converges weakly to some probability measure ν_∞ on $(\mathbb{S}, \mathcal{B}_{\mathbb{S}})$ (where ν_∞ is not necessarily in Γ and may depend on the subsequence m_k).*

Remark. For a proof of Prohorov's theorem, which is beyond the scope of these notes, see [**Dud89**, Theorem 11.5.4].

Instead of Prohorov's theorem, we prove here a bare-hands substitute for the special case $\mathbb{S} = \mathbb{R}$. When doing so, it is convenient to have the following notion of convergence of distribution functions.

Definition 3.2.35. *When a sequence F_n of distribution functions converges to a right continuous, non-decreasing function F_∞ at all continuous points of F_∞, we say that F_n* converges vaguely *to F_∞, denoted $F_n \xrightarrow{v} F_\infty$.*

In contrast with weak convergence, the vague convergence allows for the limit $F_\infty(x) = \nu_\infty((-\infty, x])$ to correspond to a measure ν_∞ such that $\nu_\infty(\mathbb{R}) < 1$.

Example 3.2.36. *Suppose $F_n = pI_{[n,\infty)} + qI_{[-n,\infty)} + (1-p-q)F$ for some $p, q \geq 0$ such that $p+q \leq 1$ and a distribution function F that is independent of n. It is easy to check that $F_n \xrightarrow{v} F_\infty$ as $n \to \infty$, where $F_\infty = q + (1-p-q)F$ is the distribution function of an $\overline{\mathbb{R}}$-valued random variable, with probability mass p at $+\infty$ and mass q at $-\infty$. If $p + q > 0$ then F_∞ is not a distribution function of any measure on $\mathbb{R}$ and F_n does not converge weakly.*

The preceding example is generic, that is, the space $\overline{\mathbb{R}}$ is compact, so the only loss of mass when dealing with weak convergence on $\mathbb{R}$ has to do with its escape to $\pm\infty$. It is thus not surprising that *every* sequence of distribution functions have vague limit points, as stated by the following theorem.

Theorem 3.2.37 (Helly's selection theorem). *For every sequence F_n of distribution functions, there is a subsequence F_{n_k} and a non-decreasing right continuous function F_∞ such that $F_{n_k}(y) \to F_\infty(y)$ as $k \to \infty$ at all continuity points y of F_∞, that is $F_{n_k} \xrightarrow{v} F_\infty$.*

Deferring the proof of Helly's theorem to the end of this section, uniform tightness is exactly what prevents probability mass from escaping to ±∞, thus assuring the existence of limit points for weak convergence.

LEMMA 3.2.38. *The sequence of distribution functions $\{F_n\}$ is uniformly tight if and only if each vague limit point of this sequence is a distribution function. That is, if and only if when $F_{n_k} \xrightarrow{v} F$, necessarily $1 - F(x) + F(-x) \to 0$ as $x \to \infty$.*

PROOF. Suppose first that $\{F_n\}$ is uniformly tight and $F_{n_k} \xrightarrow{v} F$. Fixing $\varepsilon > 0$, there exist $r_1 < -M_\varepsilon$ and $r_2 > M_\varepsilon$ that are both continuity points of F. Then, by the definition of vague convergence and the monotonicity of F_n,

$$\begin{aligned} 1 - F(r_2) + F(r_1) &= \lim_{k\to\infty} (1 - F_{n_k}(r_2) + F_{n_k}(r_1)) \\ &\le \limsup_{n\to\infty} (1 - F_n(M_\varepsilon) + F_n(-M_\varepsilon)) < \varepsilon\,. \end{aligned}$$

It follows that $\limsup_{x\to\infty}(1 - F(x) + F(-x)) \le \varepsilon$ and since $\varepsilon > 0$ is arbitrarily small, F must be a distribution function of some probability measure on $(\mathbb{R}, \mathcal{B})$.

Conversely, suppose $\{F_n\}$ is not uniformly tight, in which case by Definition 3.2.32, for some $\varepsilon > 0$ and $n_k \uparrow \infty$

$$1 - F_{n_k}(k) + F_{n_k}(-k) \ge \varepsilon \qquad \text{for all } k. \tag{3.2.6}$$

By Helly's theorem, there exists a vague limit point F to F_{n_k} as $k \to \infty$. That is, for some $k_l \uparrow \infty$ as $l \to \infty$ we have that $F_{n_{k_l}} \xrightarrow{v} F$. For any two continuity points $r_1 < 0 < r_2$ of F, we thus have by the definition of vague convergence, the monotonicity of $F_{n_{k_l}}$, and (3.2.6), that

$$\begin{aligned} 1 - F(r_2) + F(r_1) &= \lim_{l\to\infty} (1 - F_{n_{k_l}}(r_2) + F_{n_{k_l}}(r_1)) \\ &\ge \liminf_{l\to\infty} (1 - F_{n_{k_l}}(k_l) + F_{n_{k_l}}(-k_l)) \ge \varepsilon. \end{aligned}$$

Considering now $r = \min(-r_1, r_2) \to \infty$, this shows that $\inf_r(1-F(r)+F(-r)) \ge \varepsilon$, hence the vague limit point F cannot be a distribution function of a probability measure on $(\mathbb{R}, \mathcal{B})$. □

REMARK. Comparing Definitions 3.2.31 and 3.2.32 we see that if a collection Γ of probability measures on $(\mathbb{R}, \mathcal{B})$ is uniformly tight, then for any sequence $\nu_m \in \Gamma$ the corresponding sequence F_m of distribution functions is uniformly tight. In view of Lemma 3.2.38 and Helly's theorem, this implies the existence of a subsequence m_k and a distribution function F_∞ such that $F_{m_k} \xrightarrow{w} F_\infty$. By Proposition 3.2.18 we deduce that $\nu_{m_k} \overset{w}{\Rightarrow} \nu_\infty$, a probability measure on $(\mathbb{R}, \mathcal{B})$, thus proving the only direction of Prohorov's theorem that we ever use.

PROOF OF THEOREM 3.2.37. Fix a sequence of distribution function F_n. The key to the proof is to observe that there exists a sub-sequence n_k and a non-decreasing function $H : \mathbb{Q} \mapsto [0, 1]$ such that $F_{n_k}(q) \to H(q)$ for any $q \in \mathbb{Q}$.

This is done by a standard analysis argument called the principle of 'diagonal selection'. That is, let $q_1, q_2, \ldots$, be an enumeration of the set $\mathbb{Q}$ of all rational numbers. There exists then a limit point $H(q_1)$ to the sequence $F_n(q_1) \in [0, 1]$, that is a sub-sequence $n_k^{(1)}$ such that $F_{n_k^{(1)}}(q_1) \to H(q_1)$. Since $F_{n_k^{(1)}}(q_2) \in [0, 1]$, there exists a further sub-sequences $n_k^{(2)}$ of $n_k^{(1)}$ such that

$$F_{n_k^{(i)}}(q_i) \to H(q_i) \quad \text{for } i = 1, 2.$$

In the same manner we get a collection of nested sub-sequences $n_k^{(i)} \subseteq n_k^{(i-1)}$ such that

$$F_{n_k^{(i)}}(q_j) \to H(q_j), \qquad \text{for all } j \leq i.$$

The diagonal $n_k^{(k)}$ then has the property that

$$F_{n_k^{(k)}}(q_j) \to H(q_j), \qquad \text{for all } j,$$

so $n_k = n_k^{(k)}$ is our desired sub-sequence, and since each F_n is non-decreasing, the limit function H must also be non-decreasing on $\mathbb{Q}$.

Let $F_\infty(x) := \inf\{H(q) : q \in \mathbb{Q}, q > x\}$, noting that $F_\infty \in [0,1]$ is non-decreasing. Further, F_∞ is right continuous, since

$$\begin{aligned}\lim_{x_n \downarrow x} F_\infty(x_n) &= \inf\{H(q) : q \in \mathbb{Q}, q > x_n \text{ for some } n\} \\ &= \inf\{H(q) : q \in \mathbb{Q}, q > x\} = F_\infty(x).\end{aligned}$$

Suppose that x is a continuity point of the non-decreasing function F_∞. Then, for any $\varepsilon > 0$ there exists $y < x$ such that $F_\infty(x) - \varepsilon < F_\infty(y)$ and rational numbers $y < r_1 < x < r_2$ such that $H(r_2) < F_\infty(x) + \varepsilon$. It follows that

$$F_\infty(x) - \varepsilon < F_\infty(y) \leq H(r_1) \leq H(r_2) < F_\infty(x) + \varepsilon\,. \tag{3.2.7}$$

Recall that $F_{n_k}(x) \in [F_{n_k}(r_1), F_{n_k}(r_2)]$ and $F_{n_k}(r_i) \to H(r_i)$ as $k \to \infty$, for $i = 1, 2$. Thus, by (3.2.7) for all k large enough

$$F_\infty(x) - \varepsilon < F_{n_k}(r_1) \leq F_{n_k}(x) \leq F_{n_k}(r_2) < F_\infty(x) + \varepsilon,$$

which since $\varepsilon > 0$ is arbitrary implies $F_{n_k}(x) \to F_\infty(x)$ as $k \to \infty$. □

EXERCISE 3.2.39. *Suppose that the sequence of distribution functions $\{F_{X_k}\}$ is uniformly tight and $\mathbf{E}X_k^2 < \infty$ are such that $\mathbf{E}X_n^2 \to \infty$ as $n \to \infty$. Show that then also $\mathsf{Var}(X_n) \to \infty$ as $n \to \infty$.*

Hint: *If $|\mathbf{E}X_{n_l}|^2 \to \infty$ then $\sup_l \mathsf{Var}(X_{n_l}) < \infty$ yields $X_{n_l}/\mathbf{E}X_{n_l} \xrightarrow{L^2} 1$, whereas the uniform tightness of $\{F_{X_{n_l}}\}$ implies that $X_{n_l}/\mathbf{E}X_{n_l} \xrightarrow{p} 0$.*

Using Lemma 3.2.38 and Helly's theorem, you next explore the possibility of establishing weak convergence for non-negative random variables out of the convergence of the corresponding *Laplace transforms.*

EXERCISE 3.2.40.

(a) *Based on Exercise 3.2.29 show that if $Z \geq 0$ and $W \geq 0$ are such that $\mathbf{E}(e^{-sZ}) = \mathbf{E}(e^{-sW})$ for each $s > 0$, then $Z \overset{\mathcal{D}}{=} W$.*

(b) *Further, show that for any $Z \geq 0$, the function $L_Z(s) = \mathbf{E}(e^{-sZ})$ is infinitely differentiable at all $s > 0$ and for any positive integer k,*

$$\mathbf{E}[Z^k] = (-1)^k \lim_{s \downarrow 0} \frac{d^k}{ds^k} L_Z(s)\,,$$

even when (both sides are) infinite.

(c) *Suppose that $X_n \geq 0$ are such that $L(s) = \lim_n \mathbf{E}(e^{-sX_n})$ exists for all $s > 0$ and $L(s) \to 1$ for $s \downarrow 0$. Show that then the sequence of distribution functions $\{F_{X_n}\}$ is uniformly tight and that there exists a random variable $X_\infty \geq 0$ such that $X_n \xrightarrow{\mathcal{D}} X_\infty$ and $L(s) = \mathbf{E}(e^{-sX_\infty})$ for all $s > 0$.*

Hint: *To show that* $X_n \xrightarrow{\mathcal{D}} X_\infty$ *try reading and adapting the proof of Theorem 3.3.17.*

(d) *Let* $X_n = n^{-1}\sum_{k=1}^n kI_k$ *for* $I_k \in \{0,1\}$ *independent random variables, with* $\mathbf{P}(I_k = 1) = k^{-1}$. *Show that there exists* $X_\infty \geq 0$ *such that* $X_n \xrightarrow{\mathcal{D}} X_\infty$ *and* $\mathbf{E}(e^{-sX_\infty}) = \exp(\int_0^1 t^{-1}(e^{-st}-1)dt)$ *for all* $s > 0$.

Remark. The idea of using transforms to establish weak convergence shall be further developed in Section 3.3, with the *Fourier transform* instead of the Laplace transform.

3.3. Characteristic functions

This section is about the fundamental concept of characteristic function, its relevance for the theory of weak convergence, and in particular for the CLT.

In Subsection 3.3.1 we define the characteristic function, providing illustrating examples and certain general properties such as the relation between finite moments of a random variable and the degree of smoothness of its characteristic function. In Subsection 3.3.2 we recover the distribution of a random variable from its characteristic function, and building upon it, relate tightness and weak convergence with the point-wise convergence of the associated characteristic functions. We conclude with Subsection 3.3.3 in which we re-prove the CLT of Section 3.1 as an application of the theory of characteristic functions we have thus developed. The same approach will serve us well in other settings which we consider in the sequel (c.f. Sections 3.4 and 3.5).

3.3.1. Definition, examples, moments and derivatives. We start off with the definition of the characteristic function of a random variable. To this end, recall that a $\mathbb{C}$-valued random variable is a function $Z : \Omega \mapsto \mathbb{C}$ such that the real and imaginary parts of Z are measurable, and for $Z = X + iY$ with $X, Y \in \mathbb{R}$ integrable random variables (and $i = \sqrt{-1}$), let $\mathbf{E}(Z) = \mathbf{E}(X) + i\mathbf{E}(Y) \in \mathbb{C}$.

Definition 3.3.1. *The* characteristic function Φ_X *of a random variable* X *is the map* $\mathbb{R} \mapsto \mathbb{C}$ *given by*

$$\Phi_X(\theta) = \mathbf{E}[e^{i\theta X}] = \mathbf{E}[\cos(\theta X)] + i\mathbf{E}[\sin(\theta X)]$$

where $\theta \in \mathbb{R}$ *and obviously both* $\cos(\theta X)$ *and* $\sin(\theta X)$ *are integrable R.V.-s.*

We also denote by $\Phi_\mu(\theta)$ *the characteristic function associated with a probability measure* μ *on* $(\mathbb{R}, \mathcal{B})$. *That is,* $\Phi_\mu(\theta) = \mu(e^{i\theta x})$ *is the characteristic function of a R.V.* X *whose law* $\mathcal{P}_X$ *is* μ.

Here are some of the properties of characteristic functions, where the complex conjugate $x - iy$ of $z = x + iy \in \mathbb{C}$ is denoted throughout by $\overline{z}$ and the *modulus* of $z = x + iy$ is $|z| = \sqrt{x^2 + y^2}$.

Proposition 3.3.2. *Let* X *be a R.V. and* Φ_X *its characteristic function, then*

(a) $\Phi_X(0) = 1$

(b) $\Phi_X(-\theta) = \overline{\Phi_X(\theta)}$

(c) $|\Phi_X(\theta)| \leq 1$

(d) $\theta \mapsto \Phi_X(\theta)$ *is a uniformly continuous function on* $\mathbb{R}$

(e) $\Phi_{aX+b}(\theta) = e^{ib\theta}\Phi_X(a\theta)$

PROOF. For (a), $\Phi_X(0) = \mathbf{E}[e^{i0X}] = \mathbf{E}[1] = 1$. For (b), note that

$$\begin{aligned}\Phi_X(-\theta) &= \mathbf{E}\cos(-\theta X) + i\mathbf{E}\sin(-\theta X)\\ &= \mathbf{E}\cos(\theta X) - i\mathbf{E}\sin(\theta X) = \overline{\Phi_X(\theta)}\,.\end{aligned}$$

For (c), note that the function $|z| = \sqrt{x^2+y^2} : \mathbb{R}^2 \mapsto \mathbb{R}$ is convex, hence by Jensen's inequality (c.f. Exercise 1.3.20),

$$|\Phi_X(\theta)| = |\mathbf{E}e^{i\theta X}| \le \mathbf{E}|e^{i\theta X}| = 1$$

(since the modulus $|e^{i\theta x}| = 1$ for any real x and θ).

For (d), since $\Phi_X(\theta+h) - \Phi_X(\theta) = \mathbf{E}e^{i\theta X}(e^{ihX}-1)$, it follows by Jensen's inequality for the modulus function that

$$|\Phi_X(\theta+h) - \Phi_X(\theta)| \le \mathbf{E}[|e^{i\theta X}||e^{ihX}-1|] = \mathbf{E}|e^{ihX}-1| = \delta(h)$$

(using the fact that $|zv| = |z||v|$). Since $2 \ge |e^{ihX}-1| \to 0$ as $h \to 0$, by bounded convergence $\delta(h) \to 0$. As the bound $\delta(h)$ on the modulus of continuity of $\Phi_X(\theta)$ is independent of θ, we have uniform continuity of $\Phi_X(\cdot)$ on $\mathbb{R}$.

For (e) simply note that $\Phi_{aX+b}(\theta) = \mathbf{E}e^{i\theta(aX+b)} = e^{i\theta b}\mathbf{E}e^{i(a\theta)X} = e^{i\theta b}\Phi_X(a\theta)$. □

We also have the following relation between finite moments of the random variable and the derivatives of its characteristic function.

LEMMA 3.3.3. *If $\mathbf{E}|X|^n < \infty$, then the characteristic function $\Phi_X(\theta)$ of X has continuous derivatives up to the n-th order, given by*

$$\frac{d^k}{d\theta^k}\Phi_X(\theta) = \mathbf{E}[(iX)^k e^{i\theta X}]\,, \qquad \textit{for} \quad k = 1,\dots,n \tag{3.3.1}$$

PROOF. Note that for any $x, h \in \mathbb{R}$

$$e^{ihx} - 1 = ix\int_0^h e^{iux}du\,.$$

Consequently, for any $h \ne 0$, $\theta \in \mathbb{R}$ and positive integer k we have the identity

$$\begin{aligned}\Delta_{k,h}(x) &= h^{-1}\big((ix)^{k-1}e^{i(\theta+h)x} - (ix)^{k-1}e^{i\theta x}\big) - (ix)^k e^{i\theta x}\\ &= (ix)^k e^{i\theta x}h^{-1}\int_0^h (e^{iux}-1)du\,,\end{aligned} \tag{3.3.2}$$

from which we deduce that $|\Delta_{k,h}(x)| \le 2|x|^k$ for all θ and $h \ne 0$, and further that $|\Delta_{k,h}(x)| \to 0$ as $h \to 0$. Thus, for $k = 1,\dots,n$ we have by dominated convergence (and Jensen's inequality for the modulus function) that

$$|\mathbf{E}\Delta_{k,h}(X)| \le \mathbf{E}|\Delta_{k,h}(X)| \to 0 \quad \text{for} \quad h \to 0.$$

Taking $k = 1$, we have from (3.3.2) that

$$\mathbf{E}\Delta_{1,h}(X) = h^{-1}(\Phi_X(\theta+h) - \Phi_X(\theta)) - \mathbf{E}[iXe^{i\theta X}]\,,$$

so its convergence to zero as $h \to 0$ amounts to the identity (3.3.1) holding for $k = 1$. In view of this, considering now (3.3.2) for $k = 2$, we have that

$$\mathbf{E}\Delta_{2,h}(X) = h^{-1}(\Phi'_X(\theta+h) - \Phi'_X(\theta)) - \mathbf{E}[(iX)^2 e^{i\theta X}]\,,$$

and its convergence to zero as $h \to 0$ amounts to (3.3.1) holding for $k = 2$. We continue in this manner for $k = 3,\dots,n$ to complete the proof of (3.3.1). The continuity of the derivatives follows by dominated convergence from the convergence to zero of $|(ix)^k e^{i(\theta+h)x} - (ix)^k e^{i\theta x}| \le 2|x|^k$ as $h \to 0$ (with $k = 1,\dots,n$). □

The converse of Lemma 3.3.3 does not hold. That is, there exist random variables with $\mathbf{E}|X| = \infty$ for which $\Phi_X(\theta)$ is differentiable at $\theta = 0$ (c.f. Exercise 3.3.23). However, as we see next, the existence of a finite second derivative of $\Phi_X(\theta)$ at $\theta = 0$ implies that $\mathbf{E}X^2 < \infty$.

LEMMA 3.3.4. *If* $\liminf_{\theta\to 0} \theta^{-2}(2\Phi_X(0)-\Phi_X(\theta)-\Phi_X(-\theta)) < \infty$*, then* $\mathbf{E}X^2 < \infty$.

