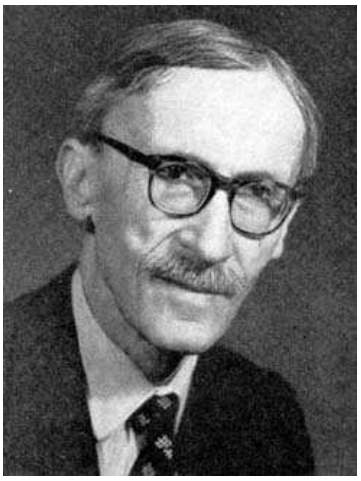
Probability Theory I

Theory of Measure and Independence



A. Kolmogorov



P. Lévy

Tommi Sottinen

tommi.sottinen@uwasa,fi
www.uwasa.fi/∼tsottine/

October 19, 2026

Preface

> For a 20-year-old, the question is what speciality should I learn? You should study very basic things that have a long shelf life, mathematics, physics, basic computer science, applied mathematics. Those are things that would be necessary to understand and develop the next generation of AI systems.
>
> — Yann LeCun

> There is nothing more practical than a theory which works.
>
> — Kurt Lewin

> Explain the theory in complete detail, starting from the very basics, and in a language the reader can understand
>
> — Michel Talagrand

> Mathematics is being lazy. Mathematics is letting the principles do the work for you so that you do not have to do the work for yourself.
>
> — George Pólya

> Statistics is the Grammar of Science.
>
> — Karl Pearson

> Teach principles not formulas.
>
> — Richard Feynman

Nobody reads the preface! So, I can write here pretty much anything. If you happen to read this preface here is a bonus for you. Send me a mathematical meme (picture) to my email `tommi.sottinen@uwasa.fi`, and you get one extra exercise point for your final grade if you remember to state your name and student number in the email. Also, you have to title your email MATH3010-meme.

These are lecture notes for University of Vaasa course MATH3010 Probability Theory I. The course is 5 ECTS consisting of 36 h lectures and

12 h exercises. In these notes we begin with the elements and end with the great limit theorems of probability. In principle we prove everything, but some proves are left as exercises. Some great theorems are left as master's theses. In the proofs we are very pedantic in the beginning of the notes and more relaxed towards the end of the notes.

The lectures are based on the Finnish language lectures T. Sottinen: Todennäköisyysteoria, 2006. Those lectures are in turn based on the following sources

1. Ya. Sinai *Probability Theory: An Introductory Course*, 1992.
2. G. Elfving ja P. Tuominen *Todennäköisyyslaskenta II*, 1990.
3. J. Jacod and P. Protter *Probability Essentials*, 2004.
4. D. Williams *Probability with Martingales*, 1991.
5. A. Shiryaev *Probability*, 1984.
6. O. Kallenberg *Foundations of modern probability*, 2002.

The sources are listed in the order from the easiest to the most difficult.

Vaasa October 19, 2026 T. S.

Contents

1 Introduction **5**
- 1.1 Sets and Limits . . . 5
- 1.2 Coin Tossing . . . 7

2 Probability Space **11**
- 2.1 σ-Algebras . . . 11
- 2.2 Probability Measure and Space . . . 13
- 2.3 Extension Theorems . . . 16
- 2.4 Measurability and Nonmeasurability . . . 19

3 Random Variables **22**
- 3.1 Definition as Measurable Function . . . 22
- 3.2 Limits and Transformations . . . 23
- 3.3 Distributions . . . 25

4 Expectation **27**
- 4.1 Definition as Measure Integral . . . 27
- 4.2 Limits and Expectations . . . 35
- 4.3 Expectations of Transformations . . . 37
- 4.4 Inequalities . . . 39

5 Independence **46**
- 5.1 Independent σ-algebras, Random Variables and Events . . . 46
- 5.2 Borel's and Cantelli's Lemmas . . . 48
- 5.3 Product Spaces and Fubini's Theorem . . . 52
- 5.4 Infinite Product Spaces . . . 55
- 5.5 Sums of Independent Random Variables . . . 56

6 Characteristic Functions **63**
- 6.1 Complex Numbers and Functions . . . 63

6.2 Complex-Valued Random Variables 67
6.3 Characteristic Function and its Properties 68

7 Modes of Convergence **78**
7.1 Almost Sure Convergence 78
7.2 L^p Convergence 81
7.3 Convergence in Probability 82
7.4 Convergence in Distribution 84
7.5 Relationships between the Modes of Convergences 86

8 Limit Theorems **89**
8.1 Laws of Large Numbers 89
8.2 Large Deviations Principle 98
8.3 Central Limit Theorem 100

Chapter 1

Introduction

We begin by recalling some notation and concepts that the students are probably familiar with in Section 1.1. After this we present the "great limit theorems of probability" in Section 1.2 in the context of coin-tossing.

1.1 Sets and Limits

The natural numbers are $\mathbb{N} = \{1, 2, 3, \ldots\}$, the integers are $\mathbb{Z} = \{\ldots, -3, -2, -1, 0, 1, 2, 3, \ldots\}$, the rational numbers are $\mathbb{Q} = \{p/q; p \in \mathbb{Z}, q \in \mathbb{N}\}$. The real numbers $\mathbb{R}$ are extension of rational numbers that are *Dedekind complete*: each $S \subset \mathbb{R}$ that is bounded above has least upper bound, the so-called *supremum*. Similarly, each $S \subset \mathbb{R}$ that is bounded below has greatest lower bound, the so-called *infinum*. We denote by $\mathbb{C} = \{x + \mathrm{i}y; x, y \in \mathbb{R}\}$ the complex numbers and finally $\mathbb{R}_+ := \{x \in \mathbb{R}\,;\, x \geq 0\}$.

A set N is *countable* if there exist an injection $f : N \to \mathbb{N}$. Finite sets and the sets $\mathbb{N}$, $\mathbb{Z}$ and $\mathbb{Q}$ are countable. The sets $\mathbb{R}$ and $\mathbb{C}$ are not countable; they are *uncountable*

Sample space is Ω. From this space Lady Fortuna chooses an element in random experiments.

The *complement* A^c of a subset $A \subset \Omega$ is $A^c := \{\omega \in \Omega\,;\, \omega \notin A\}$.

The *subtraction* of the set B from the set A the set $A \setminus B = \{\omega \in A; \omega \notin B\}$.

The totality of all subsets of E is its *power set* $\mathrm{pow}\,(E)$: $A \in \mathrm{pow}\,(E)$ if and only if $A \subset E$.

Let J be an index set and $A_j \subset \Omega$ for all $j \in J$. Then we denote

$$\begin{aligned} \bigcup_{j\in J} A_j &:= \{\omega \in \Omega\,;\, \omega \in A_j \text{ for some } j \in J\}\,, \\ \bigcap_{j\in J} A_j &:= \{\omega \in \Omega\,;\, \omega \in A_j \text{ for all } j \in J\}\,. \end{aligned}$$

In the language of probability, $\cap_{j\in J} A_j$ means "all the A_j's occur" and $\cup_{j\in J} A_j$ means "at least one A_j occurs". We also interpret $\cup_{j\in\varnothing} A_j = \varnothing$ and $\cap_{j\in\varnothing} A_j = \Omega$. The index set J need not be countable; for example $\mathbb{R} = \cup_{x\in\mathbb{R}}\{x\}$.

We also write for the union $\cup_{n=1}^{\infty} := \cup_{n\in\mathbb{N}}$ and $\cup_{k=1}^{n} := \cup_{k\in\{1,2,\ldots,n\}}$, $\cup_{k=n}^{\infty} := \cup_{k\geq n} := \cup_{\{k\in\mathbb{N}:k\geq n\}}$, and similarly for the intersection $\cap$.

The collection $A_j \subset \Omega$, $j \in J$ satisfies *De Morgan's laws*

$$\left(\bigcup_{j\in J} A_j\right)^c = \bigcap_{j\in J} A_j^c \quad \text{and} \quad \left(\bigcap_{j\in J} A_j\right)^c = \bigcup_{j\in J} A_j^c.$$

We will also need later the *Cartesian product* of sets: If A and B are sets, their Cartesian product is

$$A \times B = \{(a,b)\,;\, a \in A, b \in B\}.$$

Genaral cartesian products A^N are defined similarly. Indeed A^N is the set of functons $f\colon N \to A$.

For a function (random variable) $X : \Omega \to \mathbb{R}$ we denote

$$\{X \in B\} := X^{-1}B := \{\omega \in \Omega\,;\, X(\omega) \in B\}\,.$$

Here $\{X \in B\} \subset \Omega$ and $B \subset \mathbb{R}$.

Extended reals are $\bar{\mathbb{R}} := \mathbb{R} \cup \{-\infty, \infty\}$.

For a sequence of real numbers $(x_n)_{n\in\mathbb{N}}$ its lower and upper limits are

$$\liminf_{n\to\infty} x_n := \lim_{n\to\infty} \inf_{k\geq n} x_k \quad \text{and} \quad \limsup_{n\to\infty} x_n := \lim_{n\to\infty} \sup_{k\geq n} x_k.$$

$\liminf_{n\to\infty} x_n$ and $\limsup_{n\to\infty} x_n$ exist always as extended real numbers (exercise 1.2). Moreover, $x_n \to x$ if and only if $\liminf_{n\to\infty} x_n = \limsup_{n\to\infty} x_n = x$ (exercise 1.3). It is clear that $\liminf_{n\to\infty} x_n \leq \limsup_{n\to\infty} x_n$ for all sequences $(x_n)_{n\in\mathbb{N}}$. This is why they are called lower and upper limits

For a sequence of sets $(A_n)_{n\in\mathbb{N}}$ their *upper and lower limits* are

$$\liminf_{n\to\infty} A_n := \bigcup_{n=1}^{\infty}\bigcap_{k=n}^{\infty} A_k \quad \text{and} \quad \limsup_{n\to\infty} A_n := \bigcap_{n=1}^{\infty}\bigcup_{k=n}^{\infty} A_k.$$

In the language of probability $\limsup_{n\to\infty} A_n$ means "A_n occurs *infinitely often*" and $\liminf_{n\to\infty} A_n$ means "A_n occurs always *eventually*". Therefore we also use notations $\{A_n \text{ i.o}\} := \limsup_{n\to\infty} A_n$ and $\{A_n \text{ ev.}\} := \liminf_{n\to\infty} A_n$.

If $x_n \to x$ and $(x_n)_{n\in\mathbb{N}}$ is increasing, we denote $x_n \uparrow x$. If $x_n \to x$ and $(x_n)_{n\in\mathbb{N}}$ is decreasing, we denote $x_n \downarrow x$. In fact: if $(x_n)_{n\in\mathbb{N}}$ is increasing, then $x_n \uparrow \sup_{n\in\mathbb{N}} x_n \in \bar{\mathbb{R}}$, and if $(x_n)_{n\in\mathbb{N}}$ is decreasing, then $x_n \downarrow \inf_{n\in\mathbb{N}} x_n \in \bar{\mathbb{R}}$.

A set of sequences $(A_n)_{n\in\mathbb{N}}$ is increasing if $A_n \subset A_{n+1}$, and decreasing if $A_n \supset A_{n+1}$. Notation $A_n \uparrow A$ means that $(A_n)_{n\in\mathbb{N}}$ is increasing and $A = \cup_{n=1}^{\infty} A_n$. Notation $A_n \downarrow A$ means that $(A_n)_{n\in\mathbb{N}}$ is decreasing and $A = \cap_{n=1}^{\infty} A_n$. If $\limsup_{n\to\infty} A_n = \liminf_{n\to\infty} A_n = A$ we say that $(A_n)_{n\in\mathbb{N}}$ converges towards the set A. In this case we denote $A_n \to A$ or $\lim_{n\to\infty} A_n = A$.

The *indicator* of a set A is

$$\mathbf{1}_A(\omega) \;=\; \begin{cases} 1, & \text{if } \omega \in A, \\ 0, & \text{otherwise.} \end{cases}$$

Then $A_n \to A$ if and only if $\mathbf{1}_{A_n}(\omega) \to \mathbf{1}_A(\omega)$ for all $\omega \in \Omega$ (Exercise 1.6).

1.2 Coin Tossing

We present the great limit theorems of probability here informally in the context of fair coin tossing. More formal and general presentation is given in Chapter 8.

Let

$$X_n \;:=\; \begin{cases} 1, & \text{if } n^{\text{th}} \text{ toss is head,} \\ 0, & \text{if } n^{\text{th}} \text{ toss is tail.} \end{cases}$$

We assume that the coin tosses are independent and that the coin is fair. This means that for all $(x_1, \ldots, x_n) \in \{0,1\}^n$, $n \in \mathbb{N}$, we have

$$\mathbf{P}[X_1 = x_1, \ldots, X_n = x_n] \;=\; \frac{1}{2^n}.$$

Let

$$Y_n \;:=\; \frac{1}{n}\sum_{k=1}^{n} X_k$$

be the *relative frequency* of heads in a series of n coin tosses.

When $n \to \infty$, the probabilities $\mathbf{P}[Y_n = y]$, $y \in [0,1]$, satisfy informally

$$\mathbf{P}[Y_n = y] \simeq \mathbf{1}_{\{y=\frac{1}{2}\}}, \tag{1.2.1}$$

$$\simeq \mathrm{e}^{-n\left(y\ln y+(1-y)\ln(1-y)+\ln 2\right)}, \tag{1.2.2}$$

$$\simeq \frac{1}{\sqrt{\pi\frac{n}{2}}}\mathrm{e}^{-n\left(2(y-\frac{1}{2})^2\right)}. \tag{1.2.3}$$

The estimate (1.2.1) is the (weak) *law of large number*, the estimate (1.2.2) is the *large deviations limit* and the estimate (1.2.3) is the *central limit theorem*. These estimates can be considered the most important results of probability theory.

Next we argue the estimates (1.2.1)–(1.2.3) heuristically. Let us first note that

$$y \mapsto I(y) := y\ln y + (1-y)\ln(1-y) + \ln 2$$

is positive and attains its minimum (which is zero) at point $y = 1/2$. Therefore, the estimate (1.2.1) follows from the estimate (1.2.2). Let us then argue the estimate (1.2.2). Our tool is the *Stirling's formula*:

$$n! \simeq \sqrt{2\pi n}\, n^n \mathrm{e}^{-n}, \qquad \text{as } n \to \infty.$$

Since nY_n is binomially distributed with parameters n and $1/2$, we obtain (for $ny \in \mathbb{N}$)

$$\mathbf{P}[Y_n = y] = \mathbf{P}[nY_n = ny] = \binom{n}{ny}2^{-n} = \frac{n!\,2^{-n}}{(ny)!(n-ny)!}.$$

By applying the Stirling's formula to this, we obtain

$$\begin{aligned}\mathbf{P}[Y_n = y] &\simeq \frac{\sqrt{2\pi n}n^n\mathrm{e}^{-n}\,2^{-n}}{\sqrt{2\pi ny}(ny)^{ny}\mathrm{e}^{-ny}\sqrt{2\pi(n-ny)}(n-ny)^{n-ny}\mathrm{e}^{-(n-ny)}} \\ &= \frac{1}{\sqrt{2\pi ny(1-y)}}y^{-ny}(1-y)^{-n(1-y)}2^{-n} \\ &= \frac{1}{\sqrt{2\pi ny(1-y)}}e^{-n(y\ln y+(1-y)\ln(1-y)+\ln 2)}. \qquad (1.2.4)\end{aligned}$$

The estimate (1.2.2) follows from the estimate (1.2.4) by removing the slowly varying term

$$\frac{1}{\sqrt{2\pi ny(1-y)}}\mathrm{e}^{-n(y\ln y+(1-y)\ln(1-y)+\ln 2)} \simeq \mathrm{e}^{-n(y\ln y+(1-y)\ln(1-y)+\ln 2)}.$$

Let us then argue the estimate (1.2.3). By the estimate (1.2.1) we may assume that

$$y \simeq \frac{1}{2}.$$

The estimate (1.2.3) follows by substituting $y(1-y) \simeq 1/4$ in the slowly varying term of the estimate (1.2.4) and by using the second order Taylor's expansion

$$I(y) \simeq I\left(\frac{1}{2}\right) + I'\left(\frac{1}{2}\right)\left(y-\frac{1}{2}\right) + \frac{1}{2}I''\left(\frac{1}{2}\right)\left(y-\frac{1}{2}\right)^2$$

in the exponent of the estimate (1.2.4). Indeed, then

$$\begin{aligned} I'(y) &= \ln y - \ln(1-y), \\ I''(y) &= \frac{1}{y} + \frac{1}{1-y}, \end{aligned}$$

and

$$I\left(\frac{1}{2}\right) = 0, \qquad I'\left(\frac{1}{2}\right) = 0, \quad \text{sekä} \quad I''\left(\frac{1}{2}\right) = 4.$$

Therefore the Taylor estimate is

$$I(y) \simeq 2\left(y-\frac{1}{2}\right)^2.$$

Let us note that the large deviations estimate (1.2.2) works better when the relative frequence y is far away from its law of lage numbers estimate $y \simeq 1/2$. The central limit estimate (1.2.3) works better when the relative frequency y is close to its law of large numbers (1.2.1) estimate $y \simeq 1/2$. The different domains of the estimates (1.2.2) and (1.2.3) are clear once one notices how they were derived: the estimate (1.2.2) was derived by discarding a slowly varying term, the estimate (1.2.3) was derived by using a Taylor expansion around the typical value $y \simeq 1/2$.

The (weak) law of large numbers estimate (1.2.1) will later in these notes be stated as the weak law of large number : $\lim_{n\to\infty} \mathbf{P}[|Y_n - 1/2| > \varepsilon] = 0$ for all $\varepsilon > 0$. The *strong law of large numbers* states that $\mathbf{P}[\lim_{n\to\infty} Y_n = 1/2] = 1$. In this case we have to interpret the set $\{\lim_{n\to\infty} Y_n = 1/2\}$ as an event in the probability space that consists of infinite sequences of coin tosses. This is not trivial and here we need probability *theory*. The large deviations estimate (1.2.2) is based on the *large deviations principle*, and the central estimate (1.2.3) based on the *central limit theorems*. All these results will be shown rigorously in Chapter 8.

Exercises

1.1. Let J is an arbitrary index set. Prove the De Morgan's laws $(\cup_{j\in J} A_j)^c = \cap_{j\in J} A_j^c$ and $(\cap_{j\in J} A_j)^c = \cup_{j\in J} A_j^c$.

1.2. Let $(x_n)_{n\in\mathbb{N}}$ be a sequence of real number. Show that $\liminf_{n\to\infty} x_n$ and $\limsup_{n\to\infty} x_n$ exist always as extended real numbers.

1.3. Let $(x_n)_{n\in\mathbb{N}}$ be a sequence of real numbers. Show that $(x_n)_{n\in\mathbb{N}}$ converges if and only if $\liminf_{n\to\infty} x_n = \limsup_{n\to\infty} x_n$, and in this case $\lim_{n\to\infty} x_n = \liminf_{n\to\infty} x_n = \limsup_{n\to\infty} x_n$.

1.4. Let A be a set. For each $n \in \mathbb{N}$ we set $A_{2n-1} := A$ and $A_{2n} := A^c$. What are hte sets $\liminf_{n\to\infty} A_n$ and $\limsup_{n\to\infty} A_n$? Does the set sequence $(A_n)_{n\in\mathbb{N}}$ converge?

1.5. Let $(A_n)_{n\in\mathbb{N}}$ be a sequence of sets. Show that $\liminf_{n\to\infty} A_n \subset \limsup_{n\to\infty} A_n$.

1.6. Let $(A_n)_{n\in\mathbb{N}}$ be a sequence of events in Ω. Show that $A_n \to A$ if and only if $\mathbf{1}_{A_n}(\omega) \to \mathbf{1}_A(\omega)$ kaikilla $\omega \in \Omega$.

1.7. (a) State and (b) prove the estimates (1.2.1)–(1.2.3) of Section 1.2 for a biased coin.

This means that we assume that for all $(x_1, \ldots, x_n) \in \{0,1\}^n$, $n \in \mathbb{N}$,

$$\begin{aligned} \mathbf{P}[X_1 = x_1, \ldots, X_n = x_n] &= p^{|\{x_i\,;\, x_i=1, i\le n\}|}(1-p)^{|\{x_1\,;\, x_i=0, i\le n\}|} \\ &= p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i} \end{aligned}$$

for some $p \in [0,1]$. So the coin is biased but the tosses are still independent.

In the above $|A|$ means the number of elements in the set A.