PROOF. Note that $\theta^{-2}(2\Phi_X(0) - \Phi_X(\theta) - \Phi_X(-\theta)) = \mathbf{E}g_\theta(X)$, where

$$g_\theta(x) = \theta^{-2}(2 - e^{i\theta x} - e^{-i\theta x}) = 2\theta^{-2}[1 - \cos(\theta x)] \to x^2 \quad \text{for} \quad \theta \to 0 .$$

Since $g_\theta(x) \geq 0$ for all θ and x, it follows by Fatou's lemma that

$$\liminf_{\theta\to 0} \mathbf{E}g_\theta(X) \geq \mathbf{E}[\liminf_{\theta\to 0} g_\theta(X)] = \mathbf{E}X^2 ,$$

thus completing the proof of the lemma. □

We continue with a few explicit computations of the characteristic function.

EXAMPLE 3.3.5. *Consider a* Bernoulli *random variable* B *of parameter* p*, that is,* $\mathbf{P}(B = 1) = p$ *and* $\mathbf{P}(B = 0) = 1 - p$*. Its characteristic function is by definition*

$$\Phi_B(\theta) = \mathbf{E}[e^{i\theta B}] = pe^{i\theta} + (1 - p)e^{i0\theta} = pe^{i\theta} + 1 - p .$$

The same type of explicit formula applies to any discrete valued R.V. For example, if N *has the* Poisson *distribution of parameter* λ *then*

$$\Phi_N(\theta) = \mathbf{E}[e^{i\theta N}] = \sum_{k=0}^{\infty} \frac{(\lambda e^{i\theta})^k}{k!} e^{-\lambda} = \exp(\lambda(e^{i\theta} - 1)) . \tag{3.3.3}$$

The characteristic function has an explicit form also when the R.V. X has a probability density function f_X as in Definition 1.2.39. Indeed, then by Corollary 1.3.62 we have that

$$\Phi_X(\theta) = \int_{\mathbb{R}} e^{i\theta x} f_X(x)dx , \tag{3.3.4}$$

which is merely the Fourier transform of the density f_X (and is well defined since $\cos(\theta x)f_X(x)$ and $\sin(\theta x)f_X(x)$ are both integrable with respect to Lebesgue's measure).

EXAMPLE 3.3.6. *If* G *has the* $\mathcal{N}(\mu, v)$ *distribution, namely, the probability density function* $f_G(y)$ *is given by (3.1.1), then its characteristic function is*

$$\Phi_G(\theta) = e^{i\mu\theta - v\theta^2/2} .$$

Indeed, recall Example 1.3.68 that $G = \sigma X + \mu$ *for* $\sigma = \sqrt{v}$ *and* X *of a standard normal distribution* $\mathcal{N}(0, 1)$*. Hence, considering part (e) of Proposition 3.3.2 for* $a = \sqrt{v}$ *and* $b = \mu$*, it suffices to show that* $\Phi_X(\theta) = e^{-\theta^2/2}$*. To this end, as* X *is integrable, we have from Lemma 3.3.3 that*

$$\Phi'_X(\theta) = \mathbf{E}(iXe^{i\theta X}) = \int_{\mathbb{R}} -x \sin(\theta x) f_X(x)dx$$

(since $x\cos(\theta x)f_X(x)$ *is an integrable odd function, whose integral is thus zero). The standard normal density is such that* $f'_X(x) = -xf_X(x)$*, hence integrating by parts we find that*

$$\Phi'_X(\theta) = \int_{\mathbb{R}} \sin(\theta x) f'_X(x)dx = -\int_{\mathbb{R}} \theta \cos(\theta x) f_X(x)dx = -\theta\Phi_X(\theta)$$

(since $\sin(\theta x) f_X(x)$ is an integrable odd function). We know that $\Phi_X(0) = 1$ and since $\varphi(\theta) = e^{-\theta^2/2}$ is the unique solution of the ordinary differential equation $\varphi'(\theta) = -\theta\varphi(\theta)$ with $\varphi(0) = 1$, it follows that $\Phi_X(\theta) = \varphi(\theta)$.

Example 3.3.7. *In another example, applying the formula (3.3.4) we see that the random variable $U = U(a, b)$ whose probability density function is $f_U(x) = (b-a)^{-1}\mathbf{1}_{a<x<b}$, has the characteristic function*

$$\Phi_U(\theta) = \frac{e^{i\theta b} - e^{i\theta a}}{i\theta(b-a)}$$

(recall that $\int_a^b e^{zx}dx = (e^{zb} - e^{za})/z$ for any $z \in \mathbb{C}$). For $a = -b$ the characteristic function simplifies to $\sin(b\theta)/(b\theta)$. Or, in case $b = 1$ and $a = 0$ we have $\Phi_U(\theta) = (e^{i\theta} - 1)/(i\theta)$ for the random variable U of Example 1.1.26.

For $a = 0$ and $z = -\lambda + i\theta$, $\lambda > 0$, the same integration identity applies also when $b \to \infty$ (since the real part of z is negative). Consequently, by (3.3.4), the exponential distribution *of parameter $\lambda > 0$ whose density is $f_T(t) = \lambda e^{-\lambda t}\mathbf{1}_{t>0}$ (see Example 1.3.68), has the characteristic function $\Phi_T(\theta) = \lambda/(\lambda - i\theta)$.*

Finally, for the density $f_S(s) = 0.5e^{-|s|}$ it is not hard to check that $\Phi_S(\theta) = 0.5/(1-i\theta) + 0.5/(1+i\theta) = 1/(1+\theta^2)$ (just break the integration over $s \in \mathbb{R}$ in (3.3.4) according to the sign of s).

We next express the characteristic function of the sum of independent random variables in terms of the characteristic functions of the summands. This relation makes the characteristic function a useful tool for proving weak convergence statements involving sums of independent variables.

Lemma 3.3.8. *If X and Y are two independent random variables, then*

$$\Phi_{X+Y}(\theta) = \Phi_X(\theta)\Phi_Y(\theta)$$

Proof. By the definition of the characteristic function

$$\Phi_{X+Y}(\theta) = \mathbf{E}e^{i\theta(X+Y)} = \mathbf{E}[e^{i\theta X}e^{i\theta Y}] = \mathbf{E}[e^{i\theta X}]\mathbf{E}[e^{i\theta Y}]\,,$$

where the right-most equality is obtained by the independence of X and Y (i.e. applying (1.4.12) for the integrable $f(x) = g(x) = e^{i\theta x}$). Observing that the right-most expression is $\Phi_X(\theta)\Phi_Y(\theta)$ completes the proof. □

Here are three simple applications of this lemma.

Example 3.3.9. *If X and Y are independent and uniform on $(-1/2, 1/2)$ then by Corollary 1.4.33 the random variable $\Delta = X + Y$ has the* triangular density, *$f_\Delta(x) = (1 - |x|)\mathbf{1}_{|x|\le 1}$. Thus, by Example 3.3.7, Lemma 3.3.8, and the trigonometric identity $\cos\theta = 1 - 2\sin^2(\theta/2)$ we have that its characteristic function is*

$$\Phi_\Delta(\theta) = [\Phi_X(\theta)]^2 = \Big(\frac{2\sin(\theta/2)}{\theta}\Big)^2 = \frac{2(1-\cos\theta)}{\theta^2}\,.$$

Exercise 3.3.10. *Let X, $\widetilde{X}$ be i.i.d. random variables.*

(a) *Show that the characteristic function of $Z = X - \widetilde{X}$ is a non-negative, real-valued function.*

(b) *Prove that there do not exist $a < b$ and i.i.d. random variables X, $\widetilde{X}$ such that $X - \widetilde{X}$ is the uniform random variable on (a, b).*

In the next exercise you construct a random variable X whose law has no atoms while its characteristic function does not converge to zero for $\theta \to \infty$.

EXERCISE 3.3.11. *Let $X = 2\sum_{k=1}^{\infty} 3^{-k} B_k$ for $\{B_k\}$ i.i.d. Bernoulli random variables such that $\mathbf{P}(B_k = 1) = \mathbf{P}(B_k = 0) = 1/2$.*

(a) *Show that $\Phi_X(3^k\pi) = \Phi_X(\pi) \neq 0$ for $k = 1, 2, \ldots$.*

(b) *Recall that X has the uniform distribution on the Cantor set C, as specified in Example 1.2.42. Verify that $x \mapsto F_X(x)$ is everywhere continuous, hence the law $\mathcal{P}_X$ has no atoms (i.e. points of positive probability).*

3.3.2. Inversion, continuity and convergence. Is it possible to recover the distribution function from the characteristic function? Then answer is essentially yes.

THEOREM 3.3.12 (LÉVY'S INVERSION THEOREM). *Suppose Φ_X is the characteristic function of random variable X whose distribution function is F_X. For any real numbers $a < b$ and θ, let*

$$\psi_{a,b}(\theta) = \frac{1}{2\pi}\int_a^b e^{-i\theta u} du = \frac{e^{-i\theta a} - e^{-i\theta b}}{i2\pi\theta}\,. \tag{3.3.5}$$

Then,

$$\lim_{T\uparrow\infty} \int_{-T}^{T} \psi_{a,b}(\theta)\Phi_X(\theta)d\theta = \frac{1}{2}[F_X(b) + F_X(b^-)] - \frac{1}{2}[F_X(a) + F_X(a^-)]\,. \tag{3.3.6}$$

Furthermore, if $\int_{\mathbb{R}} |\Phi_X(\theta)|d\theta < \infty$, then X has the bounded continuous probability density function

$$f_X(x) = \frac{1}{2\pi}\int_{\mathbb{R}} e^{-i\theta x}\Phi_X(\theta)d\theta\,. \tag{3.3.7}$$

REMARK. The identity (3.3.7) is a special case of the Fourier transform inversion formula, and as such is in 'duality' with $\Phi_X(\theta) = \int_{\mathbb{R}} e^{i\theta x} f_X(x)dx$ of (3.3.4). The formula (3.3.6) should be considered its integrated version, which thereby holds even in the absence of a density for X.

Here is a simple application of the 'duality' between (3.3.7) and (3.3.4).

EXAMPLE 3.3.13. *The* Cauchy density *is $f_X(x) = 1/[\pi(1 + x^2)]$. Recall Example 3.3.7 that the density $f_S(s) = 0.5e^{-|s|}$ has the positive, integrable characteristic function $1/(1 + \theta^2)$. Thus, by (3.3.7),*

$$0.5e^{-|s|} = \frac{1}{2\pi}\int_{\mathbb{R}} \frac{1}{1+t^2} e^{-its} dt\,.$$

Multiplying both sides by two, then changing t to x and s to $-\theta$, we get (3.3.4) for the Cauchy density, resulting with its characteristic function $\Phi_X(\theta) = e^{-|\theta|}$.

When using characteristic functions for proving limit theorems we do not need the explicit formulas of Lévy's inversion theorem, but rather only the fact that the characteristic function determines the law, that is:

COROLLARY 3.3.14. *If the characteristic functions of two random variables X and Y are the same, that is $\Phi_X(\theta) = \Phi_Y(\theta)$ for all θ, then $X \overset{\mathcal{D}}{=} Y$.*

REMARK. While the real-valued *moment generating function* $M_X(s) = \mathbf{E}[e^{sX}]$ is perhaps a simpler object than the characteristic function, it has a somewhat limited scope of applicability. For example, the law of a random variable X is uniquely determined by $M_X(\cdot)$ provided $M_X(s)$ is finite for all $s \in [-\delta, \delta]$, some $\delta > 0$ (c.f. [**Bil95**, Theorem 30.1]). More generally, assuming all moments of X are finite, the *Hamburger moment problem* is about uniquely determining the law of X from a given sequence of moments $\mathbf{E}X^k$. You saw in Exercise 3.2.29 that this is always possible when X has bounded support, but unfortunately, this is not always the case when X has unbounded support. For more on this issue, see [**Dur03**, Section 2.3e].

PROOF OF COROLLARY 3.3.14. Since $\Phi_X = \Phi_Y$, comparing the right side of (3.3.6) for X and Y shows that

$$[F_X(b) + F_X(b^-)] - [F_X(a) + F_X(a^-)] = [F_Y(b) + F_Y(b^-)] - [F_Y(a) + F_Y(a^-)] \,.$$

As F_X is a distribution function, both $F_X(a) \to 0$ and $F_X(a^-) \to 0$ when $a \downarrow -\infty$. For this reason also $F_Y(a) \to 0$ and $F_Y(a^-) \to 0$. Consequently,

$$F_X(b) + F_X(b^-) = F_Y(b) + F_Y(b^-) \qquad \text{for all} \quad b \in \mathbb{R} \,.$$

In particular, this implies that $F_X = F_Y$ on the collection $\mathcal{C}$ of continuity points of both F_X and F_Y. Recall that F_X and F_Y have each at most a countable set of points of discontinuity (see Exercise 1.2.38), so the complement of $\mathcal{C}$ is countable, and consequently $\mathcal{C}$ is a dense subset of $\mathbb{R}$. Thus, as distribution functions are non-decreasing and right-continuous we know that $F_X(b) = \inf\{F_X(x) : x > b, x \in \mathcal{C}\}$ and $F_Y(b) = \inf\{F_Y(x) : x > b, x \in \mathcal{C}\}$. Since $F_X(x) = F_Y(x)$ for all $x \in \mathcal{C}$, this identity extends to all $b \in \mathbb{R}$, resulting with $X \overset{\mathcal{D}}{=} Y$. □

REMARK. In Lemma 3.1.1, it was shown directly that the sum of independent random variables of normal distributions $\mathcal{N}(\mu_k, v_k)$ has the normal distribution $\mathcal{N}(\mu, v)$ where $\mu = \sum_k \mu_k$ and $v = \sum_k v_k$. The proof easily reduces to dealing with two independent random variables, X of distribution $\mathcal{N}(\mu_1, v_1)$ and Y of distribution $\mathcal{N}(\mu_2, v_2)$ and showing that $X+Y$ has the normal distribution $\mathcal{N}(\mu_1 + \mu_2, v_1 + v_2)$. Here is an easy proof of this result via characteristic functions. First by the independence of X and Y (see Lemma 3.3.8), and their normality (see Example 3.3.6),

$$\begin{aligned}\Phi_{X+Y}(\theta) &= \Phi_X(\theta)\Phi_Y(\theta) = \exp(i\mu_1\theta - v_1\theta^2/2)\exp(i\mu_2\theta - v_2\theta^2/2) \\ &= \exp(i(\mu_1 + \mu_2)\theta - \frac{1}{2}(v_1 + v_2)\theta^2)\end{aligned}$$

We recognize this expression as the characteristic function corresponding to the $\mathcal{N}(\mu_1 + \mu_2, v_1 + v_2)$ distribution, which by Corollary 3.3.14 must indeed be the distribution of $X + Y$.

PROOF OF LÉVY'S INVERSION THEOREM. Consider the product μ of the law $\mathcal{P}_X$ of X which is a probability measure on $\mathbb{R}$ and Lebesgue's measure of $\theta \in [-T, T]$, noting that μ is a finite measure on $\mathbb{R} \times [-T, T]$ of total mass $2T$.

Fixing $a < b \in \mathbb{R}$ let $h_{a,b}(x, \theta) = \psi_{a,b}(\theta)e^{i\theta x}$, where by (3.3.5) and Jensen's inequality for the modulus function (and the uniform measure on $[a, b]$),

$$|h_{a,b}(x, \theta)| = |\psi_{a,b}(\theta)| \le \frac{1}{2\pi}\int_a^b |e^{-i\theta u}|du = \frac{b-a}{2\pi} \,.$$

Consequently, $\int |h_{a,b}|d\mu < \infty$, and applying Fubini's theorem, we conclude that

$$\begin{aligned} J_T(a,b) &:= \int_{-T}^{T} \psi_{a,b}(\theta)\Phi_X(\theta)d\theta = \int_{-T}^{T} \psi_{a,b}(\theta)\Big[\int_{\mathbb{R}} e^{i\theta x} d\mathcal{P}_X(x)\Big] d\theta \\ &= \int_{-T}^{T} \Big[\int_{\mathbb{R}} h_{a,b}(x,\theta) d\mathcal{P}_X(x)\Big] d\theta = \int_{\mathbb{R}} \Big[\int_{-T}^{T} h_{a,b}(x,\theta)d\theta\Big] d\mathcal{P}_X(x) \,. \end{aligned}$$

Since $h_{a,b}(x,\theta)$ is the difference between the function $e^{i\theta u}/(i2\pi\theta)$ at $u = x - a$ and the same function at $u = x - b$, it follows that

$$\int_{-T}^{T} h_{a,b}(x,\theta)d\theta = R(x-a,T) - R(x-b,T) \,.$$

Further, as the cosine function is even and the sine function is odd,

$$R(u,T) = \int_{-T}^{T} \frac{e^{i\theta u}}{i2\pi\theta} d\theta = \int_0^T \frac{\sin(\theta u)}{\pi\theta} d\theta = \frac{\operatorname{sgn}(u)}{\pi} S(|u|T) \,,$$

with $S(r) = \int_0^r x^{-1} \sin x \, dx$ for $r > 0$.

Even though the Lebesgue integral $\int_0^\infty x^{-1} \sin x \, dx$ does not exist, because both the integral of the positive part and the integral of the negative part are infinite, we still have that $S(r)$ is uniformly bounded on $(0,\infty)$ and

$$\lim_{r\uparrow\infty} S(r) = \frac{\pi}{2}$$

(c.f. Exercise 3.3.15). Consequently,

$$\lim_{T\uparrow\infty} [R(x-a,T) - R(x-b,T)] = g_{a,b}(x) = \begin{cases} 0 & \text{if} \;\; x < a \;\; \text{or} \;\; x > b \\ \frac{1}{2} & \text{if} \;\; x = a \;\; \text{or} \;\; x = b \\ 1 & \text{if} \;\; a < x < b \end{cases} .$$

Since $S(\cdot)$ is uniformly bounded, so is $|R(x-a,T) - R(x-b,T)|$ and by bounded convergence,

$$\begin{aligned} \lim_{T\uparrow\infty} J_T(a,b) &= \lim_{T\uparrow\infty} \int_{\mathbb{R}} [R(x-a,T) - R(x-b,T)] d\mathcal{P}_X(x) = \int_{\mathbb{R}} g_{a,b}(x) d\mathcal{P}_X(x) \\ &= \frac{1}{2}\mathcal{P}_X(\{a\}) + \mathcal{P}_X((a,b)) + \frac{1}{2}\mathcal{P}_X(\{b\}) \,. \end{aligned}$$

With $\mathcal{P}_X(\{a\}) = F_X(a) - F_X(a^-)$, $\mathcal{P}_X((a,b)) = F_X(b^-) - F_X(a)$ and $\mathcal{P}_X(\{b\}) = F_X(b) - F_X(b^-)$, we arrive at the assertion (3.3.6).