Chapter 2

Probability Space

Unfortunately, in important cases we cannot consider all the subsets of the sample space Ω as events, or the so-called "measurable" sets. This means that we cannot assing for all $A \subset \Omega$ a probability measure $\mathbf{P}[A]$ in a consistent way. We have to restrict ourselves to a proper subset of pow (Ω)*. Concretely, we will see this in Example 2.4.1. On the other hand "common sense" states that if A admits probability then A^c also admits probability and $\mathbf{P}[A^c] = 1 - \mathbf{P}[A]$. The same common sense states that if only either A or B can occur and A and B admit probabilities, then $A \cup B$ admits probability: $\mathbf{P}[A \cup B] = \mathbf{P}[A] + \mathbf{P}[B]$. Finally, it has been observed that if $A_n \downarrow A$ and A_n's admit probabilities, then it makes sense to assign the probability $\mathbf{P}[A] = \lim_{n\to\infty} \mathbf{P}[A_n]$ for the set A. Indeed, without this continuity of probabilities the law of large numbers would not hold. It follows from the discussion above that the set of events, i.e. the set of measurable sets, i.e. the set of sets that we can assign probability to is what is called a σ-algebra.*

2.1 σ-Algebras

2.1.1 Definition. A collection $\mathcal{F}$ of subsets of the sample space Ω is called a *σ-algebra* if

(i) $\Omega \in \mathcal{F}$,
(ii) if $A \in \mathcal{F}$, the $A^c \in \mathcal{F}$,
(iii) if $A_n \in \mathcal{F}$ for all $n \in \mathbb{N}$, then $\cap_{n=1}^{\infty} A_n \in \mathcal{F}$.

In Definition 2.1.1 the condition (i) can be replaced with the condition

(i') $\varnothing \in \mathcal{F}$,

because condition (ii) implies that $\varnothing^c = \Omega \in \mathcal{F}$. The condition (iii) can be replaced with the condition

(iii') if $A_n \in \mathcal{F}$ for all $n \in \mathbb{N}$, the $\cup_{n=1}^{\infty} A_n \in \mathcal{F}$,

because $\cup_{n=1}^{\infty} A_n = (\cap_{n=1}^{\infty} A_n^c)^c$ and $(\cap_{n=1}^{\infty} A_n^c)^c \in \mathcal{F}$ due to condition (ii). Finally, the condition (iii) can be split into two conditions:

(iiia) If $A, B \in \mathcal{F}$ then $A \cap B \in \mathcal{F}$, and
(iiib) If $A_n \in \mathcal{F}$ for all $n \in \mathbb{N}$ and $A_n \downarrow A$, then $A \in \mathcal{F}$.

Similar conditions can we stated in terms of unions and $A_n \uparrow A$.

2.1.2 Example. (i) The full σ-algebra $\mathrm{pow}\,(\Omega)$ is a σ-algebra.
(ii) The trivial σ-algebra $\{\Omega, \varnothing\}$ is a σ-algebra.
(iii) Let $\Omega = \{1, 2, 3\}$. Then $\{\Omega, \{2, 3\}, \{1\}, \varnothing\}$ is a σ-algebra.
(iv) Let $\Omega = \{1, 2, 3\}$. The collection $\{\Omega, \{1\}, \{2\}, \{3\}, \varnothing\}$ is not a σ-algebra.
(v) Let $A \subset \Omega$. Then $\{\Omega, A, A^c, \varnothing\}$ is a σ-algebra (generated by the set A).
(vi) The collection of open intervals of $\mathbb{R}$ is not a σ-algebra, because for example $\{0\} = \cap_{n=1}^{\infty}(-\frac{1}{n}, \frac{1}{n})$ is not an open interval.

2.1.3 Lemma. *If $\mathcal{F}_j$, $j \in J$, are σ-algebras, then $\cap_{j\in J}\mathcal{F}_j$ is also a σ-algebra.*

Proof. The proof is basically trivial. We just check the conditions of a σ-algebra: (i) $\Omega \in \cap_{j\in J}\mathcal{F}_j$, because $\Omega \in \mathcal{F}_j$ for all $j \in J$. (ii) If $A \in \cap_{j\in J}\mathcal{F}_j$, then $A \in \mathcal{F}_j$ for all $j \in J$. Therefore $A^c \in \mathcal{F}_j$, because $\mathcal{F}_j$ is a σ-algebra. Since this holds for all $j \in J$ we conclude that $A^c \in \cap_{j\in J}\mathcal{F}_j$. (iii) If A_n, $n \in \mathbb{N}$, belong to $\cap_{j\in J}\mathcal{F}_j$, then they belong to the σ-algebras $\mathcal{F}_j$ for all $j \in J$. Therefore $\cap_{n\in\mathbb{N}} A_n \in \mathcal{F}_j$ for all $j \in J$, since the collections $\mathcal{F}_j$, $j \in J$, are σ-algebras. Consequently $\cap_{n\in\mathbb{N}} A_n \in \cap_{j\in J}\mathcal{F}_j$. □

Lemma 2.1.3 makes the following definition possible.

2.1.4 Definition. The σ-algebra $\sigma(\mathcal{C})$ generated by a collection $\mathcal{C}$ is the smallest σ-algebra that contains the collection $\mathcal{C}$:

$$\sigma(\mathcal{C}) \;:=\; \bigcap \big\{\mathcal{F}\,;\, \mathcal{F} \text{ is a } \sigma\text{-algebra and } \mathcal{C} \subset \mathcal{F}\big\}.$$

In Example 2.1.2(v) the σ-algebra is generated by the collection $\{A\}$, or in other words the set A. In Example 2.1.2(iii) the σ-algebra is generated by the collection $\{\{1\}\}$, or in other words the set $\{1\}$.

2.1.5 Lemma. *If $\mathcal{C}' \subset \mathcal{C}$, then $\sigma(\mathcal{C}') \subset \sigma(\mathcal{C})$.*

Proof. On the one hand $\sigma(\mathcal{C})$ is a σ-algebra, that contains the collection $\mathcal{C}'$. On the other hand $\sigma(\mathcal{C}')$ is the smallest σ-algebra, that contains the collection $\mathcal{C}'$. Consequently, $\sigma(\mathcal{C}') \subset \sigma(\mathcal{C})$. □

Recall, that a subset $U \subset \mathbb{R}$ is *open*, if U's every point x has a positive distance from the complement of U. In other words for all $x \in U$ we have

$$\operatorname{dist}(x, U^c) \;:=\; \inf\{|x-y|\,;\, y \in U^c\} \;>\; 0.$$

In fact, open sets of $\mathbb{R}$ are countable unions of open intervals.

2.1.6 Definition. The *Borel's σ-algebra* $\mathcal{B}$ is a σ-algebra that is generated by the open sets of $\mathbb{R}$.

2.1.7 Lemma. *Each of the following collections generate $\mathcal{B}$:*

(i) $[a,b],\ a \le b,\ a,b \in \mathbb{R},$
(ii) $[a,b],\ a < b,\ a,b \in \mathbb{R},$
(iii) $(a,b],\ a \le b,\ a,b \in \mathbb{R},$
(iv) $(a,b],\ a < b,\ a,b \in \mathbb{R},$
(v) $(a,b),\ a < b,\ a,b \in \mathbb{R},$
(vi) $(-\infty,b],\ b \in \mathbb{R},$
(vii) $(-\infty,b),\ b \in \mathbb{R},$
(viii) $[a,\infty),\ a \in \mathbb{R},$
(ix) $(a,\infty),\ a \in \mathbb{R},$
(x) *any of the above, but so that* $a,b \in \mathbb{Q}$.

Proof. We only show part (v) and leave the rest as Exercise 2.3. The important observation is that every open set in $\mathbb{R}$ is a countable union of open intervals. Then the result follows from Definition 2.1.1 condition (iii'). Let $B \subset \mathbb{R}$ be open and let $B' = B \cap \mathbb{Q}$. Then

$$B \;=\; \bigcup_{q \in B'} (q - \varepsilon_q, q + \varepsilon_q),$$

where $\varepsilon_q = \operatorname{dist}(q, B^c) > 0$. □

2.2 Probability Measure and Space

2.2.1 Definition. The pair $(\Omega, \mathcal{F})$, where $\mathcal{F}$ is a σ-algebra of the sample set Ω is called a *measurable space*. The sets $A \in \mathcal{F}$ are *measurable sets* or *events*.

2.2.2 Definition. A *probability measure* is a set function $\mathbf{P} : \mathcal{F} \to [0,1]$, that satisfies

(i) $\mathbf{P}[\Omega] = 1$,
(ii) if the events $A_n \in \mathcal{F}$, $n \in \mathbb{N}$, are disjoint, then

$$\mathbf{P}\left[\bigcup_{n=1}^{\infty} A_n\right] = \sum_{n=1}^{\infty} \mathbf{P}[A_n].$$

Probability measure is a special case of a measure. In general, a *measure* is a set function $\mu : \mathcal{F} \to \mathbb{R}_+ \cup \{\infty\}$, that satisfies $\mu[\varnothing] = 0$ and

$$\mu\left[\bigcup_{n=1}^{\infty} A_n\right] = \sum_{n=1}^{\infty} \mu[A_n],$$

when the measurable sets $A_n \in \mathcal{F}$, $n \in \mathbb{N}$, are disjoint. If $\mu[\Omega] < \infty$, we say that the measure μ is *bounded* or *finite*.

The following definition combined with Definition 2.1.1 and Definition 2.2.2 is *Kolmogorov's axiomatization* of probability. It should be noted that the definition is purely technical and not at all philosophical: it does not talk about probability "an sich" but only tells us how it behaves.

2.2.3 Definition. A triplet $(\Omega, \mathcal{F}, \mathbf{P})$ is a *probability space*.

2.2.4 Example. Let $|A|$ denote the number of elements in the set A.

(i) The probability space for rolling a dice is $\Omega = \{1, \ldots, 6\}$, $\mathcal{F} = \mathrm{pow}\,(\Omega)$ ja $\mathbf{P}[A] = |A|/6$ for all $A \in \mathcal{F}$.
(ii) Tossing a fair coin n times corresponds to the probability space $\Omega = \{0,1\}^n$, $\mathcal{F} = \mathrm{pow}\,(\Omega)$ ja $\mathbf{P}[A] = |A|/2^n$ for all $A \in \mathcal{F}$.
(iii) Tossing a fair coin infinite times corresponds to the probability space $\Omega = \{0,1\}^\infty$. This means that $\omega \in \Omega$ is an infinite sequence of 0's and 1's:

$$\omega = (\omega_1, \omega_2, \ldots, \omega_n, \omega_{n+1}, \ldots).$$

The probability measure is generated by the projections $\mathbf{P}[\{\mathrm{proj}_n \omega\}] = 1/2^n$, where $\mathrm{proj}_n(\omega_1, \omega_2, \ldots) = (\omega_1, \ldots, \omega_n)$. The σ-algebra is $\mathcal{F} = \sigma(\mathrm{proj}_n A\,;\, A \in \mathrm{pow}\,(\Omega), n \in \mathbb{N})$. The sets $\mathrm{proj}_n A$ are called *cylinder sets*. In this case it is impossible to choose $\mathcal{F} = \mathrm{pow}\,(\Omega)$. We will see this later in Example 2.4.1 by interpreting the space $\{0,1\}^\infty$ as $[0,1] \subset \mathbb{R}$ by using the real numbers' binary digital expansions.