Suppose now that $\int_{\mathbb{R}} |\Phi_X(\theta)|d\theta = C < \infty$. This implies that both the real and the imaginary parts of $e^{i\theta x}\Phi_X(\theta)$ are integrable with respect to Lebesgue's measure on $\mathbb{R}$, hence $f_X(x)$ of (3.3.7) is well defined. Further, $|f_X(x)| \le C$ is uniformly bounded and by dominated convergence with respect to Lebesgue's measure on $\mathbb{R}$,

$$\lim_{h\to 0} |f_X(x+h) - f_X(x)| \le \lim_{h\to 0} \frac{1}{2\pi} \int_{\mathbb{R}} |e^{-i\theta x}||\Phi_X(\theta)||e^{-i\theta h} - 1|d\theta = 0,$$

implying that $f_X(\cdot)$ is also continuous. Turning to prove that $f_X(\cdot)$ is the density of X, note that

$$|\psi_{a,b}(\theta)\Phi_X(\theta)| \le \frac{b-a}{2\pi}|\Phi_X(\theta)| \,,$$

so by dominated convergence we have that

$$\lim_{T\uparrow\infty} J_T(a,b) = J_\infty(a,b) = \int_{\mathbb{R}} \psi_{a,b}(\theta)\Phi_X(\theta)d\theta\,. \tag{3.3.8}$$

Further, in view of (3.3.5), upon applying Fubini's theorem for the integrable function $e^{-i\theta u}I_{[a,b]}(u)\Phi_X(\theta)$ with respect to Lebesgue's measure on $\mathbb{R}^2$, we see that

$$J_\infty(a,b) = \frac{1}{2\pi}\int_{\mathbb{R}}\Big[\int_a^b e^{-i\theta u}du\Big]\Phi_X(\theta)d\theta = \int_a^b f_X(u)du\,,$$

for the bounded continuous function $f_X(\cdot)$ of (3.3.7). In particular, $J_\infty(a,b)$ must be continuous in both a and b. Comparing (3.3.8) with (3.3.6) we see that

$$J_\infty(a,b) = \frac{1}{2}[F_X(b)+F_X(b^-)] - \frac{1}{2}[F_X(a)+F_X(a^-)]\,,$$

so the continuity of $J_\infty(\cdot,\cdot)$ implies that $F_X(\cdot)$ must also be continuous everywhere, with

$$F_X(b) - F_X(a) = J_\infty(a,b) = \int_a^b f_X(u)du\,,$$

for all $a < b$. This shows that necessarily $f_X(x)$ is a non-negative real-valued function, which is the density of X. □

Exercise 3.3.15. *Integrating $\int z^{-1}e^{iz}dz$ around the contour formed by the "upper" semi-circles of radii ε and r and the intervals $[-r,-\varepsilon]$ and $[r,\varepsilon]$, deduce that $S(r) = \int_0^r x^{-1}\sin x dx$ is uniformly bounded on $(0,\infty)$ with $S(r)\to\pi/2$ as $r\to\infty$.*

Our strategy for handling the CLT and similar limit results is to establish the convergence of characteristic functions and deduce from it the corresponding convergence in distribution. One ingredient for this is of course the fact that the characteristic function uniquely determines the corresponding law. Our next result provides an important second ingredient, that is, an explicit sufficient condition for uniform tightness in terms of the limit of the characteristic functions.

Lemma 3.3.16. *Suppose $\{\nu_n\}$ are probability measures on $(\mathbb{R},\mathcal{B})$ and $\Phi_{\nu_n}(\theta) = \nu_n(e^{i\theta x})$ the corresponding characteristic functions. If $\Phi_{\nu_n}(\theta)\to\Phi(\theta)$ as $n\to\infty$, for each $\theta\in\mathbb{R}$ and further $\Phi(\theta)$ is continuous at $\theta=0$, then the sequence $\{\nu_n\}$ is uniformly tight.*

Remark. To see why continuity of the limit $\Phi(\cdot)$ at 0 is required, consider the sequence ν_n of normal distributions $\mathcal{N}(0,n^2)$. From Example 3.3.6 we see that the point-wise limit $\Phi(\theta) = I_{\theta=0}$ of $\Phi_{\nu_n}(\theta) = \exp(-n^2\theta^2/2)$ exists but is discontinuous at $\theta=0$. However, for any $M<\infty$ we know that $\nu_n([-M,M]) = \nu_1([-M/n,M/n])\to 0$ as $n\to\infty$, so clearly the sequence $\{\nu_n\}$ is not uniformly tight. Indeed, the corresponding distribution functions $F_n(x) = F_1(x/n)$ converge vaguely to $F_\infty(x) = F_1(0) = 1/2$ which is not a distribution function (reflecting escape of all the probability mass to $\pm\infty$).

Proof. We start the proof by deriving the key inequality

$$\frac{1}{r}\int_{-r}^{r}(1-\Phi_\mu(\theta))d\theta \geq \mu([-2/r,2/r]^c)\,, \tag{3.3.9}$$

which holds for every probability measure μ on $(\mathbb{R}, \mathcal{B})$ and any $r > 0$, relating the smoothness of the characteristic function at 0 with the tail decay of the corresponding probability measure at $\pm\infty$. To this end, fixing $r > 0$, note that

$$J(x) := \int_{-r}^{r} (1 - e^{i\theta x}) d\theta = 2r - \int_{-r}^{r} (\cos\theta x + i \sin\theta x) d\theta = 2r - \frac{2\sin rx}{x}\,.$$

So $J(x)$ is non-negative (since $|\sin u| \le |u|$ for all u), and bounded below by $2r - 2/|x|$ (since $|\sin u| \le 1$). Consequently,

$$J(x) \ge \max(2r - \frac{2}{|x|}, 0) \ge r I_{\{|x|>2/r\}}\,. \tag{3.3.10}$$

Now, applying Fubini's theorem for the function $1 - e^{i\theta x}$ whose modulus is bounded by 2 and the product of the probability measure μ and Lebesgue's measure on $[-r, r]$, which is a finite measure of total mass $2r$, we get the identity

$$\int_{-r}^{r} (1 - \Phi_\mu(\theta)) d\theta = \int_{-r}^{r} \Big[\int_{\mathbb{R}} (1 - e^{i\theta x}) d\mu(x) \Big] d\theta = \int_{\mathbb{R}} J(x) d\mu(x)\,.$$

Thus, the lower bound (3.3.10) and monotonicity of the integral imply that

$$\frac{1}{r}\int_{-r}^{r} (1 - \Phi_\mu(\theta)) d\theta = \frac{1}{r}\int_{\mathbb{R}} J(x) d\mu(x) \ge \int_{\mathbb{R}} I_{\{|x|>2/r\}} d\mu(x) = \mu([-2/r, 2/r]^c)\,,$$

hence establishing (3.3.9).

We turn to the application of this inequality for proving the uniform tightness. Since $\Phi_{\nu_n}(0) = 1$ for all n and $\Phi_{\nu_n}(0) \to \Phi(0)$, it follows that $\Phi(0) = 1$. Further, $\Phi(\theta)$ is continuous at $\theta = 0$, so for any $\varepsilon > 0$, there exists $r = r(\varepsilon) > 0$ such that

$$\frac{\varepsilon}{4} \ge |1 - \Phi(\theta)| \quad \text{for all} \quad \theta \in [-r, r],$$

and hence also

$$\frac{\varepsilon}{2} \ge \frac{1}{r}\int_{-r}^{r} |1 - \Phi(\theta)| d\theta\,.$$

The point-wise convergence of Φ_{ν_n} to Φ implies that $|1 - \Phi_{\nu_n}(\theta)| \to |1 - \Phi(\theta)|$. By bounded convergence with respect to Uniform measure of θ on $[-r, r]$, it follows that for some finite $n_0 = n_0(\varepsilon)$ and all $n \ge n_0$,

$$\varepsilon \ge \frac{1}{r}\int_{-r}^{r} |1 - \Phi_{\nu_n}(\theta)| d\theta\,,$$

which in view of (3.3.9) results with

$$\varepsilon \ge \frac{1}{r}\int_{-r}^{r} [1 - \Phi_{\nu_n}(\theta)] d\theta \ge \nu_n([-2/r, 2/r]^c)\,.$$

Since $\varepsilon > 0$ is arbitrary and $M = 2/r$ is independent of n, by Definition 3.2.32 this amounts to the uniform tightness of the sequence $\{\nu_n\}$. □

Building upon Corollary 3.3.14 and Lemma 3.3.16 we can finally relate the point-wise convergence of characteristic functions to the weak convergence of the corresponding measures.

Theorem 3.3.17 (Lévy's continuity theorem). *Let ν_n, $1 \le n \le \infty$ be probability measures on $(\mathbb{R}, \mathcal{B})$.*

(a) *If $\nu_n \overset{w}{\Rightarrow} \nu_\infty$, then $\Phi_{\nu_n}(\theta) \to \Phi_{\nu_\infty}(\theta)$ for each $\theta \in \mathbb{R}$.*

(b) *Conversely, if $\Phi_{\nu_n}(\theta)$ converges point-wise to a limit $\Phi(\theta)$ that is continuous at $\theta = 0$, then $\{\nu_n\}$ is a uniformly tight sequence and $\nu_n \overset{w}{\Rightarrow} \nu$ such that $\Phi_\nu = \Phi$.*

PROOF. For part (a), since both $x \mapsto \cos(\theta x)$ and $x \mapsto \sin(\theta x)$ are bounded continuous functions, the assumed weak convergence of ν_n to ν_∞ implies that $\Phi_{\nu_n}(\theta) = \nu_n(e^{i\theta x}) \to \nu_\infty(e^{i\theta x}) = \Phi_{\nu_\infty}(\theta)$ (c.f. Definition 3.2.17).
Turning to deal with part (b), recall that by Lemma 3.3.16 we know that the collection $\Gamma = \{\nu_n\}$ is uniformly tight. Hence, by Prohorov's theorem (see the remark preceding the proof of Lemma 3.2.38), for every subsequence $\nu_{n(m)}$ there is a further sub-subsequence $\nu_{n(m_k)}$ that converges weakly to some probability measure ν_∞. Though in general ν_∞ might depend on the specific choice of $n(m)$, we deduce from part (a) of the theorem that necessarily $\Phi_{\nu_\infty} = \Phi$. Since the characteristic function uniquely determines the law (see Corollary 3.3.14), here the same limit $\nu = \nu_\infty$ applies for *all choices* of $n(m)$. In particular, fixing $h \in C_b(\mathbb{R})$, the sequence $y_n = \nu_n(h)$ is such that every subsequence $y_{n(m)}$ has a further sub-subsequence $y_{n(m_k)}$ that converges to $y = \nu(h)$. Consequently, $y_n = \nu_n(h) \to y = \nu(h)$ (see Lemma 2.2.11), and since this applies for all $h \in C_b(\mathbb{R})$, we conclude that $\nu_n \overset{w}{\Rightarrow} \nu$ such that $\Phi_\nu = \Phi$. □

Here is a direct consequence of Lévy's continuity theorem.

EXERCISE 3.3.18. *Show that if $X_n \xrightarrow{\mathcal{D}} X_\infty$, $Y_n \xrightarrow{\mathcal{D}} Y_\infty$ and Y_n is independent of X_n for $1 \le n \le \infty$, then $X_n + Y_n \xrightarrow{\mathcal{D}} X_\infty + Y_\infty$.*

Combining Exercise 3.3.18 with the Portmanteau theorem and the CLT, you can now show that a finite second moment is necessary for the convergence in distribution of $n^{-1/2}\sum_{k=1}^n X_k$ for i.i.d. $\{X_k\}$.

EXERCISE 3.3.19. *Suppose $\{X_k, \widetilde{X}_k\}$ are i.i.d. and $n^{-1/2}\sum_{k=1}^n X_k \xrightarrow{\mathcal{D}} Z$ (with the limit $Z \in \mathbb{R}$).*

(a) *Set $Y_k = X_k - \widetilde{X}_k$ and show that $n^{-1/2}\sum_{k=1}^n Y_k \xrightarrow{\mathcal{D}} Z - \widetilde{Z}$, with Z and $\widetilde{Z}$ i.i.d.*

(b) *Let $U_k = Y_k I_{|Y_k| \le b}$ and $V_k = Y_k I_{|Y_k|>b}$. Show that for any $u < \infty$ and all n,*

$$\mathbf{P}(\sum_{k=1}^n Y_k \ge u\sqrt{n}) \ge \mathbf{P}(\sum_{k=1}^n U_k \ge u\sqrt{n}, \sum_{k=1}^n V_k \ge 0) \ge \frac{1}{2}\mathbf{P}(\sum_{k=1}^n U_k \ge u\sqrt{n})\,.$$

(c) *Apply the Portmanteau theorem and the CLT for the bounded i.i.d. $\{U_k\}$ to get that for any $u, b < \infty$,*

$$\mathbf{P}(Z - \widetilde{Z} \ge u) \ge \frac{1}{2}\mathbf{P}(G \ge u/\sqrt{\mathbf{E}U_1^2})\,.$$

Considering the limit $b \to \infty$ followed by $u \to \infty$ deduce that $\mathbf{E}Y_1^2 < \infty$.

(d) *Conclude that if $n^{-1/2}\sum_{k=1}^n X_k \xrightarrow{\mathcal{D}} Z$, then necessarily $\mathbf{E}X_1^2 < \infty$.*

REMARK. The trick of replacing X_k by the variables $Y_k = X_k - \widetilde{X}_k$ whose law is symmetric (i.e. $Y_k \overset{\mathcal{D}}{=} -Y_k$), is very useful in many problems. It is often called the *symmetrization trick*.

EXERCISE 3.3.20. *Provide an example of a random variable X with a bounded probability density function but for which $\int_{\mathbb{R}} |\Phi_X(\theta)| d\theta = \infty$, and another example of a random variable X whose characteristic function $\Phi_X(\theta)$ is not differentiable at $\theta = 0$.*

As you find out next, Lévy's inversion theorem can help when computing densities.

EXERCISE 3.3.21. *Suppose the random variables U_k are i.i.d. where the law of each U_k is the* uniform probability measure *on $(-1, 1)$. Considering Example 3.3.7, show that for each $n \geq 2$, the probability density function of $S_n = \sum_{k=1}^n U_k$ is*

$$f_{S_n}(s) = \frac{1}{\pi} \int_0^\infty \cos(\theta s)(\sin\theta/\theta)^n d\theta \,,$$

and deduce that $\int_0^\infty \cos(\theta s)(\sin\theta/\theta)^n d\theta = 0$ for all $s > n \geq 2$.

EXERCISE 3.3.22. *Deduce from Example 3.3.13 that if $\{X_k\}$ are i.i.d. each having the Cauchy density, then $n^{-1}\sum_{k=1}^n X_k$ has the same distribution as X_1, for any value of n.*

We next relate differentiability of $\Phi_X(\cdot)$ with the weak law of large numbers and show that it does not imply that $\mathbf{E}|X|$ is finite.

EXERCISE 3.3.23. *Let $S_n = \sum_{k=1}^n X_k$ where the i.i.d. random variables $\{X_k\}$ have each the characteristic function $\Phi_X(\cdot)$.*

(a) *Show that if $\frac{d\Phi_X}{d\theta}(0) = z \in \mathbb{C}$, then $z = ia$ for some $a \in \mathbb{R}$ and $n^{-1}S_n \xrightarrow{p} a$ as $n \to \infty$.*

(b) *Show that if $n^{-1}S_n \xrightarrow{p} a$, then $\Phi_X(\pm h_k)^{n_k} \to e^{\pm ia}$ for any $h_k \downarrow 0$ and $n_k = [1/h_k]$, and deduce that $\frac{d\Phi_X}{d\theta}(0) = ia$.*

(c) *Conclude that the weak law of large numbers holds (i.e. $n^{-1}S_n \xrightarrow{p} a$ for some non-random a), if and only if $\Phi_X(\cdot)$ is differentiable at $\theta = 0$ (this result is due to E.J.G. Pitman, see* [**Pit56**]*).*

(d) *Use Exercise 2.1.13 to provide a random variable X for which $\Phi_X(\cdot)$ is differentiable at $\theta = 0$ but $\mathbf{E}|X| = \infty$.*

As you show next, $X_n \xrightarrow{\mathcal{D}} X_\infty$ yields convergence of $\Phi_{X_n}(\cdot)$ to $\Phi_{X_\infty}(\cdot)$, uniformly over compact subsets of $\mathbb{R}$.

EXERCISE 3.3.24. *Show that if $X_n \xrightarrow{\mathcal{D}} X_\infty$ then for any r finite,*

$$\lim_{n\to\infty} \sup_{|\theta|\leq r} |\Phi_{X_n}(\theta) - \Phi_{X_\infty}(\theta)| = 0 \,.$$

Hint: *By Theorem 3.2.7 you may further assume that $X_n \xrightarrow{a.s.} X_\infty$.*

Characteristic functions of modulus one correspond to lattice or degenerate laws, as you show in the following refinement of part (c) of Proposition 3.3.2.

EXERCISE 3.3.25. *Suppose $|\Phi_Y(\theta)| = 1$ for some $\theta \neq 0$.*

(a) *Show that Y is a* $(2\pi/\theta)$-lattice random variable, *namely, that $Y \bmod (2\pi/\theta)$ is $\mathbf{P}$-degenerate.*
Hint: *Check conditions for equality when applying Jensen's inequality for $(\cos\theta Y, \sin\theta Y)$ and the convex function $g(x, y) = \sqrt{x^2 + y^2}$.*

(b) *Deduce that if in addition $|\Phi_Y(\lambda\theta)| = 1$ for some $\lambda \notin \mathcal{Q}$ then Y must be $\mathbf{P}$-degenerate, in which case $\Phi_Y(\theta) = \exp(i\theta c)$ for some $c \in \mathbb{R}$.*

Building on the preceding two exercises, you are to prove next the following *convergence of types* result.

Exercise 3.3.26. *Suppose $Z_n \xrightarrow{\mathcal{D}} Y$ and $\beta_n Z_n + \gamma_n \xrightarrow{\mathcal{D}} \widehat{Y}$ for some $\widehat{Y}$,* non-**P**-*degenerate Y, and non-random $\beta_n \geq 0$, γ_n.*

(a) *Show that $\beta_n \to \beta \geq 0$ finite.*
Hint: *Start with the finiteness of limit points of $\{\beta_n\}$.*

(b) *Deduce that $\gamma_n \to \gamma$ finite.*

(c) *Conclude that $\widehat{Y} \overset{\mathcal{D}}{=} \beta Y + \gamma$.*
Hint: *Recall Slutsky's lemma.*

Remark. This convergence of types fails for **P**-degenerate Y. For example, if $Z_n \overset{\mathcal{D}}{=} \mathcal{N}(0, n^{-3})$, then both $Z_n \xrightarrow{\mathcal{D}} 0$ and $nZ_n \xrightarrow{\mathcal{D}} 0$. Similarly, if $Z_n \overset{\mathcal{D}}{=} \mathcal{N}(0,1)$ then $\beta_n Z_n \overset{\mathcal{D}}{=} \mathcal{N}(0,1)$ for the non-converging sequence $\beta_n = (-1)^n$ (of alternating signs).

Mimicking the proof of Lévy's inversion theorem, for random variables of bounded support you get the following alternative inversion formula, based on the theory of Fourier series.

Exercise 3.3.27. *Suppose R.V. X supported on $(0,t)$ has the characteristic function Φ_X and the distribution function F_X. Let $\theta_0 = 2\pi/t$ and $\psi_{a,b}(\cdot)$ be as in (3.3.5), with $\psi_{a,b}(0) = \frac{b-a}{2\pi}$.*

(a) *Show that for any $0 < a < b < t$*

$$\lim_{T\uparrow\infty} \sum_{k=-T}^{T} \theta_0 (1 - \frac{|k|}{T}) \psi_{a,b}(k\theta_0) \Phi_X(k\theta_0) = \frac{1}{2}[F_X(b) + F_X(b^-)] - \frac{1}{2}[F_X(a) + F_X(a^-)] \,.$$

Hint: *Recall that $S_T(r) = \sum_{k=1}^{T} (1 - k/T) \frac{\sin kr}{k}$ is uniformly bounded for $r \in (0, 2\pi)$ and integer $T \geq 1$, and $S_T(r) \to \frac{\pi - r}{2}$ as $T \to \infty$.*

(b) *Show that if $\sum_k |\Phi_X(k\theta_0)| < \infty$ then X has the bounded continuous probability density function, given for $x \in (0,t)$ by*

$$f_X(x) = \frac{\theta_0}{2\pi} \sum_{k\in\mathbb{Z}} e^{-ik\theta_0 x} \Phi_X(k\theta_0) \,.$$

(c) *Deduce that if R.V.s X and Y supported on $(0,t)$ are such that $\Phi_X(k\theta_0) = \Phi_Y(k\theta_0)$ for all $k \in \mathbb{Z}$, then $X \overset{\mathcal{D}}{=} Y$.*

Here is an application of the preceding exercise for the *random walk* on the circle S^1 of radius one (c.f. Definition 5.1.6 for the random walk on $\mathbb{R}$).