The property (ii) of probability measure is called *σ-additivity*. Finite *additivity* means that for a set function $\mathbf{P}$ we have that $\mathbf{P}[A_1 \cup A_2] = \mathbf{P}[A_1] + \mathbf{P}[A_2]$, whenever A_1 and A_2 are disjoint.

The next lemma states that σ-additivity is additivity combined with the monotone continuity.

2.2.5 Lemma. *Let $\mathbf{P} : \mathcal{F} \to [0,1]$ be an additive set function with $\mathbf{P}[\Omega] = 1$. Then the following are equivalent*

(i) $\mathbf{P}$ *ia σ-additive,*
(ii) *if $A_n, A \in \mathcal{F}$, $n \in \mathbb{N}$, and $A_n \downarrow A$, then $\mathbf{P}[A_n] \downarrow \mathbf{P}[A]$,*
(iii) *if $A_n, A \in \mathcal{F}$, $n \in \mathbb{N}$, and $A_n \uparrow A$, then $\mathbf{P}[A_n] \uparrow \mathbf{P}[A]$.*

Proof. (ii) $\Leftrightarrow$ (iii): $A_n \downarrow A$ if and only if $A_n^c \uparrow A^c$. The claim follows therefore from the De Morgan's laws and the additivity of $\mathbf{P}$.

(iii) $\Leftrightarrow$ (i): Let $A_n \uparrow A$. We set $B_1 = A_1$ and $B_n := A_n \setminus A_{n-1}$ for $n \geq 2$. Then the B_n's are disjoint and $\cup_{k=1}^n B_k = A_n$. Consequently,

$$\mathbf{P}[A_n] \;=\; \mathbf{P}\left[\bigcup_{k=1}^{n} B_k\right] \;=\; \sum_{k=1}^{n} \mathbf{P}[B_k].$$

Because $A = \cup_{n=1}^{\infty} A_n = \cup_{n=1}^{\infty} B_n$, we see that (iii) $\Leftrightarrow$ (i). □

The notation $(A_n)_{n\in\mathbb{N}} \subset \mathcal{F}$ means that $A_n \in \mathcal{F}$ for all $n \in \mathbb{N}$.

The next results shows that probability is continuous

2.2.6 Theorem. *Let $(A_n)_{n\in\mathbb{N}} \subset \mathcal{F}$. If $A_n \to A$, then $A \in \mathcal{F}$ and $\mathbf{P}[A_n] \to \mathbf{P}[A]$.*

Proof. We first show that $A \in \mathcal{F}$. Because $A_n \to A$, we have in particular that $A = \liminf_{n\to\infty} A_n$. But $\liminf_{n\to\infty} A_n \in \mathcal{F}$, because it is constructed by using countable set operations on A_n, $n \in \mathbb{N}$.

Let us then prove the continuity. Let $A_n^+ = \cap_{k=n}^{\infty} A_k$ and $A_n^- = \cup_{k=n}^{\infty} A_k$. Then the sequence $(A_n^+)_{n\in\mathbb{N}}$ is increasing and the sequence on $(A_n^-)_{n\in\mathbb{N}}$ is decreasing. Moreover $\liminf_{n\to\infty} A_n = \lim_{n\to\infty} A_n^+$ ja $\limsup_{n\to\infty} A_n = \lim_{n\to\infty} A_n^-$. Since $A_n \to A$, we have $A_n^+ \uparrow A$ and $A_n^- \downarrow A$. By noticing that $A_n^+ \subset A_n \subset A_n^-$ and by using Lemma 2.2.5 we obtain that

$$\mathbf{P}[A] \;=\; \lim_{n\to\infty} \mathbf{P}[A_n^+] \;\leq\; \lim_{n\to\infty} \mathbf{P}[A_n] \;\leq\; \lim_{n\to\infty} \mathbf{P}[A_n^-] \;=\; \mathbf{P}[A].$$

But this means that $\mathbf{P}[A_n] \to \mathbf{P}[A]$. □

Finally we state the so-called *inclusion/exclusion formula*.

2.2.7 Theorem. *Let* $A_n \in \mathcal{F}$, $n \in \mathbb{N}$. *Then*

$$\mathbf{P}\left[\bigcup_{n=1}^{\infty} A_n\right] = \lim_{n\to\infty} \sum_{m=1}^{n} (-1)^{m-1} \sum_{k_1<\cdots<k_m=1}^{n} \mathbf{P}\left[\bigcap_{l=1}^{m} A_{k_l}\right].$$

Proof. Exercise 2.5. □

2.3 Extension Theorems

In modeling it is often so that the probabilities $\mathbf{P}[C]$, $C \in \mathcal{C}$, are given for a collection $\mathcal{C}$ that is not a σ-algebra. Then we are faced with two important questions

1. *uniqueness*, i.e. can we extend the probabilities $\mathbf{P}$ uniquely to the σ-algebra $\sigma(\mathcal{C})$,
2. *existence*, i.e., does there exist a probability measure $\mathbf{P}$ from the collection $\sigma(\mathcal{C})$, when the probabilities $\mathbf{P}[C]$, $C \in \mathcal{C}$, are fixed.

A typical situation is that the measurable space is $(\mathbb{R}, \mathcal{B})$ but we are only given the cumulative distribution function

$$F(x) := \mathbf{P}[(-\infty, x]],$$

i.e., we are only given the probabilities $\mathbf{P}[C]$ for C in the collection $\mathcal{C} = \{(-\infty, x]; x \in \mathbb{R}\}$.

The motto of this section is that σ-algebras are difficult, but algebras, π-classes and d-classes are easy. Indeed, in modeling the class $\mathcal{C}$ is often an algebra, a π-class, or a d-class.

2.3.1 Definition. A collection $\mathcal{A}$ of subsets of the sample set Ω is an *algebra*, if

(i) $\Omega \in \mathcal{A}$,
(ii) if $A \in \mathcal{A}$, then $A^c \in \mathcal{A}$,
(iii) if $A_1, A_2 \in \mathcal{A}$, then $A_1 \cap A_2 \in \mathcal{A}$.

2.3.2 Definition. A collection $\mathcal{I}$ of subsets of the sample space Ω is a *π-class*, if $I_1 \cap I_2 \in \mathcal{I}$, whenever $I_1, I_2 \in \mathcal{I}$.

2.3.3 Definition. A collection $\mathcal{D}$ of subsets of the sample set Ω is a *d-class* or *monotone class*, if

(i) $\Omega \in \mathcal{D}$,

(ii) If $A, B \in \mathcal{D}$ and $A \subset B$, then $B \setminus A \in \mathcal{D}$,
(iii) If $A_n \in \mathcal{D}$ for all $n \in \mathbb{N}$ and $A_n \uparrow A$, then $A \in \mathcal{D}$.

So, algebras are closed with respect to finitely many set operations, π-classes are closed under finitely many intersections and d-classes are closed under countable monotone set operations.

As with σ-algebras, also for algebras, π-classes and d-classes we can talk about generators

2.3.4 Lemma. *If $\mathcal{C}_j$, $j \in J$, are algebras (resp. π-classes, d-classes), then $\cap_{j\in J}\mathcal{C}_j$ is an algebra (resp. a π-class, a d-class).*

Proof. Exercise 2.6. □

2.3.5 Definition. The algebras, π-classes, and d-classes generated by a collection $\mathcal{C}$ are

$$\begin{aligned} a(\mathcal{C}) &:= \bigcap \{\mathcal{A}\,;\, \mathcal{A} \text{ is an algebra } \mathcal{C} \subset \mathcal{A}\}\,, \\ \pi(\mathcal{C}) &:= \bigcap \{\mathcal{I}\,;\, \mathcal{I} \text{ is a } \pi\text{-class and } \mathcal{C} \subset \mathcal{I}\}\,, \\ d(\mathcal{C}) &:= \bigcap \{\mathcal{D}\,;\, \mathcal{D} \text{ is a } d\text{-class and } \mathcal{C} \subset \mathcal{D}\}\,. \end{aligned}$$

2.3.6 Lemma. *A collection $\mathcal{F}$ is a σ-algebra if and only if its is both a π-class and a d-class.*

Proof. It is clear that σ-algebras are both π-classes and d-classes.

Let then $\mathcal{F}$ be both π-class and d-class. Let $A \in \mathcal{F}$. Then $A^c = \Omega \setminus A \in \mathcal{F}$, because $\mathcal{F}$ is a d-class. Assume furthermore that $B \in \mathcal{F}$. Then $A \cup B \in \mathcal{F}$ because of the De Morgan's law $A \cup B = (A^c \cap B^c)^c$. So, we have shown that $\mathcal{F}$ is an algebra. Let then $A_n \in \mathcal{F}$ for all $n \in \mathbb{N}$. Denote $B_n := \cup_{k=1}^n A_k$. Then $B_n \uparrow \cup_{n=1}^\infty A_n$. So, $\cup_{n=1}^\infty A_n \in \mathcal{F}$. By using the De Morgan's law $\cap_{n=1}^\infty A_n = (\cup_{n=1}^\infty A_n^c)^c$ we see that $\mathcal{F}$ is a σ-algebra. □

The following result is called *Dynkin's lemma* or the *monotone class theorem*.

2.3.7 Lemma. *If $\mathcal{I}$ is a π-class, then $d(\mathcal{I}) = \sigma(\mathcal{I})$.*

Proof. By Lemma 2.3.6 it is enough to show that $d(\mathcal{I})$ is a π-class.