Exercise 3.3.28. *Let $t = 2\pi$ and Ω denote the unit circle S^1 parametrized by the angular coordinate to yield the identification $\Omega = [0,t]$ where both end-points are considered the same point. We equip Ω with the topology induced by $[0,t]$ and the surface measure λ_Ω similarly induced by Lebesgue's measure (as in Exercise 1.4.37). In particular, R.V.-s on $(\Omega, \mathcal{B}_\Omega)$ correspond to Borel periodic functions on $\mathbb{R}$, of period t. In this context we call U of law $t^{-1}\lambda_\Omega$ a uniform R.V. and call $S_n = (\sum_{k=1}^n \xi_k) \bmod t$, with i.i.d $\xi, \xi_k \in \Omega$, a random walk.*

(a) *Verify that Exercise 3.3.27 applies for $\theta_0 = 1$ and R.V.-s on Ω.*

(b) *Show that if probability measures ν_n on $(\Omega, \mathcal{B}_\Omega)$ are such that $\Phi_{\nu_n}(k) \to \varphi(k)$ for $n \to \infty$ and fixed $k \in \mathbb{Z}$, then $\nu_n \overset{w}{\Rightarrow} \nu_\infty$ and $\varphi(k) = \Phi_{\nu_\infty}(k)$ for all $k \in \mathbb{Z}$.*
Hint: *Since Ω is compact the sequence $\{\nu_n\}$ is uniformly tight.*

(c) *Show that $\Phi_U(k) = \mathbf{1}_{k=0}$ and $\Phi_{S_n}(k) = \Phi_\xi(k)^n$. Deduce from these facts that if ξ has a density with respect to λ_Ω then $S_n \xrightarrow{\mathcal{D}} U$ as $n \to \infty$.*
Hint: *Recall part (a) of Exercise 3.3.25.*

(d) *Check that if $\xi = \alpha$ is non-random for some $\alpha/t \notin \mathbb{Q}$, then S_n does not converge in distribution, but $S_{K_n} \xrightarrow{\mathcal{D}} U$ for K_n which are uniformly chosen in $\{1, 2, \dots, n\}$, independently of the sequence $\{\xi_k\}$.*

3.3.3. Revisiting the CLT. Applying the theory of Subsection 3.3.2 we provide an alternative proof of the CLT, based on characteristic functions. One can prove many other weak convergence results for sums of random variables by properly adapting this approach, which is exactly what we will do when demonstrating the convergence to stable laws (see Exercise 3.3.33), and in proving the Poisson approximation theorem (in Subsection 3.4.1), and the multivariate CLT (in Section 3.5).

To this end, we start by deriving the analog of the bound (3.1.7) for the characteristic function.

LEMMA 3.3.29. *If a random variable X has $\mathbf{E}(X) = 0$ and $\mathbf{E}(X^2) = v < \infty$, then for all $\theta \in \mathbb{R}$,*

$$\Big|\Phi_X(\theta) - \big(1 - \frac{1}{2}v\theta^2\big)\Big| \le \theta^2 \mathbf{E} \min(|X|^2, |\theta||X|^3/6).$$

PROOF. Let $R_2(x) = e^{ix} - 1 - ix - (ix)^2/2$. Then, rearranging terms, recalling $\mathbf{E}(X) = 0$ and using Jensen's inequality for the modulus function, we see that

$$\Big|\Phi_X(\theta) - \Big(1 - \frac{1}{2}v\theta^2\Big)\Big| = \Big|\mathbf{E}\big[e^{i\theta X} - 1 - i\theta X - \frac{i^2}{2}\theta^2 X^2\big]\Big| = \Big|\mathbf{E}R_2(\theta X)\Big| \le \mathbf{E}|R_2(\theta X)|.$$

Since $|R_2(x)| \le \min(|x|^2, |x|^3/6)$ for any $x \in \mathbb{R}$ (see also Exercise 3.3.34), by monotonicity of the expectation we get that $\mathbf{E}|R_2(\theta X)| \le \mathbf{E}\min(|\theta X|^2, |\theta X|^3/6)$, completing the proof of the lemma. □

The following simple complex analysis estimate is needed for relating the approximation of the characteristic function of summands to that of their sum.

LEMMA 3.3.30. *Suppose $z_{n,k} \in \mathbb{C}$ are such that $z_n = \sum_{k=1}^n z_{n,k} \to z_\infty$ and $\eta_n = \sum_{k=1}^n |z_{n,k}|^2 \to 0$ when $n \to \infty$. Then,*

$$\varphi_n := \prod_{k=1}^n (1 + z_{n,k}) \to \exp(z_\infty) \quad \textit{for} \quad n \to \infty.$$

PROOF. Recall that the power series expansion

$$\log(1+z) = \sum_{k=1}^\infty \frac{(-1)^{k-1} z^k}{k}$$

converges for $|z| < 1$. In particular, for $|z| \leq 1/2$ it follows that

$$|\log(1+z) - z| \leq \sum_{k=2}^{\infty} \frac{|z|^k}{k} \leq |z|^2 \sum_{k=2}^{\infty} \frac{2^{-(k-2)}}{k} \leq |z|^2 \sum_{k=2}^{\infty} 2^{-(k-1)} = |z|^2 .$$

Let $\delta_n = \max\{|z_{n,k}| : k = 1, \ldots, n\}$. Note that $\delta_n^2 \leq \eta_n$, so our assumption that $\eta_n \to 0$ implies that $\delta_n \leq 1/2$ for all n sufficiently large, in which case

$$|\log \varphi_n - z_n| = |\log \prod_{k=1}^{n} (1 + z_{n,k}) - \sum_{k=1}^{n} z_{n,k}| \leq \sum_{k=1}^{n} |\log(1 + z_{n,k}) - z_{n,k}| \leq \eta_n .$$

With $z_n \to z_\infty$ and $\eta_n \to 0$, it follows that $\log \varphi_n \to z_\infty$. Consequently, $\varphi_n \to \exp(z_\infty)$ as claimed. □

We will give now an alternative proof of the CLT of Theorem 3.1.2.

Proof of Theorem 3.1.2. From Example 3.3.6 we know that $\Phi_G(\theta) = e^{-\frac{\theta^2}{2}}$ is the characteristic function of the standard normal distribution. So, by Lévy's continuity theorem it suffices to show that $\Phi_{\widehat{S}_n}(\theta) \to \exp(-\theta^2/2)$ as $n \to \infty$, for each $\theta \in \mathbb{R}$. Recall that $\widehat{S}_n = \sum_{k=1}^n X_{n,k}$, with $X_{n,k} = (X_k - \mu)/\sqrt{vn}$ i.i.d. random variables, so by independence (see Lemma 3.3.8) and scaling (see part (e) of Proposition 3.3.2), we have that

$$\varphi_n := \Phi_{\widehat{S}_n}(\theta) = \prod_{k=1}^{n} \Phi_{X_{n,k}}(\theta) = \Phi_Y(n^{-1/2}\theta)^n = (1 + z_n/n)^n,$$

where $Y = (X_1 - \mu)/\sqrt{v}$ and $z_n = z_n(\theta) := n[\Phi_Y(n^{-1/2}\theta) - 1]$. Applying Lemma 3.3.30 for $z_{n,k} = z_n/n$ it remains only to show that $z_n \to -\theta^2/2$ (for then $\eta_n = |z_n|^2/n \to 0$). Indeed, since $\mathbf{E}(Y) = 0$ and $\mathbf{E}(Y^2) = 1$, we have from Lemma 3.3.29 that

$$|z_n + \theta^2/2| = \left|n[\Phi_Y(n^{-1/2}\theta) - 1] + \theta^2/2\right| \leq \mathbf{E}V_n ,$$

for $V_n = \min(|\theta Y|^2, n^{-1/2}|\theta Y|^3/6)$. With $V_n \overset{a.s.}{\to} 0$ as $n \to \infty$ and $V_n \leq |\theta|^2|Y|^2$ which is integrable, it follows by dominated convergence that $\mathbf{E}V_n \to 0$ as $n \to \infty$, hence $z_n \to -\theta^2/2$ completing the proof of Theorem 3.1.2. □

We proceed with a brief introduction of stable laws, their domain of attraction and the corresponding limit theorems (which are a natural generalization of the CLT).

Definition 3.3.31. *Random variable Y has a* stable law *if it is non-degenerate and for any $m \geq 1$ there exist constants $d_m > 0$ and c_m, such that $Y_1 + \ldots + Y_m \overset{\mathcal{D}}{=} d_m Y + c_m$, where $\{Y_i\}$ are i.i.d. copies of Y. Such variable has a* symmetric stable law *if in addition $Y \overset{\mathcal{D}}{=} -Y$. We further say that random variable X is in the* domain of attraction *of non-degenerate Y if there exist constants $b_n > 0$ and a_n such that $Z_n(X) = (S_n - a_n)/b_n \xrightarrow{\mathcal{D}} Y$ for $S_n = \sum_{k=1}^n X_k$ and i.i.d. copies X_k of X.*

By definition, the collection of stable laws is closed under the affine map $Y \mapsto \pm\sqrt{v}Y + \mu$ for $\mu \in \mathbb{R}$ and $v > 0$ (which correspond to the centering and scale of the law, but not necessarily its mean and variance). Clearly, each stable law is in its own domain of attraction and as we see next, only stable laws have a non-empty domain of attraction.

Proposition 3.3.32. *If X is in the domain of attraction of some non-degenerate variable Y, then Y must have a stable law.*

Proof. Fix $m \geq 1$, and setting $n = km$ let $\beta_n = b_n/b_k > 0$ and $\gamma_n = (a_n - ma_k)/b_k$. We then have the representation

$$\beta_n Z_n(X) + \gamma_n = \sum_{i=1}^{m} Z_k^{(i)} ,$$

where $Z_k^{(i)} = (X_{(i-1)k+1} + \ldots + X_{ik} - a_k)/b_k$ are i.i.d. copies of $Z_k(X)$. From our assumption that $Z_k(X) \xrightarrow{\mathcal{D}} Y$ we thus deduce (by at most $m-1$ applications of Exercise 3.3.18), that $\beta_n Z_n(X)+\gamma_n \xrightarrow{\mathcal{D}} \widehat{Y}$, where $\widehat{Y} = Y_1+\ldots+Y_m$ for i.i.d. copies $\{Y_i\}$ of Y. Moreover, by assumption $Z_n(X) \xrightarrow{\mathcal{D}} Y$, hence by the convergence of types $\widehat{Y} \overset{\mathcal{D}}{=} d_m Y + c_m$ for some finite non-random $d_m \geq 0$ and c_m (c.f. Exercise 3.3.26). Recall Lemma 3.3.8 that $\Phi_{\widehat{Y}}(\theta) = [\Phi_Y(\theta)]^m$. So, with Y assumed non-degenerate the same applies to $\widehat{Y}$ (see Exercise 3.3.25), and in particular $d_m > 0$. Since this holds for any $m \geq 1$, by definition Y has a stable law. □

We have already seen two examples of symmetric stable laws, namely those associated with the zero-mean normal density and with the *Cauchy* density of Example 3.3.13. Indeed, as you show next, for each $\alpha \in (0,2)$ there corresponds the symmetric α-stable variable Y_α whose characteristic function is $\Phi_{Y_\alpha}(\theta) = \exp(-|\theta|^\alpha)$ (so the Cauchy distribution corresponds to the symmetric stable of *index* $\alpha = 1$ and the normal distribution corresponds to index $\alpha = 2$).

Exercise 3.3.33. *Fixing $\alpha \in (0,2)$, suppose $X \overset{\mathcal{D}}{=} -X$ and $\mathbf{P}(|X| > x) = x^{-\alpha}$ for all $x \geq 1$.*

(a) *Check that $\Phi_X(\theta) = 1-\gamma(|\theta|)|\theta|^\alpha$ where $\gamma(r) = \alpha\int_r^\infty (1-\cos u)u^{-(\alpha+1)}du$ converges as $r \downarrow 0$ to $\gamma(0)$ finite and positive.*

(b) *Setting $\varphi_{\alpha,0}(\theta) = \exp(-|\theta|^\alpha)$, $b_n = (\gamma(0)n)^{1/\alpha}$ and $\widehat{S}_n = b_n^{-1}\sum_{k=1}^n X_k$ for i.i.d. copies X_k of X, deduce that $\Phi_{\widehat{S}_n}(\theta) \to \varphi_{\alpha,0}(\theta)$ as $n \to \infty$, for any fixed $\theta \in \mathbb{R}$.*

(c) *Conclude that X is in the domain of attraction of a symmetric stable variable Y_α, whose characteristic function is $\varphi_{\alpha,0}(\cdot)$.*

(d) *Fix $\alpha = 1$ and show that with probability one $\limsup_{n\to\infty} \widehat{S}_n = \infty$ and $\liminf_{n\to\infty} \widehat{S}_n = -\infty$.*
Hint: *Recall Kolmogorov's 0-1 law. The same proof applies for any $\alpha > 0$ once we verify that Y_α has unbounded support.*

(e) *Show that if $\alpha = 1$ then $\frac{1}{n\log n}\sum_{k=1}^n |X_k| \to 1$ in probability but not almost surely (in contrast, X is integrable when $\alpha > 1$, in which case the strong law of large numbers applies).*

Remark. While outside the scope of these notes, one can show that (up to scaling) *any* symmetric stable variable must be of the form Y_α for some $\alpha \in (0,2]$. Further, for any $\alpha \in (0,2)$ the necessary and sufficient condition for $X \overset{\mathcal{D}}{=} -X$ to be in the domain of attraction of Y_α is that the function $L(x) = x^\alpha \mathbf{P}(|X| > x)$ is *slowly varying* at ∞ (that is, $L(ux)/L(x) \to 1$ for $x \to \infty$ and fixed $u > 0$). Indeed, as shown for example in [**Bre92**, Theorem 9.32], up to the mapping $Y \mapsto \sqrt{v}Y + \mu$, the collection of all stable laws forms a two parameter family $Y_{\alpha,\kappa}$, parametrized by the index $\alpha \in (0,2]$ and *skewness* $\kappa \in [-1,1]$. The corresponding characteristic functions are

$$\varphi_{\alpha,\kappa}(\theta) = \exp(-|\theta|^\alpha(1 + i\kappa\text{sgn}(\theta)g_\alpha(\theta)))\,, \tag{3.3.11}$$

where $g_1(r) = (2/\pi)\log|r|$ and $g_\alpha = \tan(\pi\alpha/2)$ is constant for all $\alpha \neq 1$ (in particular, $g_2 = 0$ so the parameter κ is irrelevant when $\alpha = 2$). Further, in case $\alpha < 2$ the domain of attraction of $Y_{\alpha,\kappa}$ consists precisely of the random variables X for which $L(x) = x^\alpha \mathbf{P}(|X| > x)$ is slowly varying at ∞ and $(\mathbf{P}(X > x) - \mathbf{P}(X < -x))/\mathbf{P}(|X| > x) \to \kappa$ as $x \to \infty$ (for example, see [**Bre92**, Theorem 9.34]). To complete this picture, we recall [**Fel71**, Theorem XVII.5.1], that X is in the domain of attraction of the normal variable Y_2 if and only if $L(x) = \mathbf{E}[X^2 I_{|X|\le x}]$ is slowly varying (as is of course the case whenever $\mathbf{E}X^2$ is finite).

As shown in the following exercise, controlling the modulus of the remainder term for the n-th order Taylor approximation of e^{ix} one can generalize the bound on $\Phi_X(\theta)$ beyond the case $n = 2$ of Lemma 3.3.29.

EXERCISE 3.3.34. *For any $x \in \mathbb{R}$ and non-negative integer n, let*

$$R_n(x) = e^{ix} - \sum_{k=0}^{n} \frac{(ix)^k}{k!}\,.$$

(a) *Show that $R_n(x) = \int_0^x iR_{n-1}(y)dy$ for all $n \ge 1$ and deduce by induction on n that*

$$|R_n(x)| \le \min\Big(\frac{2|x|^n}{n!}, \frac{|x|^{n+1}}{(n+1)!}\Big) \quad \textit{for all} \quad x \in \mathbb{R}, n = 0,1,2,\ldots.$$

(b) *Conclude that if $\mathbf{E}|X|^n < \infty$ then*

$$\big|\Phi_X(\theta) - \sum_{k=0}^{n} \frac{(i\theta)^k \mathbf{E}X^k}{k!}\big| \le |\theta|^n \mathbf{E}\Big[\min\big(\frac{2|X|^n}{n!}, \frac{|\theta||X|^{n+1}}{(n+1)!}\big)\Big]\,.$$

By solving the next exercise you generalize the proof of Theorem 3.1.2 via characteristic functions to the setting of Lindeberg's CLT.

EXERCISE 3.3.35. *Consider $\widehat{S}_n = \sum_{k=1}^n X_{n,k}$ for mutually independent random variables $X_{n,k}$, $k = 1,\ldots,n$, of zero mean and variance $v_{n,k}$, such that $v_n = \sum_{k=1}^n v_{n,k} \to 1$ as $n \to \infty$.*

(a) *Fixing $\theta \in \mathbb{R}$ show that*

$$\varphi_n = \Phi_{\widehat{S}_n}(\theta) = \prod_{k=1}^{n}(1 + z_{n,k})\,,$$

where $z_{n,k} = \Phi_{X_{n,k}}(\theta) - 1$.

(b) *With $z_\infty = -\theta^2/2$, use Lemma 3.3.29 to verify that $|z_{n,k}| \le 2\theta^2 v_{n,k}$ and further, for any $\varepsilon > 0$,*

$$|z_n - v_n z_\infty| \le \sum_{k=1}^{n} |z_{n,k} - v_{n,k} z_\infty| \le \theta^2 g_n(\varepsilon) + \frac{|\theta|^3}{6}\varepsilon v_n\,,$$

where $z_n = \sum_{k=1}^n z_{n,k}$ and $g_n(\varepsilon)$ is given by (3.1.4).

(c) *Recall that Lindeberg's condition $g_n(\varepsilon) \to 0$ implies that $r_n^2 = \max_k v_{n,k} \to 0$ as $n \to \infty$. Deduce that then $z_n \to z_\infty$ and $\eta_n = \sum_{k=1}^n |z_{n,k}|^2 \to 0$ when $n \to \infty$.*

(d) *Applying Lemma 3.3.30, conclude that* $\widehat{S}_n \xrightarrow{\mathcal{D}} G$.

We conclude this section with an exercise that reviews various techniques one may use for establishing convergence in distribution for sums of independent random variables.

EXERCISE 3.3.36. *Throughout this problem* $S_n = \sum_{k=1}^n X_k$ *for mutually independent random variables* $\{X_k\}$.

(a) *Suppose that* $\mathbf{P}(X_k = k^\alpha) = \mathbf{P}(X_k = -k^\alpha) = 1/(2k^\beta)$ *and* $\mathbf{P}(X_k = 0) = 1 - k^{-\beta}$. *Show that for any fixed* $\alpha \in \mathbb{R}$ *and* $\beta > 1$, *the series* $S_n(\omega)$ *converges almost surely as* $n \to \infty$.

(b) *Consider the setting of part (a) when* $0 \le \beta < 1$ *and* $\gamma = 2\alpha - \beta + 1$ *is positive. Find non-random* b_n *such that* $b_n^{-1} S_n \xrightarrow{\mathcal{D}} Z$ *and* $0 < F_Z(z) < 1$ *for some* $z \in \mathbb{R}$. *Provide also the characteristic function* $\Phi_Z(\theta)$ *of* Z.

(c) *Repeat part (b) in case* $\beta = 1$ *and* $\alpha > 0$ *(see Exercise 3.1.11 for* $\alpha = 0$*).*

(d) *Suppose now that* $\mathbf{P}(X_k = 2k) = \mathbf{P}(X_k = -2k) = 1/(2k^2)$ *and* $\mathbf{P}(X_k = 1) = \mathbf{P}(X_k = -1) = 0.5(1 - k^{-2})$. *Show that* $S_n/\sqrt{n} \xrightarrow{\mathcal{D}} G$.