Step 1: Denote

$$\mathcal{D}_1 := \{B \in d(\mathcal{I})\,;\, B \cap I \in d(\mathcal{I}) \text{ for all } I \in \mathcal{I}\}\,.$$

Obviously $\mathcal{I} \subset \mathcal{D}_1$. We show that $\mathcal{D}_1$ is a d-class. Clearly $\Omega \in \mathcal{D}_1$. Let then $A, B \in \mathcal{D}_1$, $A \subset B$ and $I \in \mathcal{I}$. Then

$$\begin{aligned}(B \setminus A) \cap I &= B \cap I \cap A^c \\ &= (B \cap I) \cap (A^c \cap I^c) \\ &= (B \cap I) \setminus (A \cap I).\end{aligned}$$

Because $(B \cap I) \in d(\mathcal{I})$ ja $(A \cap I) \in d(\mathcal{I})$, we have $(B \cap I) \setminus (A \cap I) \in d(\mathcal{I})$. Therefore $B \setminus A \in \mathcal{D}_1$. Finally, let $(A_n)_{n\in\mathbb{N}} \subset \mathcal{D}_1$ and $A_n \uparrow A$. Then for all $I \in \mathcal{I}$ we have $A_n \cap I \in d(\mathcal{I})$ and $A_n \cap I \uparrow A \cap I$. We see that $A \cap I \in d(\mathcal{I})$. Consequently, $A \in \mathcal{D}_1$. We have shown that $\mathcal{D}_1$ is a d-class. Now, because $\mathcal{I} \subset \mathcal{D}_1$, we have $d(\mathcal{I}) \subset \mathcal{D}_1$. Consequently, $d(\mathcal{I}) = \mathcal{D}_1$.

Step 2: Denote

$$\mathcal{D}_2 \;:=\; \{A \in d(\mathcal{I})\,;\, A \cap B \in d(\mathcal{I}) \text{ for all } B \in d(\mathcal{I})\}\,.$$

By Step 1, $\mathcal{D}_1 = d(\mathcal{I})$. Therefore, $\mathcal{I} \subset \mathcal{D}_2$. Now, as in Step 1, it is easy to see that $\mathcal{D}_2$ is a d-class. Moreover, $\mathcal{I} \subset \mathcal{D}_2 \subset \mathcal{D}_1 \subset d(\mathcal{I})$. Therefore $\mathcal{D}_2 = d(\mathcal{I})$. Finally, we note that the claim follows by noticing that $\mathcal{D}_2$ is clearly a π-class. □

The next result is the so-called *Dynkin's extension theorem*. It guarantees the uniqueness of the extension of the probability measure.

2.3.8 Theorem. *Let $\mathcal{I}$ be a π-class and let $\mathcal{F} = \sigma(\mathcal{I})$. Let $\mathbf{P}$ and $\mathbf{P}'$ be two probability measures on $(\Omega, \mathcal{F})$. If $\mathbf{P} = \mathbf{P}'$ on the π-class $\mathcal{I}$, then $\mathbf{P} = \mathbf{P}'$ on the σ-algebra $\mathcal{F}$.*

Proof. Denote

$$\mathcal{D} \;:=\; \{A \in \mathcal{F}\,;\, \mathbf{P}[A] = \mathbf{P}'[A]\}\,.$$

We show that $\mathcal{D}$ is a d-class. Obviously $\Omega \in \mathcal{D}$. Let then $A, B \in \mathcal{D}$ and $A \subset B$. Then

$$\mathbf{P}[B \setminus A] \;=\; \mathbf{P}[B] - \mathbf{P}[A] \;=\; \mathbf{P}'[B] - \mathbf{P}'[A] \;=\; \mathbf{P}'[B \setminus A].$$

Consequently, $B \setminus A \in \mathcal{D}$. Finally, let $(A_n)_{n\in\mathbb{N}} \subset \mathcal{D}$, $A_n \uparrow A$. By the monotone continuity of probability

$$\mathbf{P}[A] \;=\; \lim_{n\to\infty} \mathbf{P}[A_n] \;=\; \lim_{n\to\infty} \mathbf{P}'[A_n] \;=\; \mathbf{P}'[A].$$

Consequently, $A \in \mathcal{D}$.

Now $\mathcal{D}$ is a d-class and $\mathcal{I} \subset \mathcal{D}$ by assumption. Therefore, by Dynkin's lemma 2.3.7, $\mathcal{D} \supset d(\mathcal{I}) = \sigma(\mathcal{I}) = \mathcal{F}$. □

Dynkin's extension theorem 2.3.8 guaranteed only the uniqueness of the extension of the probability measure. The next *Carathéodory's extension theorem* guarantees also the existence.

2.3.9 Theorem. *Let $\mathcal{A}$ be an algebra and let $\mathbf{P}_0 : \mathcal{A} \to [0,1]$ be a σ-additive set function satisfying $\mathbf{P}_0[\Omega] = 1$. Then there exist a unique probability measure $\mathbf{P} : \sigma(\mathcal{A}) \to [0,1]$ such, that $\mathbf{P}[A] = \mathbf{P}_0[A]$, whenever $A \in \mathcal{A}$.*

Proof. This is a good topic for a master's thesis. □

2.3.10 Corollary. *For the measurable space $(\mathbb{R}, \mathcal{B})$ there exists a unique* Lebesgue's measure *ℓ, that is defined by setting $\ell[(a,b)] = |b-a|$ for all $a, b \in \mathbb{R}$.*

Proof. This is also a good topic for a master's thesis. □

Finally, we talk about the *completeness* of a probability space $(\Omega, \mathcal{F}, \mathbf{P})$. Let $\mathcal{N}$ be a *null set*:

> $N \in \mathcal{N}$ if and only there exists an $A \in \mathcal{F}$ such, that $N \subset A$ and $\mathbf{P}[A] = 0$.

It is philosophically satisfying to say that then $\mathbf{P}[N] = 0$ for all $N \in \mathcal{N}$. Mathematically this can be achieved by considering the completed σ-algebra $\mathcal{F}^*$, that consists of the sets A^*, for which there exists such $A^-, A^+ \in \mathcal{F}$, that $A^- \subset A^* \subset A^+$ and $\mathbf{P}[A^+ \setminus A^-] = 0$. Then $\mathcal{F}^*$ is a σ-algebra. In fact, it can be shown tha $\mathcal{F}^* = \sigma(\mathcal{F}, \mathcal{N})$. We then set $\mathbf{P}^*[A^*] = \mathbf{P}[A^+] = \mathbf{P}[A^-]$. The probability space $(\Omega, \mathcal{F}^*, \mathbf{P}^*)$ is the *completion* of the probability space $(\Omega, \mathcal{F}, \mathbf{P})$

2.4 Measurability and Nonmeasurability

This section is all about nonmeasurability. We show that in a relatively natural setting there are sets that cannot admit a probability. This justifies the use of σ-algebras $\mathcal{F}$ instead of the entire collection of all subsets $\mathrm{pow}(\Omega)$.

2.4.1 Example. Let $\Omega = [0,1]$ and let $\mathbf{P}$ be the uniform probability measure, or the Lebesgue's measure. Then

$$\mathbf{P}[(a,b)] \;=\; |b-a|,$$

for all $a, b \in \Omega$. We show that there exist a set $A \in \mathrm{pow}(\Omega)$ that cannot admit a probability.

Let $x, y \in [0,1]$. In what follows $x+y$ is the sum $x+y$, if $x+y \leq 1$ and the sum $x+y-1$ otherwise.

Let $B \subset \Omega$ and $\omega \in \Omega$. Denote

$$B + \omega \;:=\; \{\omega' \in \Omega\,;\, \omega' = b + \omega, b \in B\}\,.$$

We note that $\mathbf{P}[B] = \mathbf{P}[B+\omega]$ for all $\omega \in \Omega$, since $\mathbf{P}$ is translation invariant. Moreover, B admits a probability if and only if $B+\omega$ admits a probability for all $\omega \in \Omega$.

We then define an equivalence relation $\sim$ on Ω by setting $\omega \sim \omega'$ if $\omega - \omega' \in \mathbb{Q}$. Now, $\sim$ splits the sample space Ω into equivalence classes. Let then A se a set, where we have picked one element form each of these equivalence classes. This set exists by the axiom of choice. We show that this A does not admit a probability.

Let $q_1, q_2, \ldots$ be the rational numbers in the set Ω. Consider the sets $A_n := A + q_n$. These sets form a partition of the sample set Ω (Exercise 2.7).

Now, if A admits a probability, then each of the sets A_n, $n \in \mathbb{N}$, also admit a probability. Since the sets A_n, $n \in \mathbb{N}$, form a partition of Ω we have

$$1 \;=\; \mathbf{P}[\Omega] \;=\; \mathbf{P}\left[\bigcup_{n=1}^{\infty} A_n\right] \;=\; \sum_{n=1}^{\infty} \mathbf{P}[A_n]. \tag{2.4.2}$$

But the probability $\mathbf{P}$ is translation invariant. Therefore, $\mathbf{P}[A_n]$ is a constant. Consequently, the equation (2.4.2) is impossible. We have reached a contradiction. The assumption we have to abandon is that A admits a probability.

Exercises

2.1. Let $\Omega = \{1, 2, \ldots, n\}$. How many different σ-algebras the set Ω has?

2.2. Let $\mathcal{F}$ a σ-algebra with infinite number of elements. Show that $\mathcal{F}$ cannot be countable.

2.3. Prove the parts (i)–(iv) and (vi)–(x) of Lemma 2.1.7.

2.4. Let $A_n \in \mathcal{F}$, $n \in \mathbb{N}$. Show that

$$\mathbf{P}\left[\bigcup_{n=1}^{\infty} A_n\right] \;\leq\; \sum_{n=1}^{\infty} \mathbf{P}[A_n].$$

2.5. Prove Theorem 2.2.7.

2.6. Prove Lemma 2.3.4.

2.7. Show that in Example 2.4.1 the sets A_n, $n \in \mathbb{N}$, form a partition of the sample space Ω. In other words, show that A_n's are disjoint and that $\cup_{n=1}^{\infty} A_n = \Omega$.

Chapter 3

Random Variables

The idea of a random variable, or rather a random function, is the following: The probability space $(\Omega, \mathcal{F}, \mathbf{P})$ is an abstraction that does not need to relate in any way to our observable reality. The real observations come from random experiments, where we observe a random variable $X\colon \Omega \to \mathbb{R}$. We do not observe $\mathbf{P}$, but the distribution probability measure of X: $\mathbf{P}_X[\,\cdot\,] := \mathbf{P}[X \in \cdot\,] := \mathbf{P} \circ X^{-1}[\,\cdot\,]$. In general, the random function X can be a measurable function from the measurable space $(\Omega, \mathcal{F})$ to any measurable space $(M, \mathcal{M})$. Measurability means that $\{X \in M\} \in \mathcal{F}$ for all $M \in \mathcal{M}$. In this course we only consider real-valued (and complex-valued) random variables, i.e., for us $(\mathbb{M}, \mathcal{M}) = (\mathbb{R}, \mathcal{B})$ for real-valued random variables, and a complex-valued random variable is $Z = X + \mathrm{i}Y$, where X and Y are real-valued random variables.

3.1 Definition as Measurable Function

3.1.1 Definition. A function $X : \Omega \to \mathbb{R}$ is a *random variable* or a *measurable function*, if $\{X \in B\} \in \mathcal{F}$ for all $B \in \mathcal{B}$.

Whether a function $X\colon \Omega \to \mathbb{R}$, or rather $X\colon (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B})$ is a random variable depends of how large the σ-algebra $\mathcal{F}$ is. The next example discusses this point.

3.1.2 Example. We consider functions $X : \Omega \to \mathbb{R}$.

(i) If $\mathcal{F} = \mathrm{pow}\,(\Omega)$, then all functions are random variables.
(ii) If $\mathcal{F} = \{\Omega, \varnothing\}$, then only constant functions are random variables.
(iii) Let $\Omega = \{1, 2, 3, 4, 5, 6\}$ and X be the result of a dice rolling. If $\mathcal{F} = \mathrm{pow}\,(\Omega)$, then X is a random variable from the measurable space $(\Omega, \mathcal{F})$. If $\mathcal{G} = \{\Omega, \{1, 3, 5\}, \{2, 4, 6\}, \varnothing\}$, then X is not a random

variable from the measurable space $(\Omega, \mathcal{G})$. However, the function $Y := \mathbf{1}_{\{\text{the dice roll is even}\}}$ is a random variable from the measurable space $(\Omega, \mathcal{G})$.

There is a relatively simple characterization of a random variable in terms of level sets:

3.1.3 Theorem. *A function $X\colon \Omega \to \mathbb{R}$ is a random variable if and only if $\{X \le x\} \in \mathcal{F}$ for all $x \in \mathbb{R}$.*

Theorem 3.1.3 follows from the following more general theorem.

3.1.4 Theorem. *Let $\mathcal{B}_0$ be a class of subsets of $\mathbb{R}$ such that $\sigma(\mathcal{B}_0) = \mathcal{B}$. The function $X\colon \Omega \to \mathbb{R}$ is a random variable if and only if $\{X \in B_0\} \in \mathcal{F}$ for all $B_0 \in \mathcal{B}_0$.*

Proof. The necessity is clear, and we show the sufficiency. Suppose that $\{X \in B_0\} \in \mathcal{F}$ for all $B_0 \in \mathcal{B}_0$. We need to show that $\{X \in B\} \in \mathcal{F}$ for all $B \in \mathcal{B}$. First note that the pre-images of all countable set operations commute. Define $\mathcal{G} = \{B \in \mathcal{B}; \{X \in B\} \in \mathcal{F}\}$. Then $\mathcal{B}_0 \subset \mathcal{G}$, and since the pre-images commute, we have that $\mathcal{G}$ is also a σ-algebra. Therefore $\mathcal{G} \supset \sigma(\mathcal{B}_0)$, and also $\mathcal{B} \supset \mathcal{G}$. Therefore $\mathcal{B} = \mathcal{G}$, and thus $\{X \in B\} \in \mathcal{F}$ for all $B \in \mathcal{B}$. □

3.2 Limits and Transformations

3.2.1 Theorem. *Let X_n, $n \in \mathbb{N}$, be random variables. Then $\sup_{n\in\mathbb{N}} X_n$, $\inf_{n\in\mathbb{N}} X_n$, $\liminf_{n\to\infty} X_n$, and $\limsup_{n\to\infty} X_n$ are also random variables.*

Proof. Since $\{\sup_{n\in\mathbb{N}} X_n \le a\} = \cap_{n=1}^{\infty}\{X_n \le a\}$ (Exercise 3.1) and $\{X_n \le a\} \in \mathcal{F}$ for all $a \in \mathbb{R}$, we see that $\sup_{n\in\mathbb{N}} X_n$ a random variable. In the same way we see that also $\inf_{n\in\mathbb{N}} X_n$ is a random variable. But $\limsup_{n\to\infty} X_n = \inf_{n\ge1}\sup_{k\ge n} X_k$. Therefore $\limsup_{n\to\infty} X_n$ is a random variable. In the same way we see that also $\liminf_{n\to\infty} X_n$ is a random variable. □

3.2.2 Corollary. *Let X_n, $n \in \mathbb{N}$, be random variables. If $X_n \to X$, then X is also a random variable.*

3.2.3 Definition. A function $f\colon \mathbb{R} \to \mathbb{R}$ is *Borel measurable*, if $\{f \in B\} \in \mathcal{B}$ for all $B \in \mathcal{B}$, i.e., the pre-image of a Borel's set is a Borel's set.

Recall from topology that a function $f\colon \mathbb{R} \to \mathbb{R}$ is continuous if the pre-image $\{f \in U\} \subset \mathbb{R}$ is an open set for every open set $U \subset \mathbb{R}$.

3.2.4 Lemma. *Let $f\colon \mathbb{R} \to \mathbb{R}$ be continuous. Then f is Borel measurable.*

Proof. We only need to show that $\{f < x\} \in \mathcal{B}$ for all $x \in \mathbb{R}$. But the function f is continuous. Hence the pre-image $\{f < x\}$ of the open set $(-\infty, x)$ is open. But $\mathcal{B}$ contains the open sets. Therefore $\{f < x\} \in \mathcal{B}$. □

In fact, it is pretty difficult to construct a function $f\colon \mathbb{R} \to \mathbb{R}$ that is not Borel measurable. We do not give examples of non-Borel measurable functions.

3.2.5 Theorem. *Let X be a random variable and let $f\colon \mathbb{R} \to \mathbb{R}$ be Borel measurable. Then $f(X)$ is a random variable.*

Proof. Let $B \in \mathcal{B}$. Then $\{f(X) \in B\} = \{X \in f^{-1}B\}$. Since f is Borel measurable, we have that $B_f := f^{-1}B = \{f \in B\} \in \mathcal{B}$ and since X is a random variable, we have that $\{f(X) \in B\} = \{X \in B_f\} \in \mathcal{F}$. □

3.2.6 Theorem. *Let X and Y be random variables. Then also $X+Y$, XY, $\max(X,Y)$, and $\min(X,Y)$ are random variables. Also, if $Y \neq 0$, then X/Y is also a random variable.*

Proof. Let us first consider $X+Y$. We show that $\{X+Y > z\} \in \mathcal{F}$ for all $z \in \mathbb{R}$ an then the result follows from Theorem 3.1.4. Now,

$$\{X+Y > z\} \;=\; \bigcup_{q\in\mathbb{Q}} \Big(\{X > q\} \cap \{Y > z-q\}\Big),$$

which is a countable union of intersection sets in $\mathcal{F}$. Therefore we conclude that $\{X+Y > z\} \in \mathcal{F}$.

Let us then consider XY. We use a trick here. We first note that

$$XY = \frac{1}{2}\left[(X+Y)^2 - X^2 - Y^2\right].$$

Since the sums of random variables are random variables, and constant times random variables are clearly random variables, it is enough to show that X^2 is a random variable if X is a random variable. But this is obvious, since

$$\left\{X^2 \leq x\right\} \;=\; \left\{X \leq \sqrt{x}\right\} \cap \left\{X \geq -\sqrt{x}\right\}.$$

So see that $\max(X,Y)$ is a random variable it is enough to note that

$$\{\max(X,Y) \geq z\} \;=\; \{X \geq z\} \cup \{Y \geq z\}.$$

The fact that $\min(X, Y)$ is a random variable follows from the fact that

$$\{\min(X, Y) \leq z\} \;=\; \{X \leq z\} \cap \{Y \leq z\}\,.$$

Let us then consider X/Y with $Y \neq 0$. Now it is enough to show that $1/Y$ is a random variable. But this follows from the formulas

$$\left\{\frac{1}{Y} \leq y\right\} \;=\; \begin{cases} \left\{\frac{1}{y} \leq Y < 0\right\}, & \text{if } y < 0, \\ \{Y < 0\}\,, & \text{if } y = 0, \\ \{Y < 0\} \cap \left\{Y \geq \frac{1}{y}\right\}, & \text{if } y > 0. \end{cases}$$

□

3.3 Distributions

3.3.1 Definition. The *distribution* $\mathbf{P}_X$ of a random variable X is the probability measure it induces on $(\mathbb{R}, \mathcal{B})$ given by

$$\mathbf{P}_X[B] \;:=\; \mathbf{P}[X \in B], \qquad B \in \mathcal{B}.$$

3.3.2 Lemma. *Let X be a random variable. Then $\mathbf{P}_X$ is a probability measure on the measurable space $(\mathbb{R}, \mathcal{B})$.*

Proof. Exercise 3.5. □

3.3.3 Example. Let $B \in \mathcal{B}$.

(i) If X is a deterministic "random" variable, i.e., $\mathbf{P}[X = x_0] = 1$, then its distribution is *Dirac's point mass* at point x_0:

$$\mathbf{P}_X[B] \;=\; \delta_{x_0}[B] \;:=\; \begin{cases} 1, & \text{if } x_0 \in B, \\ 0, & \text{otherwise.} \end{cases}$$

(ii) Let X be the result od N-sided dice roll. Then its distribution is a sum of Dirac's point masses:

$$\mathbf{P}_X[B] \;=\; \frac{1}{N} \sum_{k=1}^{N} \delta_k[B].$$

In other words, X is uniformly distributed in the set $\{1, \ldots, N\}$.

(iii) A random variable X is *discrete*, if its range is discrete. We can represent its distribution by using Dirac's point massses. Indeed, the the range of X be $\{a_1, a_2, \ldots\}$ and denote $p_k := \mathbf{P}[X = a_k]$. Then

$$\mathbf{P}_X[B] = \sum_{k=1}^{\infty} p_k \delta_{a_k}[B]. \tag{3.3.4}$$

In fact, X is discrete if and only if its distribution admits the representation (3.3.4).

(iv) A random variable X is *continuous*, if its distribution admits the representation

$$\mathbf{P}_X[B] = \int_B f_X(x)\,\mathrm{d}x$$

for all $B \in \mathcal{B}$.The function $f\colon \mathbb{R} \to \mathbb{R}_+$ is the *density function* of the random variable X.

3.3.5 Theorem. *A function $F : \mathbb{R} \to \mathbb{R}$ is the* cumulative distribution function *of a random variable X, i.e., $F(x) = \mathbf{P}[X \le x]$, if and only if*

(i) *F is non-decreasing: $F(x) \le F(y)$ for all $x \le y$,*
(ii) *F is right-continuous: $F(x+) := \lim_{y \to x+} F(y) = F(x)$,*
(iii) *$\lim_{x\to\infty} F(x) = 1$ and $\lim_{x\to-\infty} F(x) = 0$.*

Proof. Exercise 3.6 □

Exercises

3.1. Let $a \in \mathbb{R}$ and let X_n, $n \in \mathbb{N}$, be random variables. Show that $\{\sup_{n\in\mathbb{N}} X_n \le a\} = \cap_{n=1}^{\infty}\{X_n \le a\}$.