3.4. Poisson approximation and the Poisson process

Subsection 3.4.1 deals with the Poisson approximation theorem and few of its applications. It leads naturally to the introduction of the Poisson process in Subsection 3.4.2, where we also explore its relation to sums of i.i.d. Exponential variables and to order statistics of i.i.d. uniform random variables.

3.4.1. Poisson approximation.

The Poisson approximation theorem is about the law of the sum S_n of a large number ($= n$) of independent random variables. In contrast to the CLT that also deals with such objects, here all variables are non-negative integer valued and the variance of S_n remains bounded, allowing for the approximation in law of S_n by an integer valued variable. The Poisson distribution results when the number of terms in the sum grows while the probability that each of them is non-zero decays. As such, the Poisson approximation is about counting the number of occurrences among many independent rare events.

THEOREM 3.4.1 (POISSON APPROXIMATION). *Let* $S_n = \sum\limits_{k=1}^{n} Z_{n,k}$, *where for each* n *the random variables* $Z_{n,k}$ *for* $1 \le k \le n$, *are mutually independent, each taking value in the set of nonnegative integers. Suppose that* $p_{n,k} = \mathbf{P}(Z_{n,k} = 1)$ *and* $\varepsilon_{n,k} = \mathbf{P}(Z_{n,k} \ge 2)$ *are such that as* $n \to \infty$,

(a) $\sum\limits_{k=1}^{n} p_{n,k} \to \lambda < \infty$,

(b) $\max\limits_{k=1,\cdots,n} \{p_{n,k}\} \to 0$,

(c) $\sum\limits_{k=1}^{n} \varepsilon_{n,k} \to 0$.

Then, $S_n \xrightarrow{\mathcal{D}} N_\lambda$ *of a Poisson distribution with parameter* λ, *as* $n \to \infty$.

Proof. The first step of the proof is to apply truncation by comparing S_n with

$$\overline{S}_n = \sum_{k=1}^{n} \overline{Z}_{n,k} \,,$$

where $\overline{Z}_{n,k} = Z_{n,k} I_{Z_{n,k} \le 1}$ for $k = 1, \dots, n$. Indeed, observe that,

$$\begin{aligned} \mathbf{P}(\overline{S}_n \neq S_n) &\le \sum_{k=1}^{n} \mathbf{P}(\overline{Z}_{n,k} \neq Z_{n,k}) = \sum_{k=1}^{n} \mathbf{P}(Z_{n,k} \ge 2) \\ &= \sum_{k=1}^{n} \varepsilon_{n,k} \to 0 \quad \text{for } n \to \infty, \qquad \text{by assumption} \quad (c) \,. \end{aligned}$$

Hence, $(\overline{S}_n - S_n) \xrightarrow{p} 0$. Consequently, the convergence $\overline{S}_n \xrightarrow{\mathcal{D}} N_\lambda$ of the sums of truncated variables imply that also $S_n \xrightarrow{\mathcal{D}} N_\lambda$ (c.f. Exercise 3.2.8).

As seen in the context of the CLT, characteristic functions are a powerful tool for the convergence in distribution of sums of independent random variables (see Subsection 3.3.3). This is also evident in our proof of the Poisson approximation theorem. That is, to prove that $\overline{S}_n \xrightarrow{\mathcal{D}} N_\lambda$, if suffices by Levy's continuity theorem to show the convergence of the characteristic functions $\Phi_{\overline{S}_n}(\theta) \to \Phi_{N_\lambda}(\theta)$ for each $\theta \in \mathbb{R}$.

To this end, recall that $\overline{Z}_{n,k}$ are independent Bernoulli variables of parameters $p_{n,k}$, $k = 1, \dots, n$. Hence, by Lemma 3.3.8 and Example 3.3.5 we have that for $z_{n,k} = p_{n,k}(e^{i\theta} - 1)$,

$$\Phi_{\overline{S}_n}(\theta) = \prod_{k=1}^{n} \Phi_{\overline{Z}_{n,k}}(\theta) = \prod_{k=1}^{n} (1 - p_{n,k} + p_{n,k} e^{i\theta}) = \prod_{k=1}^{n} (1 + z_{n,k}) \,.$$

Our assumption (a) implies that for $n \to \infty$

$$z_n := \sum_{k=1}^{n} z_{n,k} = (\sum_{k=1}^{n} p_{n,k})(e^{i\theta} - 1) \to \lambda(e^{i\theta} - 1) := z_\infty \,.$$

Further, since $|z_{n,k}| \le 2p_{n,k}$, our assumptions (a) and (b) imply that for $n \to \infty$,

$$\eta_n = \sum_{k=1}^{n} |z_{n,k}|^2 \le 4 \sum_{k=1}^{n} p_{n,k}^2 \le 4 (\max_{k=1,\dots,n} \{p_{n,k}\}) (\sum_{k=1}^{n} p_{n,k}) \to 0 \,.$$

Applying Lemma 3.3.30 we conclude that when $n \to \infty$,

$$\Phi_{\overline{S}_n}(\theta) \to \exp(z_\infty) = \exp(\lambda(e^{i\theta} - 1)) = \Phi_{N_\lambda}(\theta)$$

(see (3.3.3) for the last identity), thus completing the proof. □

Remark. Recall Example 3.2.25 that the weak convergence of the laws of the integer valued S_n to that of N_λ also implies their convergence in total variation. In the setting of the Poisson approximation theorem, taking $\lambda_n = \sum_{k=1}^n p_{n,k}$, the more quantitative result

$$||\mathcal{P}_{\overline{S}_n} - \mathcal{P}_{N_{\lambda_n}}||_{tv} = \sum_{k=0}^{\infty} |\mathbf{P}(\overline{S}_n = k) - \mathbf{P}(N_{\lambda_n} = k)| \le 2 \min(\lambda_n^{-1}, 1) \sum_{k=1}^{n} p_{n,k}^2$$

due to Stein (1987) also holds (see also [**Dur03**, (2.6.5)] for a simpler argument, due to Hodges and Le Cam (1960), which is just missing the factor $\min(\lambda_n^{-1}, 1)$).

For the remainder of this subsection we list applications of the Poisson approximation theorem, starting with

EXAMPLE 3.4.2 (POISSON APPROXIMATION FOR THE BINOMIAL). *Take independent variables $Z_{n,k} \in \{0,1\}$, so $\varepsilon_{n,k} = 0$, with $p_{n,k} = p_n$ that does not depend on k. Then, the variable $S_n = \overline{S}_n$ has the Binomial distribution of parameters (n, p_n). By Stein's result, the Binomial distribution of parameters (n, p_n) is approximated well by the Poisson distribution of parameter $\lambda_n = np_n$, provided $p_n \to 0$. In case $\lambda_n = np_n \to \lambda < \infty$, Theorem 3.4.1 yields that the Binomial (n, p_n) laws converge weakly as $n \to \infty$ to the Poisson distribution of parameter λ. This is in agreement with Example 3.1.7 where we approximate the Binomial distribution of parameters (n, p) by the normal distribution, for in Example 3.1.8 we saw that, upon the same scaling, N_{λ_n} is also approximated well by the normal distribution when $\lambda_n \to \infty$.*

Recall the occupancy problem where we distribute at random r distinct balls among n distinct boxes and each of the possible n^r assignments of balls to boxes is equally likely. In Example 2.1.10 we considered the asymptotic fraction of empty boxes when $r/n \to \alpha$ and $n \to \infty$. Noting that the number of balls $M_{n,k}$ in the k-th box follows the Binomial distribution of parameters (r, n^{-1}), we deduce from Example 3.4.2 that $M_{n,k} \xrightarrow{\mathcal{D}} N_\alpha$. Thus, $\mathbf{P}(M_{n,k} = 0) \to \mathbf{P}(N_\alpha = 0) = e^{-\alpha}$. That is, for large n each box is empty with probability about $e^{-\alpha}$, which may explain (though not prove) the result of Example 2.1.10. Here we use the Poisson approximation theorem to tackle a different regime, in which $r = r_n$ is of order $n \log n$, and consequently, there are fewer empty boxes.

PROPOSITION 3.4.3. *Let S_n denote the number of empty boxes. Assuming $r = r_n$ is such that $ne^{-r/n} \to \lambda \in [0, \infty)$, we have that $S_n \xrightarrow{\mathcal{D}} N_\lambda$ as $n \to \infty$.*

PROOF. Let $Z_{n,k} = I_{M_{n,k}=0}$ for $k = 1, \ldots, n$, that is $Z_{n,k} = 1$ if the k-th box is empty and $Z_{n,k} = 0$ otherwise. Note that $S_n = \sum_{k=1}^n Z_{n,k}$, with each $Z_{n,k}$ having the Bernoulli distribution of parameter $p_n = (1 - n^{-1})^r$. Our assumption about r_n guarantees that $np_n \to \lambda$. If the occupancy $Z_{n,k}$ of the various boxes were mutually independent, then the stated convergence of S_n to N_λ would have followed from Theorem 3.4.1. Unfortunately, this is not the case, so we present a bare-hands approach showing that the dependence is weak enough to retain the same conclusion. To this end, first observe that for any $l = 1, 2, \ldots, n$, the probability that given boxes $k_1 < k_2 < \ldots < k_l$ are all empty is,

$$\mathbf{P}(Z_{n,k_1} = Z_{n,k_2} = \cdots = Z_{n,k_l} = 1) = (1 - \frac{l}{n})^r .$$

Let $p_l = p_l(r, n) = \mathbf{P}(S_n = l)$ denote the probability that exactly l boxes are empty out of the n boxes into which the r balls are placed at random. Then, considering all possible choices of the locations of these $l \geq 1$ empty boxes we get the identities $p_l(r, n) = b_l(r, n) p_0(r, n - l)$ for

$$b_l(r, n) = \binom{n}{l} \big(1 - \frac{l}{n}\big)^r . \tag{3.4.1}$$

Further, $p_0(r, n) = 1 - \mathbf{P}(\text{ at least one empty box})$, so that by the inclusion-exclusion formula,

$$p_0(r, n) = \sum_{l=0}^{n} (-1)^l b_l(r, n) . \tag{3.4.2}$$

According to part (b) of Exercise 3.4.4, $p_0(r,n) \to e^{-\lambda}$. Further, for fixed l we have that $(n-l)e^{-r/(n-l)} \to \lambda$, so as before we conclude that $p_0(r,n-l) \to e^{-\lambda}$. By part (a) of Exercise 3.4.4 we know that $b_l(r,n) \to \lambda^l/l!$ for fixed l, hence $p_l(r,n) \to e^{-\lambda}\lambda^l/l!$. As $p_l = \mathbf{P}(S_n = l)$, the proof of the proposition is thus complete. □

The following exercise provides the estimates one needs during the proof of Proposition 3.4.3 (for more details, see [**Dur03**, Theorem 2.6.6]).

Exercise 3.4.4. *Assuming $ne^{-r/n} \to \lambda$, show that*

(a) *$b_l(r,n)$ of (3.4.1) converges to $\lambda^l/l!$ for each fixed l.*
(b) *$p_0(r,n)$ of (3.4.2) converges to $e^{-\lambda}$.*

Finally, here is an application of Proposition 3.4.3 to the coupon collector's problem of Example 2.1.8, where T_n denotes the number of independent trials, it takes to have at least one representative of each of the n possible values (and each trial produces a value U_i that is distributed uniformly on the set of n possible values).

Example 3.4.5 (Revisiting the coupon collector's problem). *For any $x \in \mathbb{R}$, we have that*

$$\lim_{n\to\infty} \mathbf{P}(T_n - n\log n \le nx) = \exp(-e^{-x}), \tag{3.4.3}$$

which is an improvement over our weak law result that $T_n/n\log n \to 1$. Indeed, to derive (3.4.3) view the first r trials of the coupon collector as the random placement of r balls into n distinct boxes that correspond to the n possible values. From this point of view, the event $\{T_n \le r\}$ corresponds to filling all n boxes with the r balls, that is, having none empty. Taking $r = r_n = [n\log n + nx]$ we have that $ne^{-r/n} \to \lambda = e^{-x}$, and so it follows from Proposition 3.4.3 that $\mathbf{P}(T_n \le r_n) \to \mathbf{P}(N_\lambda = 0) = e^{-\lambda}$, as stated in (3.4.3).

Note that though $T_n = \sum_{k=1}^n X_{n,k}$ with $X_{n,k}$ independent, the convergence in distribution of T_n, given by (3.4.3), is to a non-normal limit. This should not surprise you, for the terms $X_{n,k}$ with k near n are large and do not satisfy Lindeberg's condition.

Exercise 3.4.6. *Recall that τ_ℓ^n denotes the first time one has ℓ distinct values when collecting coupons that are uniformly distributed on $\{1,2,\ldots,n\}$. Using the Poisson approximation theorem show that if $n \to \infty$ and $\ell = \ell(n)$ is such that $n^{-1/2}\ell \to \lambda \in [0,\infty)$, then $\tau_\ell^n - \ell \xrightarrow{\mathcal{D}} N$ with N a Poisson random variable of parameter $\lambda^2/2$.*

3.4.2. Poisson Process.

The Poisson process is a continuous time stochastic process $\omega \mapsto N_t(\omega)$, $t \ge 0$ which belongs to the following class of counting processes.

Definition 3.4.7. *A* counting process *is a mapping $\omega \longmapsto N_t(\omega)$, where $N_t(\omega)$ is a piecewise constant, nondecreasing, right continuous function of $t \ge 0$, with $N_0(\omega) = 0$ and (countably) infinitely many jump discontinuities, each of whom is of size one.*

Associated with each sample path *$N_t(\omega)$ of such a process are the jump times $0 = T_0 < T_1 < \cdots < T_n < \cdots$ such that $T_k = \inf\{t \ge 0 : N_t \ge k\}$ for each k, or equivalently*

$$N_t = \sup\{k \ge 0 : T_k \le t\}.$$

In applications we find such N_t as counting the number of discrete events occurring in the interval $[0,t]$ for each $t \geq 0$, with T_k denoting the arrival or occurrence time of the k-th such event.

Remark. It is possible to extend the notion of counting processes to discrete events indexed on $\mathbb{R}^d$, $d \geq 2$. This is done by assigning random integer counts N_A to Borel subsets A of $\mathbb{R}^d$ in an additive manner, that is, $N_{A\cup B} = N_A + N_B$ whenever A and B are disjoint. Such processes are called *point processes.* See also Exercise 7.1.13 for more about *Poisson point process* and inhomogeneous Poisson processes of non-constant rate.

Among all counting processes we characterize the Poisson process by the joint distribution of its jump (arrival) times $\{T_k\}$.

Definition 3.4.8. *The* Poisson process *of* rate $\lambda > 0$ *is the unique counting process with the gaps between jump times $\tau_k = T_k - T_{k-1}$, $k = 1, 2, \dots$ being i.i.d. random variables, each having the* exponential distribution *of parameter λ.*

Thus, from Exercise 1.4.46 we deduce that the k-th arrival time T_k of the Poisson process of rate λ has the *gamma density* of parameters $\alpha = k$ and λ,

$$f_{T_k}(u) = \frac{\lambda^k u^{k-1}}{(k-1)!} e^{-\lambda u} \mathbf{1}_{u>0} \, .$$

As we have seen in Example 2.3.7, counting processes appear in the context of renewal theory. In particular, as shown in Exercise 2.3.8, the Poisson process of rate λ satisfies the strong law of large numbers $t^{-1} N_t \overset{a.s.}{\to} \lambda$.

Recall that a random variable N has the Poisson(μ) law if

$$\mathbf{P}(N = n) = \frac{\mu^n}{n!} e^{-\mu}, \quad n = 0, 1, 2, \dots \, .$$

Our next proposition, which is often used as an alternative definition of the Poisson process, also explains its name.

Proposition 3.4.9. *For any ℓ and any $0 = t_0 < t_1 < \cdots < t_\ell$, the increments N_{t_1}, $N_{t_2} - N_{t_1}$, $\dots$, $N_{t_\ell} - N_{t_{\ell-1}}$, are independent random variables and for some $\lambda > 0$ and all $t > s \geq 0$, the increment $N_t - N_s$ has the* Poisson($\lambda(t-s)$) *law.*

Thus, the Poisson process has independent increments, each having a Poisson law, where the parameter of the count $N_t - N_s$ is proportional to the length of the corresponding interval $[s,t]$.

The proof of Proposition 3.4.9 relies on the *lack of memory* of the exponential distribution. That is, if the law of a random variable T is exponential (of some parameter $\lambda > 0$), then for all $t, s \geq 0$,

$$\mathbf{P}(T > t+s | T > t) = \frac{\mathbf{P}(T > t+s)}{\mathbf{P}(T > t)} = \frac{e^{-\lambda(t+s)}}{e^{-\lambda t}} = e^{-\lambda s} = \mathbf{P}(T > s) \, . \tag{3.4.4}$$

Indeed, the key to the proof of Proposition 3.4.9 is the following lemma.

Lemma 3.4.10. *Fixing $t > 0$, the variables $\{\tau'_j\}$ with $\tau'_1 = T_{N_t+1} - t$, and $\tau'_j = T_{N_t+j} - T_{N_t+j-1}$, $j \geq 2$ are i.i.d. each having the exponential distribution of parameter λ. Further, the collection $\{\tau'_j\}$ is independent of N_t which has the Poisson distribution of parameter λt.*

Remark. Note that in particular, $E_t = T_{N_t+1} - t$ which counts the time till next arrival occurs, hence called *the excess life time* at t, follows the exponential distribution of parameter λ.

Proof. Fixing $t > 0$ and $n \geq 1$ let $H_n(x) = \mathbf{P}(t \geq T_n > t - x)$. With $H_n(x) = \int_0^x f_{T_n}(t-y)dy$ and T_n independent of τ_{n+1}, we get by Fubini's theorem (for $I_{t \geq T_n > t - \tau_{n+1}}$), and the integration by parts of Lemma 1.4.30 that

$$\begin{aligned}\mathbf{P}(N_t = n) = \mathbf{P}(t \geq T_n > t - \tau_{n+1}) &= \mathbf{E}[H_n(\tau_{n+1})] \\ &= \int_0^t f_{T_n}(t-y)\mathbf{P}(\tau_{n+1} > y)dy \\ &= \int_0^t \frac{\lambda^n (t-y)^{n-1}}{(n-1)!} e^{-\lambda(t-y)} e^{-\lambda y} dy = e^{-\lambda t}\frac{(\lambda t)^n}{n!} \,. \qquad (3.4.5)\end{aligned}$$

As this applies for any $n \geq 1$, it follows that N_t has the Poisson distribution of parameter λt. Similarly, observe that for any $s_1 \geq 0$ and $n \geq 1$,

$$\begin{aligned}\mathbf{P}(N_t = n, \tau_1' > s_1) &= \mathbf{P}(t \geq T_n > t - \tau_{n+1} + s_1) \\ &= \int_0^t f_{T_n}(t-y)\mathbf{P}(\tau_{n+1} > s_1 + y)dy \\ &= e^{-\lambda s_1}\mathbf{P}(N_t = n) = \mathbf{P}(\tau_1 > s_1)\mathbf{P}(N_t = n) \,.\end{aligned}$$

Since $T_0 = 0$, $\mathbf{P}(N_t = 0) = e^{-\lambda t}$ and $T_1 = \tau_1$, in view of (3.4.4) this conclusion extends to $n = 0$, proving that τ_1' is independent of N_t and has the same exponential law as τ_1.