3.2. Let $a \in \mathbb{R}$ and let X_n, $n \in \mathbb{N}$, be random variables. Which of the relations $=$, $\subset$, $\supset$ hold between the following pairs of events:

(a) $\{\inf_{n\in\mathbb{N}} X_n \le a\}$ and $\cup_{n=1}^{\infty}\{X_n \le a\}$,
(b) $\{\limsup_{n\to\infty} X_n \le a\}$ and $\{\liminf_{n\to\infty} X_n \le a\}$?

3.3. Let X be a random variable and $A \in \mathcal{F}$. Show that $X\mathbf{1}_A$ is a random variable.

3.4. Let X be a random variable and $A \notin \mathcal{F}$. Is $X\mathbf{1}_A$ a random variable?

3.5. Prove Lemma 3.3.2.

3.6. Prove Theorem 3.3.5.

Chapter 4

Expectation

The concept of the expectation of a random variable can be derived heuristically in many ways.

A physicist thinks about the distribution of a random variable and the center of its probability mass. In other words, the expectation is the point where the probability mass on the real line is in equilibrium.

For a gambler, the expectation is the amount one should be willing to pay in taking a part of the gamble given by the random variable.

For a measure theorist, or a probabilist, the expectation is the measure integral of the random variable with respect to the probability measure.

In these notes we take the probabilist's point of view.

4.1 Definition as Measure Integral

We start by considering simple random variables and extend the results to general random variables by using limit arguments.

4.1.1 Definition. A random variable X is *simple*, if

$$X = \sum_{k=1}^{n} a_k \mathbf{1}_{A_k}, \tag{4.1.2}$$

for some $a_k \in \mathbb{R}$ and $A_k \in \mathcal{F}$. If $A_k = \{X = a_k\}$, the representation (4.1.2) is *minimal*. We denote by $\mathcal{E}$ the class on simple random variables.

In the minimal representation the sets $\{X = a_k\}$, $k = 1, \ldots, n$, form a partition of the sample space Ω: $\{X = a_k\}$'s are disjoint and $\cup_{k=1}^{n}\{X = a_k\} = \Omega$.

4.1.3 Example. Let X be the number of heads is a series of n coin tosses. Then X is a simple random variable and we can represent

$$X \;=\; \sum_{k=1}^{n} \mathbf{1}_{A_k},$$

where A_k is the event "k^{th} toss is head". This representation is not minimal. The minimal representation is

$$X \;=\; \sum_{k=0}^{n} k\mathbf{1}_{\{X=k\}}.$$

4.1.4 Lemma. *$\mathcal{E}$ is a* vector algebra*: if $X, Y \in \mathcal{E}$ and $a, b \in \mathbb{R}$, the $aX+bY \in \mathcal{E}$ and $XY \in \mathcal{E}$.*

Proof. The claim follows by noticing that $X \in \mathcal{E}$ if and only if the *range* $X\Omega = \{X(\omega); \omega \in \Omega\}$ is finite. □

4.1.5 Definition. Olkoon $X \in \mathcal{E}$ with representation (4.1.2). Then its *expectation* is

$$\mathbf{E}[X] \;:=\; \sum_{k=1}^{n} a_k \mathbf{P}[A_k].$$

It should be noted that the expectation is well-defined in the sense that it does not depend on the particular representation (4.1.2). This will follow from the fact that $\mathbf{E}$ is a positive linear operator, which we shall see in Lemma 4.1.7. In particular, we note that

$$\mathbf{E}[\mathbf{1}_A] \;=\; \mathbf{P}[A].$$

Also, we note that from the minimal representation it follows that

$$\mathbf{E}[X] \;=\; \sum_{k=1}^{n} a_k \mathbf{P}[X = a_k].$$

4.1.6 Example. (i) Let X be uniformly distributed on the set $\{1, \ldots, N\}$, i.e. $\mathbf{P}[X = k] = \frac{1}{N}$, when $k \in \{1, \ldots, N\}$, and 0 otherwise. Then

$$\mathbf{E}[X] \;=\; \sum_{k=1}^{N} k\mathbf{P}[X = k] \;=\; \frac{1}{N}\sum_{k=1}^{N} k \;=\; \frac{N+1}{2}.$$

(ii) Let X be binomially distributed with parameters n ja p, i.e.

$$\mathbf{P}[X = k] \;=\; \binom{n}{k} p^k q^{n-k}$$

($q := 1 - p$), when $k \in \{0, \ldots, n\}$ and 0 otherwise. Then

$$\begin{aligned}
\mathbf{E}[X] &= \sum_{k=1}^{n} k \binom{n}{k} p^k q^{n-k} \\
&= \sum_{k=1}^{n} k \frac{n!}{k!(n-k)!} p^k q^{n-k} \\
&= np \sum_{k=1}^{n} \frac{(n-1)!}{(k-1)!((n-1)-(k-1))!} p^{k-1} q^{(n-1)-(k-1)} \\
&= np \sum_{k'=1}^{n'} \binom{n'}{k'} p^{k'} q^{n'-k'} \\
&= np
\end{aligned}$$

by the Newton's binomial formula. In the above we used the change-of-variables $n' := n - 1$ ja $k' := k - 1$.

4.1.7 Lemma. $\mathbf{E} : \mathcal{E} \to \mathbb{R}$ *is a positive linear operator:*

(i) *if* $X, Y \in \mathcal{E}$ *and* $a, b \in \mathbb{R}$, *then* $\mathbf{E}[aX + bY] = a\mathbf{E}[X] + b\mathbf{E}[Y]$,
(ii) *if* $X \in \mathcal{E}$ *and* $X \geq 0$, *then* $\mathbf{E}[X] \geq 0$.

Proof. (i) Let

$$X \;=\; \sum_{k=1}^{n} a_k \mathbf{1}_{A_k} \quad \text{and} \quad Y \;=\; \sum_{l=1}^{m} b_l \mathbf{1}_{B_l},$$

where the A_k's and the B_l's from partitions of the sample space Ω (consider th eminimal representations, for example). Then

$$aX + bY \;=\; \sum_{k=1}^{n} \sum_{l=1}^{m} (aa_k + bb_l) \mathbf{1}_{A_k \cap B_l}$$

and

$$\begin{aligned}
\mathbf{E}[aX+bY] &= \sum_{k=1}^{n}\sum_{l=1}^{m}(aa_k+bb_l)\mathbf{P}[A_k\cap B_l] \\
&= \sum_{k=1}^{n}\sum_{l=1}^{m}aa_k\mathbf{P}[A_k\cap B_l]+\sum_{k=1}^{n}\sum_{l=1}^{m}bb_l\mathbf{P}[A_k\cap B_l] \\
&= a\sum_{k=1}^{n}a_k\sum_{l=1}^{m}\mathbf{P}[A_k\cap B_l]+b\sum_{l=1}^{m}b_l\sum_{k=1}^{n}\mathbf{P}[A_k\cap B_l] \\
&= a\sum_{k=1}^{n}a_k\mathbf{P}[A_k]+b\sum_{l=1}^{m}b_l\mathbf{P}[B_l] \\
&= a\mathbf{E}[X]+b\mathbf{E}[Y].
\end{aligned}$$

(ii) This claim is trivial. □

4.1.8 Corollary. *Let $X, Y \in \mathcal{E}$. If $X \le Y$, then $\mathbf{E}[X] \le \mathbf{E}[Y]$.*

We denote $X^+ := \max(X, 0)$ and $X^- := -\min(X, 0)$. Then X^+ and X^- are positive. Moreover, $X = X^+ - X^-$ and $|X| = X^+ + X^-$.

To define the expectation $\mathbf{E}[X]$ of a genaral random variable, it is tempting to define $\mathbf{E}[X] = \lim_{n\to\infty}\mathbf{E}[X_n]$, where $(X_n)_{n\in\mathbb{N}} \subset \mathcal{E}$ and $X_n \to X$. But this requires that $\lim_{n\to\infty}\mathbf{E}[X_n] = \mathbf{E}[\lim_{n\to\infty}X_n]$, which is unfortunately not true in general. Therefore, we have the following definition that does not use limits.

4.1.9 Definition. The *expectation* of a positive random variable X is

$$\mathbf{E}[X] := \sup\{\mathbf{E}[Y]\,;\, Y\in\mathcal{E}, Y\le X\}$$

if this supremum is finite. If $\mathbf{E}[|X|] < \infty$, we say that X is *integrable* and define

$$\mathbf{E}[X] := \mathbf{E}[X^+] - \mathbf{E}[X^-].$$

The class of integrable random variables is $L^1 = L^1(\Omega, \mathcal{F}, \mathbf{P})$.

For the expectation $\mathbf{E}[X]$ we also use the notations

$$\int_\Omega X(\omega)\,\mathbf{P}[\mathrm{d}\omega], \quad \int_\Omega X(\omega)\,\mathrm{d}\mathbf{P}[\omega], \quad \int_\Omega X\mathrm{d}\mathbf{P},$$

$$\int_{-\infty}^{\infty} x\mathbf{P}_X[\mathrm{d}x] \quad \text{and} \quad \int_{-\infty}^{\infty} x\,\mathrm{d}F_X(x).$$

The last two of these notations emphasize the fact that th expectation depend of the random variable only via its distribution: if $\mathbf{P}_X = \mathbf{P}_Y$, then $\mathbf{E}[X] = \mathbf{E}[Y]$.

4.1.10 Lemma. *L^1 is a vector space: if $X, Y \in L^1$ and $a, b \in \mathbb{R}$, then $aX + bY \in L^1$.*

Proof. Exercise 4.2. □

4.1.11 Theorem. *Let X be a positive random variable.*

(i) *There exists a sequence $(X_n)_{n\in\mathbb{N}} \subset \mathcal{E}$ such that $X_n \uparrow X$.*
(ii) *Let $(X_n)_{n\in\mathbb{N}} \subset \mathcal{E}$ be such that $X_n \uparrow X$. Then $\mathbf{E}[X_n] \uparrow \mathbf{E}[X]$, and $\mathbf{E}[X] < \infty$ if and only if $X \in L^1$.*

Proof. (i) We can set

$$X_n \;:=\; \sum_{k=1}^{n2^n} (k-1)2^{-n} \mathbf{1}_{\{(k-1)2^{-n} \le X < k2^{-n}\}} + n\mathbf{1}_{\{X \ge n\}}.$$

(ii) Since $\mathbf{E}[X] = \sup\{\mathbf{E}[Y]; Y \in \mathcal{E}, Y \le X\}$, we have $\mathbf{E}[X_n] \le \mathbf{E}[X]$ for all $n \in \mathbb{N}$. Consequently, $\lim_{n\to\infty} \mathbf{E}[X_n] \le \mathbf{E}[X]$. Therefore, it is enough to show that $\mathbf{E}[X] \le \lim_{n\to\infty} \mathbf{E}[X_n]$.