Next, fix arbitrary integer $k \geq 2$ and non-negative $s_j \geq 0$ for $j = 1, \ldots, k$. Then, for any $n \geq 0$, since $\{\tau_{n+j}, j \geq 2\}$ are i.i.d. and independent of (T_n, τ_{n+1}),

$$\begin{aligned}&\mathbf{P}(N_t = n, \tau_j' > s_j, j = 1, \ldots, k) \\ &= \mathbf{P}(t \geq T_n > t - \tau_{n+1} + s_1, T_{n+j} - T_{n+j-1} > s_j, j = 2, \ldots, k) \\ &= \mathbf{P}(t \geq T_n > t - \tau_{n+1} + s_1)\prod_{j=2}^k \mathbf{P}(\tau_{n+j} > s_j) = \mathbf{P}(N_t = n)\prod_{j=1}^k \mathbf{P}(\tau_j > s_j).\end{aligned}$$

Since $s_j \geq 0$ and $n \geq 0$ are arbitrary, this shows that the random variables N_t and τ_j', $j = 1, \ldots, k$ are mutually independent (c.f. Corollary 1.4.12), with each τ_j' having an exponential distribution of parameter λ. As k is arbitrary, the independence of N_t and the countable collection $\{\tau_j'\}$ follows by Definition 1.4.3. □

Proof of Proposition 3.4.9. Fix $t, s_j \geq 0$, $j = 1, \ldots, k$, and non-negative integers n and m_j, $1 \leq j \leq k$. The event $\{N_{s_j} = m_j, 1 \leq j \leq k\}$ is of the form $\{(\tau_1, \ldots, \tau_r) \in H\}$ for $r = m_k + 1$ and

$$H = \bigcap_{j=1}^k \{\underline{x} \in [0,\infty)^r : x_1 + \cdots + x_{m_j} \leq s_j < x_1 + \cdots + x_{m_j+1}\} \,.$$

Since the event $\{(\tau_1', \ldots, \tau_r') \in H\}$ is merely $\{N_{t+s_j} - N_t = m_j, 1 \leq j \leq k\}$, it follows form Lemma 3.4.10 that

$$\begin{aligned}&\mathbf{P}(N_t = n, N_{t+s_j} - N_t = m_j, 1 \leq j \leq k) = \mathbf{P}(N_t = n, (\tau_1', \ldots, \tau_r') \in H) \\ &= \mathbf{P}(N_t = n)\mathbf{P}((\tau_1, \ldots, \tau_r) \in H) = \mathbf{P}(N_t = n)\mathbf{P}(N_{s_j} = m_j, 1 \leq j \leq k) \,.\end{aligned}$$

By induction on ℓ this identity implies that if $0 = t_0 < t_1 < t_2 < \cdots < t_\ell$, then

$$\mathbf{P}(N_{t_i} - N_{t_{i-1}} = n_i, 1 \le i \le \ell) = \prod_{i=1}^{\ell} \mathbf{P}(N_{t_i - t_{i-1}} = n_i) \tag{3.4.6}$$

(the case $\ell = 1$ is trivial, and to advance the induction to $\ell + 1$ set $k = \ell$, $t = t_1$, $n = n_1$ and $s_j = t_{j+1} - t_1$, $m_j = \sum_{i=2}^{j+1} n_i$).

Considering (3.4.6) for $\ell = 2$, $t_2 = t > s = t_1$, and summing over the values of n_1 we see that $\mathbf{P}(N_t - N_s = n_2) = \mathbf{P}(N_{t-s} = n_2)$, hence by (3.4.5) we conclude that $N_t - N_s$ has the Poisson distribution of parameter $\lambda(t - s)$, as claimed. □

The Poisson process is also related to the *order statistics* $\{V_{n,k}\}$ for the uniform measure, as stated in the next two exercises.

EXERCISE 3.4.11. *Let $U_1, U_2, \ldots, U_n$ be i.i.d. with each U_i having the uniform measure on $(0, 1]$. Denote by $V_{n,k}$ the k-th smallest number in $\{U_1, \ldots, U_n\}$.*

(a) *Show that $(V_{n,1}, \ldots, V_{n,n})$ has the same law as $(T_1/T_{n+1}, \ldots, T_n/T_{n+1})$, where $\{T_k\}$ are the jump (arrival) times for a Poisson process of rate λ (see Subsection 1.4.2 for the definition of the* law *$\mathcal{P}_{\underline{X}}$ of a random vector $\underline{X}$).*

(b) *Taking $\lambda = 1$, deduce that $nV_{n,k} \xrightarrow{\mathcal{D}} T_k$ as $n \to \infty$ while k is fixed, where T_k has the* gamma density *of parameters $\alpha = k$ and $s = 1$.*

EXERCISE 3.4.12. *Fixing any positive integer n and $0 \le t_1 \le t_2 \le \cdots \le t_n \le t$, show that*

$$\mathbf{P}(T_k \le t_k,\ k = 1, \ldots, n | N_t = n) = \frac{n!}{t^n} \int_0^{t_1} \int_{x_1}^{t_2} \cdots (\int_{x_{n-1}}^{t_n} dx_n) dx_{n-1} \cdots dx_1 .$$

That is, conditional on the event $N_t = n$, the first n jump times $\{T_k : k = 1, \ldots, n\}$ have the same law as the order statistics *$\{V_{n,k} : k = 1, \ldots, n\}$ of a sample of n i.i.d random variables $U_1, \ldots, U_n$, each of which is uniformly distributed in $[0, t]$.*

Here is an application of Exercise 3.4.12.

EXERCISE 3.4.13. *Consider a Poisson process N_t of rate λ and jump times $\{T_k\}$.*

(a) *Compute the values of $g(n) = \mathbf{E}(I_{N_t = n} \sum_{k=1}^{n} T_k)$.*

(b) *Compute the value of $v = \mathbf{E}(\sum_{k=1}^{N_t} (t - T_k))$.*

(c) *Suppose that T_k is the arrival time to the train station of the k-th passenger on a train that departs the station at time t. What is the meaning of N_t and of v in this case?*

The representation of the *order statistics* $\{V_{n,k}\}$ in terms of the jump times of a Poisson process is very useful when studying the large n asymptotics of their spacings $\{R_{n,k}\}$. For example,

EXERCISE 3.4.14. *Let $R_{n,k} = V_{n,k} - V_{n,k-1}$, $k = 1, \ldots, n$, denote the spacings between $V_{n,k}$ of Exercise 3.4.11 (with $V_{n,0} = 0$). Show that as $n \to \infty$,*

$$\frac{n}{\log n} \max_{k=1,\ldots,n} R_{n,k} \xrightarrow{p} 1 , \tag{3.4.7}$$

and further for each fixed $x \ge 0$,

$$G_n(x) := n^{-1} \sum_{k=1}^{n} I_{\{R_{n,k} > x/n\}} \xrightarrow{p} e^{-x}\,, \tag{3.4.8}$$

$$B_n(x) := \mathbf{P}(\min_{k=1,\ldots,n} R_{n,k} > x/n^2) \to e^{-x}\,. \tag{3.4.9}$$

As we show next, the Poisson approximation theorem provides a characterization of the Poisson process that is very attractive for modeling real-world phenomena.

Corollary 3.4.15. *If* N_t *is a Poisson process of rate* $\lambda > 0$, *then for any fixed* k, $0 < t_1 < t_2 < \cdots < t_k$ *and nonnegative integers* $n_1, n_2, \cdots, n_k$,

$$\mathbf{P}(N_{t_k+h} - N_{t_k} = 1 | N_{t_j} = n_j,\ j \le k) = \lambda h + o(h),$$
$$\mathbf{P}(N_{t_k+h} - N_{t_k} \ge 2 | N_{t_j} = n_j,\ j \le k) = o(h),$$

where $o(h)$ *denotes a function* $f(h)$ *such that* $h^{-1} f(h) \to 0$ *as* $h \downarrow 0$.

Proof. Fixing k, the t_j and the n_j, denote by A the event $\{N_{t_j} = n_j,\ j \le k\}$. For a Poisson process of rate λ the random variable $N_{t_k+h} - N_{t_k}$ is independent of A with $\mathbf{P}(N_{t_k+h} - N_{t_k} = 1) = e^{-\lambda h}\lambda h$ and $\mathbf{P}(N_{t_k+h} - N_{t_k} \ge 2) = 1 - e^{-\lambda h}(1 + \lambda h)$. Since $e^{-\lambda h} = 1 - \lambda h + o(h)$ we see that the Poisson process satisfies this corollary. □

Our next exercise explores the phenomenon of *thinning*, that is, the partitioning of Poisson variables as sums of mutually independent Poisson variables of smaller parameter.

Exercise 3.4.16. *Suppose* $\{X_i\}$ *are i.i.d. with* $\mathbf{P}(X_i = j) = p_j$ *for* $j = 0, 1, \ldots, k$ *and* N *a Poisson random variable of parameter* λ *that is independent of* $\{X_k\}$. *Let*

$$N_j = \sum_{i=1}^{N} I_{X_i = j} \quad j = 0, \ldots, k\,.$$

(a) *Show that the variables* N_j, $j = 0, 1, \ldots, k$ *are mutually independent with* N_j *having a Poisson distribution of parameter* λp_j.

(b) *Show that the sub-sequence of jump times* $\{\widetilde{T}_k\}$ *obtained by independently keeping with probability* p *each of the jump times* $\{T_k\}$ *of a Poisson process* N_t *of rate* λ, *yields in turn a Poisson process* $\widetilde{N}_t$ *of rate* λp.

We conclude this section noting the *superposition* property, namely that the sum of two independent Poisson processes is yet another Poisson process.

Exercise 3.4.17. *Suppose* $N_t = N_t^{(1)} + N_t^{(2)}$ *where* $N_t^{(1)}$ *and* $N_t^{(2)}$ *are two independent Poisson processes of rates* $\lambda_1 > 0$ *and* $\lambda_2 > 0$, *respectively. Show that* N_t *is a Poisson process of rate* $\lambda_1 + \lambda_2$.

3.5. Random vectors and the multivariate CLT

The goal of this section is to extend the CLT to random vectors, that is, $\mathbb{R}^d$-valued random variables. Towards this end, we revisit in Subsection 3.5.1 the theory of weak convergence, this time in the more general setting of $\mathbb{R}^d$-valued random variables. Subsection 3.5.2 is devoted to the extension of characteristic functions and Lévy's theorems to the multivariate setting, culminating with the Cramér-wold reduction of convergence in distribution of random vectors to that of their one dimensional linear projections. Finally, in Subsection 3.5.3 we introduce the important concept of Gaussian random vectors and prove the multivariate CLT.

3.5.1. Weak convergence revisited. Recall Definition 3.2.17 of weak convergence for a sequence of probability measures on a topological space $\mathbb{S}$, which suggests the following definition for convergence in distribution of $\mathbb{S}$-valued random variables.

DEFINITION 3.5.1. *We say that* $(\mathbb{S}, \mathcal{B}_{\mathbb{S}})$*-valued random variables* X_n converge in distribution *to a* $(\mathbb{S}, \mathcal{B}_{\mathbb{S}})$*-valued random variable* X_∞*, denoted by* $X_n \xrightarrow{\mathcal{D}} X_\infty$*, if* $\mathcal{P}_{X_n} \overset{w}{\Rightarrow} \mathcal{P}_{X_\infty}$.

As already remarked, the *Portmanteau theorem* about equivalent characterizations of the weak convergence holds also when the probability measures ν_n are on a Borel measurable space $(\mathbb{S}, \mathcal{B}_{\mathbb{S}})$ with $(\mathbb{S}, \rho)$ any metric space (and in particular for $\mathbb{S} = \mathbb{R}^d$).

THEOREM 3.5.2 (PORTMANTEAU THEOREM). *The following five statements are equivalent for any probability measures* ν_n*,* $1 \le n \le \infty$ *on* $(\mathbb{S}, \mathcal{B}_{\mathbb{S}})$*, with* $(\mathbb{S}, \rho)$ *any metric space.*

(a) $\nu_n \overset{w}{\Rightarrow} \nu_\infty$
(b) *For every closed set* F*, one has* $\limsup\limits_{n\to\infty} \nu_n(F) \le \nu_\infty(F)$
(c) *For every open set* G*, one has* $\liminf\limits_{n\to\infty} \nu_n(G) \ge \nu_\infty(G)$
(d) *For every* ν_∞*-continuity set* A*, one has* $\lim\limits_{n\to\infty} \nu_n(A) = \nu_\infty(A)$
(e) *If the Borel function* $g : \mathbb{S} \mapsto \mathbb{R}$ *is such that* $\nu_\infty(\mathbf{D}_g) = 0$*, then* $\nu_n \circ g^{-1} \overset{w}{\Rightarrow} \nu_\infty \circ g^{-1}$ *and if in addition* g *is bounded then* $\nu_n(g) \to \nu_\infty(g)$*.*

REMARK. For $\mathbb{S} = \mathbb{R}$, the equivalence of (a)–(d) is the content of Theorem 3.2.21 while Proposition 3.2.19 derives (e) out of (a) (in the context of convergence in distribution, that is, $X_n \xrightarrow{\mathcal{D}} X_\infty$ and $\mathbf{P}(X_\infty \in \mathbf{D}_g) = 0$ implying that $g(X_n) \xrightarrow{\mathcal{D}} g(X_\infty)$). In addition to proving the converse of the continuous mapping property, we extend the validity of this equivalence to any metric space $(\mathbb{S}, \rho)$, for we shall apply it again in Subsection 9.2, considering there $\mathbb{S} = C([0, \infty))$, the metric space of all continuous functions on $[0, \infty)$.

PROOF. The derivation of $(b) \Rightarrow (c) \Rightarrow (d)$ in Theorem 3.2.21 applies for any topological space. The direction $(e) \Rightarrow (a)$ is also obvious since $h \in C_b(\mathbb{S})$ has $\mathbf{D}_h = \emptyset$ and $C_b(\mathbb{S})$ is a subset of the bounded Borel functions on the same space (c.f. Exercise 1.2.20). So taking $g \in C_b(\mathbb{S})$ in (e) results with (a). It thus remains only to show that $(a) \Rightarrow (b)$ and that $(d) \Rightarrow (e)$, which we proceed to show next.

$(a) \Rightarrow (b)$. Fixing $A \in \mathcal{B}_{\mathbb{S}}$ let $\rho_A(x) = \inf_{y \in A} \rho(x, y) : \mathbb{S} \mapsto [0, \infty)$. Since $|\rho_A(x) - \rho_A(x')| \le \rho(x, x')$ for any x, x', it follows that $x \mapsto \rho_A(x)$ is a continuous function on $(\mathbb{S}, \rho)$. Consequently, $h_r(x) = (1 - r\rho_A(x))_+ \in C_b(\mathbb{S})$ for all $r \ge 0$. Further, $\rho_A(x) = 0$ for all $x \in A$, implying that $h_r \ge I_A$ for all r. Thus, applying part (a) of the Portmanteau theorem for h_r we have that

$$\limsup_{n\to\infty} \nu_n(A) \le \lim_{n\to\infty} \nu_n(h_r) = \nu_\infty(h_r)\,.$$

As $\rho_A(x) = 0$ if and only if $x \in \overline{A}$ it follows that $h_r \downarrow I_{\overline{A}}$ as $r \to \infty$, resulting with

$$\limsup_{n\to\infty} \nu_n(A) \le \nu_\infty(\overline{A})\,.$$

Taking $A = \overline{A} = F$ a closed set, we arrive at part (b) of the theorem.
$(d) \Rightarrow (e)$. Fix a Borel function $g : \mathbb{S} \mapsto \mathbb{R}$ with $K = \sup_x |g(x)| < \infty$ such that $\nu_\infty(\mathbf{D}_g) = 0$. Clearly, $\{\alpha \in \mathbb{R} : \nu_\infty \circ g^{-1}(\{\alpha\}) > 0\}$ is a countable set. Thus, fixing $\varepsilon > 0$ we can pick $\ell < \infty$ and $\alpha_0 < \alpha_1 < \cdots < \alpha_\ell$ such that $\nu_\infty \circ g^{-1}(\{\alpha_i\}) = 0$ for $0 \le i \le \ell$, $\alpha_0 < -K < K < \alpha_\ell$ and $\alpha_i - \alpha_{i-1} < \varepsilon$ for $1 \le i \le \ell$. Let $A_i = \{x : \alpha_{i-1} < g(x) \le \alpha_i\}$ for $i = 1, \dots, \ell$, noting that $\partial A_i \subset \{x : g(x) = \alpha_{i-1}$, or $g(x) = \alpha_i\} \cup \mathbf{D}_g$. Consequently, by our assumptions about $g(\cdot)$ and $\{\alpha_i\}$ we have that $\nu_\infty(\partial A_i) = 0$ for each $i = 1, \dots, \ell$. It thus follows from part (d) of the Portmanteau theorem that

$$\sum_{i=1}^{\ell} \alpha_i \nu_n(A_i) \to \sum_{i=1}^{\ell} \alpha_i \nu_\infty(A_i)$$

as $n \to \infty$. Our choice of α_i and A_i is such that $g \le \sum_{i=1}^{\ell} \alpha_i I_{A_i} \le g + \varepsilon$, resulting with

$$\nu_n(g) \le \sum_{i=1}^{\ell} \alpha_i \nu_n(A_i) \le \nu_n(g) + \varepsilon$$

for $n = 1, 2, \dots, \infty$. Considering first $n \to \infty$ followed by $\varepsilon \downarrow 0$, we establish that $\nu_n(g) \to \nu_\infty(g)$. More generally, recall that $\mathbf{D}_{h \circ g} \subseteq \mathbf{D}_g$ for any $g : \mathbb{S} \mapsto \mathbb{R}$ and $h \in C_b(\mathbb{R})$. Thus, by the preceding proof $\nu_n(h \circ g) \to \nu_\infty(h \circ g)$ as soon as $\nu_\infty(\mathbf{D}_g) = 0$. This applies for every $h \in C_b(\mathbb{R})$, so in this case $\nu_n \circ g^{-1} \overset{w}{\Rightarrow} \nu_\infty \circ g^{-1}$. □

We next show that the relation of Exercise 3.2.6 between convergences in probability and in distribution also extends to any metric space $(\mathbb{S}, \rho)$, a fact we will later use in Subsection 9.2, when considering the metric space of all continuous functions on $[0, \infty)$.

Corollary 3.5.3. *If random variables X_n, $1 \le n \le \infty$ on the same probability space and taking value in a metric space $(\mathbb{S}, \rho)$ are such that $\rho(X_n, X_\infty) \xrightarrow{p} 0$, then $X_n \xrightarrow{\mathcal{D}} X_\infty$.*

Proof. Fixing $h \in C_b(\mathbb{S})$ and $\varepsilon > 0$, we have by continuity of $h(\cdot)$ that $G_r \uparrow \mathbb{S}$, where

$$G_r = \{y \in \mathbb{S} : |h(x) - h(y)| \le \varepsilon \text{ whenever } \rho(x, y) \le r^{-1}\}.$$

By definition, if $X_\infty \in G_r$ and $\rho(X_n, X_\infty) \le r^{-1}$ then $|h(X_n) - h(X_\infty)| \le \varepsilon$. Hence, for any $n, r \ge 1$,

$$\mathbf{E}[|h(X_n) - h(X_\infty)|] \le \varepsilon + 2\|h\|_\infty (\mathbf{P}(X_\infty \notin G_r) + \mathbf{P}(\rho(X_n, X_\infty) > r^{-1})),$$

where $\|h\|_\infty = \sup_{x \in \mathbb{S}} |h(x)|$ is finite (by the boundedness of h). Considering $n \to \infty$ followed by $r \to \infty$ we deduce from the convergence in probability of $\rho(X_n, X_\infty)$ to zero, that

$$\limsup_{n \to \infty} \mathbf{E}[|h(X_n) - h(X_\infty)|] \le \varepsilon + 2\|h\|_\infty \lim_{r \to \infty} \mathbf{P}(X_\infty \notin G_r) = \varepsilon.$$

Since this applies for any $\varepsilon > 0$, it follows by the triangle inequality that $\mathbf{E}h(X_n) \to \mathbf{E}h(X_\infty)$ for all $h \in C_b(\mathbb{S})$, i.e. $X_n \xrightarrow{\mathcal{D}} X_\infty$. □

Remark. The notion of *distribution function* for an $\mathbb{R}^d$-valued random vector $\underline{X} = (X_1, \dots, X_d)$ is

$$F_{\underline{X}}(\underline{x}) = \mathbf{P}(X_1 \le x_1, \dots, X_d \le x_d).$$

Inducing a partial order on $\mathbb{R}^d$ by $\underline{x} \leq \underline{y}$ if and only if $\underline{x} - \underline{y}$ has only non-negative coordinates, each distribution function $F_{\underline{X}}(\underline{x})$ has the three properties listed in Theorem 1.2.36. Unfortunately, these three properties are not sufficient for a given function $F : \mathbb{R}^d \mapsto [0, 1]$ to be a distribution function. For example, since the measure of each rectangle $A = \prod_{i=1}^d (a_i, b_i]$ should be positive, the additional constraint of the form $\Delta_A F = \sum_{j=1}^{2^d} \pm F(\underline{x}_j) \geq 0$ should hold if $F(\cdot)$ is to be a distribution function. Here $\underline{x}_j$ enumerates the 2^d corners of the rectangle A and each corner is taken with a positive sign if and only if it has an even number of coordinates from the collection $\{a_1, \ldots, a_d\}$. Adding the fourth property that $\Delta_A F \geq 0$ for each rectangle $A \subset \mathbb{R}^d$, we get the necessary and sufficient conditions for $F(\cdot)$ to be a distribution function of some $\mathbb{R}^d$-valued random variable (c.f. [**Dur03**, Theorem A.1.6] or [**Bil95**, Theorem 12.5] for a detailed proof).

Recall Definition 3.2.31 of *uniform tightness*, where for $\mathbb{S} = \mathbb{R}^d$ we can take $K_\varepsilon = [-M_\varepsilon, M_\varepsilon]^d$ with no loss of generality. Though Prohorov's theorem about uniform tightness (i.e. Theorem 3.2.34) is beyond the scope of these notes, we shall only need in the sequel the fact that a uniformly tight sequence of probability measures has at least one limit point. This can be proved for $\mathbb{S} = \mathbb{R}^d$ in a manner similar to what we have done in Theorem 3.2.37 and Lemma 3.2.38 for $\mathbb{S} = \mathbb{R}^1$, using the corresponding concept of distribution function $F_{\underline{X}}(\cdot)$ (see [**Dur03**, Theorem 2.9.2] for more details).

3.5.2. Characteristic function.

We start by extending the useful notion of characteristic function to the context of $\mathbb{R}^d$-valued random variables (which we also call hereafter random vectors).

Definition 3.5.4. *Adopting the notation* $(\underline{x}, \underline{y}) = \sum_{i=1}^d x_i y_i$ *for* $\underline{x}, \underline{y} \in \mathbb{R}^d$*, a* random vector $\underline{X} = (X_1, X_2, \cdots, X_d)$ *with values in* $\mathbb{R}^d$ *has the* characteristic function

$$\Phi_{\underline{X}}(\underline{\theta}) = \mathbf{E}[e^{i(\underline{\theta}, \underline{X})}] ,$$

where $\underline{\theta} = (\theta_1, \theta_2, \cdots, \theta_d) \in \mathbb{R}^d$ *and* $i = \sqrt{-1}$.

Remark. The characteristic function $\Phi_{\underline{X}} : \mathbb{R}^d \mapsto \mathbb{C}$ exists for any $\underline{X}$ since

$$e^{i(\underline{\theta}, \underline{X})} = \cos(\underline{\theta}, \underline{X}) + i \sin(\underline{\theta}, \underline{X}) , \tag{3.5.1}$$

with both real and imaginary parts being bounded (hence integrable) random variables. Actually, it is easy to check that all five properties of Proposition 3.3.2 hold, where part (e) is modified to $\Phi_{\mathbf{A}^t \underline{X} + \underline{b}}(\underline{\theta}) = \exp(i(\underline{b}, \underline{\theta}))\Phi_{\underline{X}}(\mathbf{A}\underline{\theta})$, for any non-random $d \times d$-dimensional matrix $\mathbf{A}$ and $\underline{b} \in \mathbb{R}^d$ (with $\mathbf{A}^t$ denoting the transpose of the matrix $\mathbf{A}$).

Here is the extension of the notion of probability density function (as in Definition 1.2.39) to a random vector.

Definition 3.5.5. *Suppose* $f_{\underline{X}}$ *is a non-negative Borel measurable function with* $\int_{\mathbb{R}^d} f_{\underline{X}}(\underline{x}) d\underline{x} = 1$*. We say that a random vector* $\underline{X} = (X_1, \ldots, X_d)$ *has a* probability density function $f_{\underline{X}}(\cdot)$ *if for every* $\underline{b} = (b_1, \ldots, b_d)$,

$$F_{\underline{X}}(\underline{b}) = \int_{-\infty}^{b_1} \cdots \int_{-\infty}^{b_d} f_{\underline{X}}(x_1, \ldots, x_d) dx_d \cdots dx_1$$

(such $f_{\underline{X}}$ is sometimes called the joint density of $X_1, \ldots, X_d$). This is the same as saying that the law of $\underline{X}$ is of the form $f_{\underline{X}}\lambda^d$ with λ^d the d-fold product Lebesgue measure on $\mathbb{R}^d$ (i.e. the $d > 1$ extension of Example 1.3.60).

EXAMPLE 3.5.6. *We have the following extension of the Fourier transform formula (3.3.4) to random vectors $\underline{X}$ with density,*

$$\Phi_{\underline{X}}(\underline{\theta}) = \int_{\mathbb{R}^d} e^{i(\underline{\theta},\underline{x})} f_{\underline{X}}(\underline{x}) d\underline{x}$$

(this is merely a special case of the extension of Corollary 1.3.62 to $h : \mathbb{R}^d \mapsto \mathbb{R}$).

We next state and prove the corresponding extension of Lévy's inversion theorem.

THEOREM 3.5.7 (LÉVY'S INVERSION THEOREM). *Suppose $\Phi_{\underline{X}}(\underline{\theta})$ is the characteristic function of random vector $\underline{X} = (X_1, \ldots, X_d)$ whose law is $\mathcal{P}_{\underline{X}}$, a probability measure on $(\mathbb{R}^d, \mathcal{B}_{\mathbb{R}^d})$. If $A = [a_1, b_1] \times \cdots \times [a_d, b_d]$ with $\mathcal{P}_{\underline{X}}(\partial A) = 0$, then*

$$\mathcal{P}_{\underline{X}}(A) = \lim_{T\to\infty} \int_{[-T,T]^d} \prod_{j=1}^{d} \psi_{a_j,b_j}(\theta_j) \Phi_{\underline{X}}(\underline{\theta}) d\underline{\theta} \tag{3.5.2}$$

for $\psi_{a,b}(\cdot)$ of (3.3.5). Further, the characteristic function determines the law *of a random vector. That is, if $\Phi_{\underline{X}}(\underline{\theta}) = \Phi_{\underline{Y}}(\underline{\theta})$ for all $\underline{\theta}$ then $\underline{X}$ has the same law as $\underline{Y}$.*

PROOF. We derive (3.5.2) by adapting the proof of Theorem 3.3.12. First apply Fubini's theorem with respect to the product of Lebesgue's measure on $[-T, T]^d$ and the law of $\underline{X}$ (both of which are finite measures on $\mathbb{R}^d$) to get the identity

$$J_T(\underline{a}, \underline{b}) := \int_{[-T,T]^d} \prod_{j=1}^{d} \psi_{a_j,b_j}(\theta_j) \Phi_{\underline{X}}(\underline{\theta}) d\underline{\theta} = \int_{\mathbb{R}^d} \Big[\prod_{j=1}^{d} \int_{-T}^{T} h_{a_j,b_j}(x_j, \theta_j) d\theta_j \Big] d\mathcal{P}_{\underline{X}}(\underline{x})$$

(where $h_{a,b}(x, \theta) = \psi_{a,b}(\theta) e^{i\theta x}$). In the course of proving Theorem 3.3.12 we have seen that for $j = 1, \ldots, d$ the integral over θ_j is uniformly bounded in T and that it converges to $g_{a_j,b_j}(x_j)$ as $T \uparrow \infty$. Thus, by bounded convergence it follows that

$$\lim_{T\uparrow\infty} J_T(\underline{a}, \underline{b}) = \int_{\mathbb{R}^d} g_{\underline{a},\underline{b}}(\underline{x}) d\mathcal{P}_{\underline{X}}(\underline{x}) ,$$

where

$$g_{\underline{a},\underline{b}}(\underline{x}) = \prod_{j=1}^{d} g_{a_j,b_j}(x_j) ,$$

is zero on A^c and one on A^o (see the explicit formula for $g_{a,b}(x)$ provided there). So, our assumption that $\mathcal{P}_{\underline{X}}(\partial A) = 0$ implies that the limit of $J_T(\underline{a}, \underline{b})$ as $T \uparrow \infty$ is merely $\mathcal{P}_{\underline{X}}(A)$, thus establishing (3.5.2).

Suppose now that $\Phi_{\underline{X}}(\underline{\theta}) = \Phi_{\underline{Y}}(\underline{\theta})$ for all $\underline{\theta}$. Adapting the proof of Corollary 3.3.14 to the current setting, let $\mathcal{J} = \{\alpha \in \mathbb{R} : \mathbf{P}(X_j = \alpha) > 0$ or $\mathbf{P}(Y_j = \alpha) > 0$ for some $j = 1, \ldots, d\}$ noting that if all the coordinates $\{a_j, b_j, j = 1, \ldots, d\}$ of a rectangle A are from the complement of $\mathcal{J}$ then both $\mathcal{P}_{\underline{X}}(\partial A) = 0$ and $\mathcal{P}_{\underline{Y}}(\partial A) = 0$. Thus, by (3.5.2) we have that $\mathcal{P}_{\underline{X}}(A) = \mathcal{P}_{\underline{Y}}(A)$ for any A in the collection $\mathcal{C}$ of rectangles with coordinates in the complement of $\mathcal{J}$. Recall that $\mathcal{J}$ is countable, so for any rectangle A there exists $A_n \in \mathcal{C}$ such that $A_n \downarrow A$, and by continuity from above of both $\mathcal{P}_{\underline{X}}$ and $\mathcal{P}_{\underline{Y}}$ it follows that $\mathcal{P}_{\underline{X}}(A) = \mathcal{P}_{\underline{Y}}(A)$ for *every rectangle* A. In view of Proposition 1.1.39 and Exercise 1.1.21 this implies that the probability measures $\mathcal{P}_{\underline{X}}$ and $\mathcal{P}_{\underline{Y}}$ agree on all Borel subsets of $\mathbb{R}^d$. □

We next provide the ingredients needed when using characteristic functions enroute to the derivation of a convergence in distribution result for random vectors. To this end, we start with the following analog of Lemma 3.3.16.

LEMMA 3.5.8. *Suppose the random vectors $\underline{X}_n$, $1 \le n \le \infty$ on $\mathbb{R}^d$ are such that $\Phi_{\underline{X}_n}(\underline{\theta}) \to \Phi_{\underline{X}_\infty}(\underline{\theta})$ as $n \to \infty$ for each $\underline{\theta} \in \mathbb{R}^d$. Then, the corresponding sequence of laws $\{\mathcal{P}_{\underline{X}_n}\}$ is uniformly tight.*

PROOF. Fixing $\underline{\theta} \in \mathbb{R}^d$ consider the sequence of random variables $Y_n = (\underline{\theta}, \underline{X}_n)$. Since $\Phi_{Y_n}(\alpha) = \Phi_{\underline{X}_n}(\alpha\underline{\theta})$ for $1 \le n \le \infty$, we have that $\Phi_{Y_n}(\alpha) \to \Phi_{Y_\infty}(\alpha)$ for all $\alpha \in \mathbb{R}$. The uniform tightness of the laws of Y_n then follows by Lemma 3.3.16. Considering $\underline{\theta}_1, \dots, \underline{\theta}_d$ which are the unit vectors in the d different coordinates, we have the uniform tightness of the laws of $X_{n,j}$ for the sequence of random vectors $\underline{X}_n = (X_{n,1}, X_{n,2}, \dots, X_{n,d})$ and each fixed coordinate $j = 1, \dots, d$. For the compact sets $K_\varepsilon = [-M_\varepsilon, M_\varepsilon]^d$ and all n,

$$\mathbf{P}(\underline{X}_n \notin K_\varepsilon) \le \sum_{j=1}^d \mathbf{P}(|X_{n,j}| > M_\varepsilon) .$$

As d is finite, this leads from the uniform tightness of the laws of $X_{n,j}$ for each $j = 1, \dots, d$ to the uniform tightness of the laws of $\underline{X}_n$. □

Equipped with Lemma 3.5.8 we are ready to state and prove Lévy's continuity theorem.

THEOREM 3.5.9 (LÉVY'S CONTINUITY THEOREM). *Let $\underline{X}_n$, $1 \le n \le \infty$ be random vectors with characteristic functions $\Phi_{\underline{X}_n}(\underline{\theta})$. Then, $\underline{X}_n \xrightarrow{\mathcal{D}} \underline{X}_\infty$ if and only if $\Phi_{\underline{X}_n}(\underline{\theta}) \to \Phi_{\underline{X}_\infty}(\underline{\theta})$ as $n \to \infty$ for each fixed $\underline{\theta} \in \mathbb{R}^d$.*

PROOF. This is a re-run of the proof of Theorem 3.3.17, adapted to $\mathbb{R}^d$-valued random variables. First, both $\underline{x} \mapsto \cos((\underline{\theta}, \underline{x}))$ and $\underline{x} \mapsto \sin((\underline{\theta}, \underline{x}))$ are bounded continuous functions, so if $\underline{X}_n \xrightarrow{\mathcal{D}} \underline{X}_\infty$, then clearly as $n \to \infty$,

$$\Phi_{\underline{X}_n}(\underline{\theta}) = \mathbf{E}\Big[e^{i(\underline{\theta}, \underline{X}_n)}\Big] \to \mathbf{E}\Big[e^{i(\underline{\theta}, \underline{X}_\infty)}\Big] = \Phi_{\underline{X}_\infty}(\underline{\theta}).$$

For the converse direction, assuming that $\Phi_{\underline{X}_n} \to \Phi_{\underline{X}_\infty}$ point-wise, we know from Lemma 3.5.8 that the collection $\{\mathcal{P}_{\underline{X}_n}\}$ is uniformly tight. Hence, by Prohorov's theorem, for every subsequence $n(m)$ there is a further sub-subsequence $n(m_k)$ such that $\mathcal{P}_{\underline{X}_{n(m_k)}}$ converges weakly to some probability measure $\mathcal{P}_{\underline{Y}}$, possibly dependent upon the choice of $n(m)$. As $\underline{X}_{n(m_k)} \xrightarrow{\mathcal{D}} \underline{Y}$, we have by the preceding part of the proof that $\Phi_{\underline{X}_{n(m_k)}} \to \Phi_{\underline{Y}}$, and necessarily $\Phi_{\underline{Y}} = \Phi_{\underline{X}_\infty}$. The characteristic function determines the law (see Theorem 3.5.7), so $\underline{Y} \overset{\mathcal{D}}{=} \underline{X}_\infty$ is independent of the choice of $n(m)$. Thus, fixing $h \in C_b(\mathbb{R}^d)$, the sequence $y_n = \mathbf{E}h(\underline{X}_n)$ is such that every subsequence $y_{n(m)}$ has a further sub-subsequence $y_{n(m_k)}$ that converges to y_∞. Consequently, $y_n \to y_\infty$ (see Lemma 2.2.11). This applies for all $h \in C_b(\mathbb{R}^d)$, so we conclude that $\underline{X}_n \xrightarrow{\mathcal{D}} \underline{X}_\infty$, as stated. □

REMARK. As in the case of Theorem 3.3.17, it is not hard to show that if $\Phi_{\underline{X}_n}(\underline{\theta}) \to \Phi(\underline{\theta})$ as $n \to \infty$ and $\Phi(\underline{\theta})$ is continuous at $\underline{\theta} = \underline{0}$ then Φ is necessarily the characteristic function of some random vector $\underline{X}_\infty$ and consequently $\underline{X}_n \xrightarrow{\mathcal{D}} \underline{X}_\infty$.

The proof of the multivariate CLT is just one of the results that rely on the following immediate corollary of Lévy's continuity theorem.

COROLLARY 3.5.10 (CRAMÉR-WOLD DEVICE). *A sufficient condition for* $\underline{X}_n \xrightarrow{\mathcal{D}} \underline{X}_\infty$ *is that* $(\underline{\theta}, \underline{X}_n) \xrightarrow{\mathcal{D}} (\underline{\theta}, \underline{X}_\infty)$ *for each* $\underline{\theta} \in \mathbb{R}^d$.

PROOF. Since $(\underline{\theta}, \underline{X}_n) \xrightarrow{\mathcal{D}} (\underline{\theta}, \underline{X}_\infty)$ it follows by Lévy's continuity theorem (for $d = 1$, that is, Theorem 3.3.17), that

$$\lim_{n\to\infty} \mathbf{E}\Big[e^{i(\underline{\theta},\underline{X}_n)}\Big] = \mathbf{E}\Big[e^{i(\underline{\theta},\underline{X}_\infty)}\Big] .$$

As this applies for any $\underline{\theta} \in \mathbb{R}^d$, we get that $\underline{X}_n \xrightarrow{\mathcal{D}} \underline{X}_\infty$ by applying Lévy's continuity theorem in $\mathbb{R}^d$ (i.e., Theorem 3.5.9), now in the converse direction. □

REMARK. Beware that it is not enough to consider only finitely many values of $\underline{\theta}$ in the Cramér-Wold device. For example, consider the random vectors $\underline{X}_n = (X_n, Y_n)$ with $\{X_n, Y_{2n}\}$ i.i.d. and $Y_{2n+1} = X_{2n+1}$. Convince yourself that in this case $X_n \xrightarrow{\mathcal{D}} X_1$ and $Y_n \xrightarrow{\mathcal{D}} Y_1$ but the random vectors $\underline{X}_n$ do not converge in distribution (to any limit).

The computation of the characteristic function is much simplified in the presence of independence.

EXERCISE 3.5.11. *Show that if* $\underline{Y} = (Y_1, \dots, Y_d)$ *with* Y_k *mutually independent R.V., then for all* $\underline{\theta} = (\theta_1, \dots, \theta_d) \in \mathbb{R}^d$,

$$\Phi_{\underline{Y}}(\underline{\theta}) = \prod_{k=1}^{d} \Phi_{Y_k}(\theta_k) \tag{3.5.3}$$

Conversely, show that if (3.5.3) holds for all $\underline{\theta} \in \mathbb{R}^d$, *the random variables* Y_k, $k = 1, \dots, d$ *are mutually independent of each other.*

3.5.3. Gaussian random vectors and the multivariate CLT.

Recall the following linear algebra concept.

DEFINITION 3.5.12. *An* $d \times d$ *matrix* $\mathbf{A}$ *with entries* A_{jk} *is called* nonnegative definite *(or positive semidefinite) if* $A_{jk} = A_{kj}$ *for all* j, k, *and for any* $\underline{\theta} \in \mathbb{R}^d$

$$(\underline{\theta}, \mathbf{A}\underline{\theta}) = \sum_{j=1}^{d}\sum_{k=1}^{d} \theta_j A_{jk} \theta_k \geq 0.$$

We are ready to define the class of multivariate normal distributions via the corresponding characteristic functions.

DEFINITION 3.5.13. *We say that a random vector* $\underline{X} = (X_1, X_2, \cdots, X_d)$ *is* Gaussian, *or alternatively that it has a* multivariate normal *distribution if*

$$\Phi_{\underline{X}}(\underline{\theta}) = e^{-\frac{1}{2}(\underline{\theta}, \mathbf{V}\underline{\theta})} e^{i(\underline{\theta}, \underline{\mu})}, \tag{3.5.4}$$

for some nonnegative definite $d \times d$ *matrix* $\mathbf{V}$, *some* $\underline{\mu} = (\mu_1, \dots, \mu_d) \in \mathbb{R}^d$ *and all* $\underline{\theta} = (\theta_1, \dots, \theta_d) \in \mathbb{R}^d$. *We denote such a law by* $\mathcal{N}(\underline{\mu}, \mathbf{V})$.

REMARK. For $d = 1$ this definition coincides with Example 3.3.6.

Our next proposition proves that the multivariate $\mathcal{N}(\underline{\mu}, \mathbf{V})$ distribution is well defined and further links the vector $\underline{\mu}$ and the matrix $\mathbf{V}$ to the first two moments of this distribution.

PROPOSITION 3.5.14. *The formula (3.5.4) corresponds to the characteristic function of a probability measure on $\mathbb{R}^d$. Further, the parameters $\underline{\mu}$ and $\mathbf{V}$ of the Gaussian random vector $\underline{X}$ are merely $\mu_j = \mathbf{E}X_j$ and $V_{jk} = \mathsf{Cov}(X_j, X_k)$, $j, k = 1, \ldots, d$.*

PROOF. Any nonnegative definite matrix $\mathbf{V}$ can be written as $\mathbf{V} = \mathbf{U}^t\mathbf{D}^2\mathbf{U}$ for some orthogonal matrix $\mathbf{U}$ (i.e., such that $\mathbf{U}^t\mathbf{U} = \mathbf{I}$, the $d \times d$-dimensional identity matrix), and some diagonal matrix $\mathbf{D}$. Consequently,

$$(\underline{\theta}, \mathbf{V}\underline{\theta}) = (\mathbf{A}\underline{\theta}, \mathbf{A}\underline{\theta})$$

for $\mathbf{A} = \mathbf{D}\mathbf{U}$ and all $\underline{\theta} \in \mathbb{R}^d$. We claim that (3.5.4) is the characteristic function of the random vector $\underline{X} = \mathbf{A}^t\underline{Y} + \underline{\mu}$, where $\underline{Y} = (Y_1, \ldots, Y_d)$ has i.i.d. coordinates Y_k, each of which has the standard normal distribution. Indeed, by Exercise 3.5.11 $\Phi_{\underline{Y}}(\underline{\theta}) = \exp(-\frac{1}{2}(\underline{\theta}, \underline{\theta}))$ is the product of the characteristic functions $\exp(-\theta_k^2/2)$ of the standard normal distribution (see Example 3.3.6), and by part (e) of Proposition 3.3.2, $\Phi_{\underline{X}}(\underline{\theta}) = \exp(i(\underline{\theta}, \underline{\mu}))\Phi_{\underline{Y}}(\mathbf{A}\underline{\theta})$, yielding the formula (3.5.4).

We have just shown that $\underline{X}$ has the $\mathcal{N}(\underline{\mu}, \mathbf{V})$ distribution if $\underline{X} = \mathbf{A}^t\underline{Y} + \underline{\mu}$ for a Gaussian random vector $\underline{Y}$ (whose distribution is $\mathcal{N}(\underline{0}, \mathbf{I})$), such that $\mathbf{E}Y_j = 0$ and $\mathsf{Cov}(Y_j, Y_k) = \mathbf{1}_{j=k}$ for $j, k = 1, \ldots, d$. It thus follows by linearity of the expectation and the bi-linearity of the covariance that $\mathbf{E}X_j = \mu_j$ and $\mathsf{Cov}(X_j, X_k) = [\mathbf{E}\mathbf{A}^t\underline{Y}(\mathbf{A}^t\underline{Y})^t]_{jk} = (\mathbf{A}^t\mathbf{I}\mathbf{A})_{jk} = V_{jk}$, as claimed. □

Definition 3.5.13 allows for $\mathbf{V}$ that is non-invertible, so for example the constant random vector $\underline{X} = \underline{\mu}$ is considered a Gaussian random vector though it obviously does not have a density. The reason we make this choice is to have the collection of multivariate normal distributions closed with respect to L^2-convergence, as we prove below to be the case.

PROPOSITION 3.5.15. *Suppose Gaussian random vectors $\underline{X}_n$ converge in L^2 to a random vector $\underline{X}_\infty$, that is, $\mathbf{E}[\|\underline{X}_n - \underline{X}_\infty\|^2] \to 0$ as $n \to \infty$. Then, $\underline{X}_\infty$ is a Gaussian random vector, whose parameters are the limits of the corresponding parameters of $\underline{X}_n$.*

PROOF. Recall that the convergence in L^2 of $\underline{X}_n$ to $\underline{X}_\infty$ implies that $\underline{\mu}_n = \mathbf{E}\underline{X}_n$ converge to $\underline{\mu}_\infty = \mathbf{E}\underline{X}_\infty$ and the element-wise convergence of the covariance matrices $\mathbf{V}_n$ to the corresponding covariance matrix $\mathbf{V}_\infty$. Further, the L^2-convergence implies the corresponding convergence in probability and hence, by bounded convergence $\Phi_{\underline{X}_n}(\underline{\theta}) \to \Phi_{\underline{X}_\infty}(\underline{\theta})$ for each $\underline{\theta} \in \mathbb{R}^d$. Since $\Phi_{\underline{X}_n}(\underline{\theta}) = e^{-\frac{1}{2}(\underline{\theta}, \mathbf{V}_n\underline{\theta})}e^{i(\underline{\theta}, \underline{\mu}_n)}$, for any $n < \infty$, it follows that the same applies for $n = \infty$. It is a well known fact of linear algebra that the element-wise limit $\mathbf{V}_\infty$ of nonnegative definite matrices $\mathbf{V}_n$ is necessarily also nonnegative definite. In view of Definition 3.5.13, we see that the limit $\underline{X}_\infty$ is a Gaussian random vector, whose parameters are the limits of the corresponding parameters of $\underline{X}_n$. □

One of the main reasons for the importance of the multivariate normal distribution is the following CLT (which is the multivariate extension of Proposition 3.1.2).

THEOREM 3.5.16 (MULTIVARIATE CLT). *Let* $\widehat{\underline{S}}_n = n^{-\frac{1}{2}} \sum_{k=1}^{n} (\underline{X}_k - \underline{\mu})$, *where* $\{\underline{X}_k\}$ *are i.i.d. random vectors with finite second moments and such that* $\underline{\mu} = \mathbf{E}\underline{X}_1$. *Then,* $\widehat{\underline{S}}_n \xrightarrow{\mathcal{D}} \underline{G}$, *with* $\underline{G}$ *having the* $\mathcal{N}(\underline{0}, \mathbf{V})$ *distribution and where* $\mathbf{V}$ *is the* $d \times d$*-dimensional covariance matrix of* $\underline{X}_1$.

PROOF. Consider the i.i.d. random vectors $\underline{Y}_k = \underline{X}_k - \underline{\mu}$ each having also the covariance matrix $\mathbf{V}$. Fixing an arbitrary vector $\underline{\theta} \in \mathbb{R}^d$ we proceed to show that $(\underline{\theta}, \widehat{\underline{S}}_n) \xrightarrow{\mathcal{D}} (\underline{\theta}, \underline{G})$, which in view of the Cramér-Wold device completes the proof of the theorem. Indeed, note that $(\underline{\theta}, \widehat{\underline{S}}_n) = n^{-\frac{1}{2}} \sum_{k=1}^{n} Z_k$, where $Z_k = (\underline{\theta}, \underline{Y}_k)$ are i.i.d. $\mathbb{R}$-valued random variables, having zero mean and variance

$$v_{\underline{\theta}} = \mathsf{Var}(Z_1) = \mathbf{E}[(\underline{\theta}, \underline{Y}_1)^2] = (\underline{\theta}, \, \mathbf{E}[\underline{Y}_1 \underline{Y}_1^t] \, \underline{\theta}) = (\underline{\theta}, \mathbf{V}\underline{\theta}) \, .$$

Observing that the CLT of Proposition 3.1.2 thus applies to $(\underline{\theta}, \widehat{\underline{S}}_n)$, it remains only to verify that the resulting limit distribution $\mathcal{N}(0, v_{\underline{\theta}})$ is indeed the law of $(\underline{\theta}, \underline{G})$. To this end note that by Definitions 3.5.4 and 3.5.13, for any $s \in \mathbb{R}$,

$$\Phi_{(\underline{\theta}, \underline{G})}(s) = \Phi_{\underline{G}}(s\underline{\theta}) = e^{-\frac{1}{2}s^2 (\underline{\theta}, \mathbf{V}\underline{\theta})} = e^{-v_{\underline{\theta}} s^2/2} \, ,$$

which is the characteristic function of the $\mathcal{N}(0, v_{\underline{\theta}})$ distribution (see Example 3.3.6). Since the characteristic function uniquely determines the law (see Corollary 3.3.14), we are done. □

Here is an explicit example for which the multivariate CLT applies.

EXAMPLE 3.5.17. *The* simple random walk *on* $\mathbf{Z}^d$ *is* $\underline{S}_n = \sum_{k=1}^{n} \underline{X}_k$ *where* $\underline{X}$, $\underline{X}_k$ *are i.i.d. random vectors such that*

$$\mathbf{P}(\underline{X} = +e_i) = \mathbf{P}(\underline{X} = -e_i) = \frac{1}{2d} \qquad i = 1, \ldots, d,$$

and e_i *is the unit vector in the* i*-th direction,* $i = 1, \ldots, d$. *In this case* $\mathbf{E}\underline{X} = \underline{0}$ *and if* $i \neq j$ *then* $\mathbf{E}X_i X_j = 0$, *resulting with the covariance matrix* $\mathbf{V} = (1/d)\mathbf{I}$ *for the multivariate normal limit in distribution of* $n^{-1/2}\underline{S}_n$.

Building on Lindeberg's CLT for weighted sums of i.i.d. random variables, the following multivariate normal limit is the basis for the convergence of random walks to *Brownian motion*, to which Section 9.2 is devoted.

EXERCISE 3.5.18. *Suppose* $\{\xi_k\}$ *are i.i.d. with* $\mathbf{E}\xi_1 = 0$ *and* $\mathbf{E}\xi_1^2 = 1$. *Consider the random functions* $\widehat{S}_n(t) = n^{-1/2} S(nt)$ *where* $S(t) = \sum_{k=1}^{[t]} \xi_k + (t - [t])\xi_{[t]+1}$ *and* $[t]$ *denotes the integer part of* t.

(a) *Verify that Lindeberg's* CLT *applies for* $\widehat{S}_n = \sum_{k=1}^{n} a_{n,k}\xi_k$ *whenever the non-random* $\{a_{n,k}\}$ *are such that* $r_n = \max\{|a_{n,k}| : k = 1, \ldots, n\} \to 0$ *and* $v_n = \sum_{k=1}^{n} a_{n,k}^2 \to 1$.

(b) *Let* $c(s,t) = \min(s,t)$ *and fixing* $0 = t_0 \leq t_1 < \cdots < t_d$, *denote by* $\mathbf{C}$ *the* $d \times d$ *matrix of entries* $C_{jk} = c(t_j, t_k)$. *Show that for any* $\underline{\theta} \in \mathbb{R}^d$,

$$\sum_{r=1}^{d} (t_r - t_{r-1}) (\sum_{j=1}^{r} \theta_j)^2 = (\underline{\theta}, \mathbf{C}\underline{\theta}) \, ,$$

(c) *Using the Cramér-Wold device deduce that* $(\widehat{S}_n(t_1), \ldots, \widehat{S}_n(t_d)) \xrightarrow{\mathcal{D}} \underline{G}$ *with* $\underline{G}$ *having the* $\mathcal{N}(\underline{0}, \mathbf{C})$ *distribution.*

As we see in the next exercise, there is more to a Gaussian random vector than each coordinate having a normal distribution.

EXERCISE 3.5.19. *Suppose X_1 has a standard normal distribution and S is independent of X_1 and such that $\mathbf{P}(S=1)=\mathbf{P}(S=-1)=1/2$.*

(a) *Check that $X_2 = SX_1$ also has a standard normal distribution.*
(b) *Check that X_1 and X_2 are uncorrelated random variables, each having the standard normal distribution, while $\underline{X}=(X_1,X_2)$ is not a Gaussian random vector and where X_1 and X_2 are not independent variables.*

Motivated by the proof of Proposition 3.5.14 here is an important property of Gaussian random vectors which may also be considered to be an alternative to Definition 3.5.13.

EXERCISE 3.5.20. *A random vector $\underline{X}$ has the multivariate normal distribution if and only if $(\sum_{i=1}^d a_{ji}X_i, j=1,\ldots,m)$ is a Gaussian random vector for any non-random coefficients $a_{11},a_{12},\ldots,a_{md}\in\mathbb{R}$.*

The classical definition of the multivariate normal density applies for a strict subset of the distributions we consider in Definition 3.5.13.

DEFINITION 3.5.21. *We say that $\underline{X}$ has a* non-degenerate *multivariate normal distribution if the matrix $\mathbf{V}$ is invertible, or alternatively, when $\mathbf{V}$ is (strictly) positive definite matrix, that is $(\underline{\theta},\mathbf{V}\underline{\theta})>0$ whenever $\underline{\theta}\neq\underline{0}$.*

We next relate the density of a random vector with its characteristic function, and provide the density for the non-degenerate multivariate normal distribution.

EXERCISE 3.5.22.

(a) *Show that if $\int_{\mathbb{R}^d}|\Phi_{\underline{X}}(\underline{\theta})|d\underline{\theta}<\infty$, then $\underline{X}$ has the bounded continuous probability density function*

$$f_{\underline{X}}(\underline{x})=\frac{1}{(2\pi)^d}\int_{\mathbb{R}^d}e^{-i(\underline{\theta},\underline{x})}\Phi_{\underline{X}}(\underline{\theta})d\underline{\theta}\,. \tag{3.5.5}$$

(b) *Show that a random vector $\underline{X}$ with a non-degenerate multivariate normal distribution $\mathcal{N}(\underline{\mu},\mathbf{V})$ has the probability density function*

$$f_{\underline{X}}(\underline{x})=(2\pi)^{-d/2}(\det\mathbf{V})^{-1/2}\exp\Big(-\frac{1}{2}(\underline{x}-\underline{\mu},\mathbf{V}^{-1}(\underline{x}-\underline{\mu}))\Big)\,.$$

Here is an application to the uniform distribution over the sphere in $\mathbb{R}^n$, as $n\to\infty$.

EXERCISE 3.5.23. *Suppose $\{Y_k\}$ are i.i.d. random variables with $\mathbf{E}Y_1^2=1$ and $\mathbf{E}Y_1=0$. Let $W_n=n^{-1}\sum_{k=1}^n Y_k^2$ and $X_{n,k}=Y_k/\sqrt{W_n}$ for $k=1,\ldots,n$.*

(a) *Noting that $W_n\overset{a.s.}{\to}1$ deduce that $X_{n,1}\xrightarrow{\mathcal{D}}Y_1$.*
(b) *Show that $n^{-1/2}\sum_{k=1}^n X_{n,k}\xrightarrow{\mathcal{D}}G$ whose distribution is $\mathcal{N}(0,1)$.*
(c) *Show that if $\{Y_k\}$ are standard normal random variables, then the random vector $\underline{X}_n=(X_{n,1},\ldots,X_{n,n})$ has the uniform distribution over the surface of the sphere of radius $\sqrt{n}$ in $\mathbb{R}^n$ (i.e., the unique measure supported on this sphere and invariant under orthogonal transformations), and interpret the preceding results for this special case.*

We conclude the section with the following exercise, which is a *multivariate, Lindeberg's type* CLT.

EXERCISE 3.5.24. *Let $\underline{y}^t$ denotes the transpose of the vector $\underline{y} \in \mathbb{R}^d$ and $\|\underline{y}\|$ its Euclidean norm. The independent random vectors $\{\underline{Y}_k\}$ on $\mathbb{R}^d$ are such that $\underline{Y}_k \overset{\mathcal{D}}{=} -\underline{Y}_k$,*

$$\lim_{n\to\infty} \sum_{k=1}^{n} \mathbf{P}(\|\underline{Y}_k\| > \sqrt{n}) = 0,$$

and for some symmetric, (strictly) positive definite matrix $\mathbf{V}$ and any fixed $\varepsilon \in (0,1]$,

$$\lim_{n\to\infty} n^{-1} \sum_{k=1}^{n} \mathbf{E}(\underline{Y}_k \underline{Y}_k^t I_{\|\underline{Y}_k\| \le \varepsilon\sqrt{n}}) = \mathbf{V}.$$

(a) *Let $\underline{T}_n = \sum_{k=1}^n \underline{X}_{n,k}$ for $\underline{X}_{n,k} = n^{-1/2}\underline{Y}_k I_{\|\underline{Y}_k\| \le \sqrt{n}}$. Show that $\underline{T}_n \xrightarrow{\mathcal{D}} \underline{G}$, with $\underline{G}$ having the $\mathcal{N}(\underline{0}, \mathbf{V})$ multivariate normal distribution.*

(b) *Let $\underline{\widehat{S}}_n = n^{-1/2}\sum_{k=1}^n \underline{Y}_k$ and show that $\underline{\widehat{S}}_n \xrightarrow{\mathcal{D}} \underline{G}$.*

(c) *Show that $(\underline{\widehat{S}}_n)^t \mathbf{V}^{-1} \underline{\widehat{S}}_n \xrightarrow{\mathcal{D}} Z$ and identify the law of Z.*

Bibliography

[And1887] Désiré André, *Solution directe du probléme résolu par M. Bertrand*, Comptes Rendus Acad. Sci. Paris, **105**, (1887), 436–437.

[Bil95] Patrick Billingsley, *Probability and measure*, third edition, John Wiley and Sons, 1995.

[Bre92] Leo Breiman, *Probability*, Classics in Applied Mathematics, Society for Industrial and Applied Mathematics, 1992.

[Bry95] Wlodzimierz Bryc, *The normal distribution*, Springer-Verlag, 1995.

[Doo53] Joseph Doob, *Stochastic processes*, Wiley, 1953.

[Dud89] Richard Dudley, *Real analysis and probability*, Chapman and Hall, 1989.

[Dur03] Rick Durrett, *Probability: Theory and Examples*, third edition, Thomson, Brooks/Cole, 2003.

[DKW56] Aryeh Dvortzky, Jack Kiefer and Jacob Wolfowitz, *Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator*, Ann. Math. Stat., **27**, (1956), 642–669.

[Dyn65] Eugene Dynkin, *Markov processes*, volumes 1-2, Springer-Verlag, 1965.

[Fel71] William Feller, *An introduction to probability theory and its applications, volume II*, second edition, John Wiley and sons, 1971.

[Fel68] William Feller, *An introduction to probability theory and its applications, volume I*, third edition, John Wiley and sons, 1968.

[Fre71] David Freedman, *Brownian motion and diffusion*, Holden-Day, 1971.

[GS01] Geoffrey Grimmett and David Stirzaker, *Probability and random processes*, 3rd ed., Oxford University Press, 2001.

[Hun56] Gilbert Hunt, *Some theorems concerning Brownian motion*, Trans. Amer. Math. Soc., **81**, (1956), 294–319.

[KaS97] Ioannis Karatzas and Steven E. Shreve, *Brownian motion and stochastic calculus*, Springer Verlag, third edition, 1997.

[Lev37] Paul Lévy, *Théorie de l'addition des variables aléatoires*, Gauthier-Villars, Paris, (1937).

[Lev39] Paul Lévy, *Sur certains processus stochastiques homogénes*, Compositio Math., **7**, (1939), 283–339.

[KT75] Samuel Karlin and Howard M. Taylor, *A first course in stochastic processes*, 2nd ed., Academic Press, 1975.

[Mas90] Pascal Massart, *The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality*, Ann. Probab. **18**, (1990), 1269–1283.

[MP09] Peter Mörters and Yuval Peres, *Brownian motion*, Cambridge University Press, 2010.

[Num84] Esa Nummelin, *General irreducible Markov chains and non-negative operators*, Cambridge University Press, 1984.

[Oks03] Bernt Oksendal, *Stochastic differential equations: An introduction with applications*, 6th ed., Universitext, Springer Verlag, 2003.

[PWZ33] Raymond E.A.C. Paley, Norbert Wiener and Antoni Zygmund, *Note on random functions*, Math. Z. **37**, (1933), 647–668.

[Pit56] E. J. G. Pitman, *On the derivative of a characteristic function at the origin*, Ann. Math. Stat. **27** (1956), 1156–1160.

[SW86] Galen R. Shorak and Jon A. Wellner, *Empirical processes with applications to statistics*, Wiley, 1986.

[Str93] Daniel W. Stroock, *Probability theory: an analytic view*, Cambridge university press, 1993.

[Wil91] David Williams, *Probability with martingales*, Cambridge university press, 1991.