We show that if $Y \in \mathcal{E}$ and $Y \le X$, then $\mathbf{E}[Y] \le \lim_{n\to\infty} \mathbf{E}[X_n]$. The inequality $\mathbf{E}[X] \le \lim_{n\to\infty} \mathbf{E}[X_n]$ follows from this, because

$$\begin{aligned} \mathbf{E}[X] &= \sup\{\mathbf{E}[Y]\,;\, Y \in \mathcal{E}, Y \le X\} \\ &\le \sup\Big\{ \lim_{n\to\infty} \mathbf{E}[X_n]\,;\, Y \in \mathcal{E}, Y \le X \Big\} \\ &= \lim_{n\to\infty} \mathbf{E}[X_n]. \end{aligned}$$

Let then

$$Y \;=\; \sum_{k=1}^{m} a_k \mathbf{1}_{A_k},$$

where $A_k = \{Y = a_k\}$. Let $\varepsilon \in (0,1)$ be arbitrary. Denote

$$Y_{n,\varepsilon} \;:=\; (1-\varepsilon) Y \mathbf{1}_{\left\{(1-\varepsilon)Y \le X_n\right\}}.$$

Let

$$A_{n,k} \;:=\; A_k \cap \{(1-\varepsilon)Y \le X_n\}.$$

Since $Y_{n,\varepsilon} \le X_n$, we have that $\mathbf{E}[Y_{n,\varepsilon}] \le \mathbf{E}[X_n]$. Morevover,

$$\mathbf{E}[Y_{n,\varepsilon}] \;=\; (1-\varepsilon) \sum_{k=1}^{m} a_k \mathbf{P}[A_{n,k}] \tag{4.1.12}$$

Now $A_{n,k} \uparrow A_k$ as $n \to \infty$ because $X_n \uparrow X$ and $Y \leq X$. By taking the limits in (4.1.12) and by using the continuity of the probability we obtain that

$$\begin{aligned} \lim_{n\to\infty} \mathbf{E}[Y_{n,\varepsilon}] &= (1-\varepsilon) \sum_{k=1}^{m} a_k \lim_{n\to\infty} \mathbf{P}[A_{n,k}] \\ &= (1-\varepsilon) \sum_{k=1}^{m} a_k \mathbf{P}[A_k] \\ &= (1-\varepsilon)\mathbf{E}[Y]. \end{aligned}$$

Since $\mathbf{E}[Y_{n,\varepsilon}] \leq \mathbf{E}[X_n]$, we see that $\lim_{n\to\infty} \mathbf{E}[Y_{n,\varepsilon}] \leq \lim_{n\to\infty} \mathbf{E}[X_n]$. Therefore $(1-\varepsilon)\mathbf{E}[Y] \leq \lim_{n\to\infty} \mathbf{E}[X_n]$. The claim $\mathbf{E}[Y] \leq \lim_{n\to\infty} \mathbf{E}[X_n]$ follows now from the fact that $\varepsilon \in (0,1)$ was arbitrary.

Finally, we note that the proof shows that $\mathbf{E}[X] < \infty$ if and only if $\sup_{n\in\mathbb{N}} \mathbf{E}[X_n] < \infty$. □

4.1.13 Corollary. *Let $X \in L^1$. Then there exists a sequence $(X_n)_{n\in\mathbb{N}} \subset \mathcal{E}$ such that $X_n \to X$ and $\mathbf{E}[X_n] \to \mathbf{E}[X]$.*

Proof. This claim follows from Theorem 4.1.11 applied to the positive and negative parts, X^+ and X^-, of $X = X^+ - X^-$. □

By using Theorem 4.1.11 we obtain the familiar definitions of expectations for discrete and continuous random variables.

4.1.14 Corollary. (i) *Let X be a discrete random variable with range $\{a_1, a_2, \ldots\}$. Then $X \in L^1$ if and only if*

$$\sum_{k=1}^{\infty} |a_k| \mathbf{P}[X = a_k] < \infty$$

and in this case

$$\mathbf{E}[X] = \sum_{k=1}^{\infty} a_k \mathbf{P}[X = a_k].$$

(ii) *Let X be a continuous random variable with density function f_X. Then $X \in L^1$ if and only if*

$$\int_{-\infty}^{\infty} |x| f_X(x)\,\mathrm{d}x < \infty,$$

and in this case

$$\mathbf{E}[X] = \int_{-\infty}^{\infty} x f_X(x)\,\mathrm{d}x.$$

Proof. (i) The claim follows by splitting $X = X^+ - X^-$ and by approximating the positive and negative parts with the simple random variables

$$\begin{aligned} X_n^+ &:= X^+ \mathbf{1}_{\{X^+ \in \{a_1, a_2, \ldots, a_n\}\}}, \\ X_n^- &:= X^- \mathbf{1}_{\{X^- \in \{a_1, a_2, \ldots, a_n\}\}}. \end{aligned}$$

(ii) The claim follows by splitting $X = X^+ - X^-$ and by approximating the positive and negative parts with the simple random variables

$$\begin{aligned} X_n^+ &:= \sum_{k=1}^{n2^n} (k-1)2^{-n} \mathbf{1}_{\{(k-1)2^{-n} \le X^+ < k2^{-n}\}} \\ X_n^- &:= \sum_{k=1}^{n2^n} (k-1)2^{-n} \mathbf{1}_{\{(k-1)2^{-n} \le X^- < k2^{-n}\}} \end{aligned}$$

and by using the definition of th Riemann-integral. □

4.1.15 Theorem. $\mathbf{E} : L^1 \to \mathbb{R}$ *is a positive linear operator:*

(i) *if* $X, Y \in L^1$ *and* $a, b \in \mathbb{R}$, *then* $\mathbf{E}[aX + bY] = a\mathbf{E}[X] + b\mathbf{E}[Y]$,
(ii) *if* $X \in L^1$ *and* $X \ge 0$, *then* $\mathbf{E}[X] \ge 0$.

Proof. (ii) Let $X \in L^1$ be positive and $Y \in \mathcal{E}$ such that $Y \le X$. Then $Y\mathbf{1}_{\{Y \ge 0\}} \in \mathcal{E}$ and $Y\mathbf{1}_{\{Y \ge 0\}} \le X$. Thefore

$$\begin{aligned} \mathbf{E}[X] &= \sup\{\mathbf{E}[Y]\,;\, Y \in \mathcal{E}, Y \le X\} \\ &\ge \sup\left\{\mathbf{E}\left[Y\mathbf{1}_{\{Y \ge 0\}}\right]\,;\, Y \in \mathcal{E}, Y \le X\right\}. \end{aligned}$$

But $\mathbf{E}[Y\mathbf{1}_{\{Y \ge 0\}}] \ge 0$ by Lemma 4.1.7. Consequently, $\mathbf{E}[X] \ge 0$.

(i) We begin with positive $X \in L^1$ and $a \ge 0$. Let $(X_n)_{n \in \mathbb{N}} \subset \mathcal{E}$ be such that $X_n \uparrow X$. Then $(aX_n)_{n \in \mathbb{N}} \subset \mathcal{E}$ and $aX_n \uparrow aX$. Therefore, by Lemma 4.1.7 and by Theorem 4.1.11,

$$\mathbf{E}[aX] = \lim_{n \to \infty} \mathbf{E}[aX_n] = a \lim_{n \to \infty} \mathbf{E}[X_n] = a\mathbf{E}[X].$$

Let then $X \in L^1$ and $a \ge 0$. Now $(aX)^\pm = aX^\pm$, and by the previous derivation

$$\begin{aligned} \mathbf{E}[aX] &= \mathbf{E}[(aX)^+] - \mathbf{E}[(aX)^-] \\ &= a\mathbf{E}[X^+] - a\mathbf{E}[X^-] \\ &= a\left(\mathbf{E}[X^+] - \mathbf{E}[X^-]\right) \\ &= a\mathbf{E}[X]. \end{aligned}$$

If $a < 0$, the claim $\mathbf{E}[aX] = a\mathbf{E}[X]$ follows as above, but now $(aX)^{\pm} = -aX^{\mp}$.

We have shown that $\mathbf{E}[aX] = a\mathbf{E}[X]$. Next we show that $\mathbf{E}[X+Y] = \mathbf{E}[X] + \mathbf{E}[Y]$.

Let first $X, Y \in L^1$ bbe positive and $(X_n)_{n\in\mathbb{N}}, (Y_n)_{n\in\mathbb{N}} \subset \mathcal{E}$ such that $X_n \uparrow X$ and $Y_n \uparrow Y$. Then $(X_n + Y_n)_{n\in\mathbb{N}} \subset \mathcal{E}$ and $X_n + Y_n \uparrow X + Y$. Since the expectation is linear on $\mathcal{E}$ (Lemma 4.1.7), we have by Theorem 4.1.11 that

$$\begin{aligned}\mathbf{E}[X+Y] &= \lim_{n\to\infty} \mathbf{E}[X_n + Y_n] \\ &= \lim_{n\to\infty} (\mathbf{E}[X_n] + \mathbf{E}[Y_n]) \\ &= \lim_{n\to\infty} \mathbf{E}[X_n] + \lim_{n\to\infty} \mathbf{E}[Y_n] \\ &= \mathbf{E}[X] + \mathbf{E}[Y].\end{aligned}$$

Finally, let $X, Y \in L^1$. By definition

$$\mathbf{E}[X+Y] = \mathbf{E}[(X+Y)^+] - \mathbf{E}[(X+Y)^-].$$

Then, by definition and the derivations above,

$$\begin{aligned}\mathbf{E}[X] + \mathbf{E}[Y] &= \mathbf{E}[X^+] - \mathbf{E}[X^-] + \mathbf{E}[Y^+] - \mathbf{E}[Y^-] \\ &= \mathbf{E}[X^+ + Y^+] - \mathbf{E}[X^- + Y^-].\end{aligned}$$

We see that $\mathbf{E}[X+Y] = \mathbf{E}[X] + \mathbf{E}[Y]$ if and only if

$$\mathbf{E}[(X+Y)^+] - \mathbf{E}[(X+Y)^-] = \mathbf{E}[X^+ + Y^+] - \mathbf{E}[X^- + Y^-].$$

By reorganizing the terms we obtain

$$\mathbf{E}[(X+Y)^+ + X^- + Y^-] = \mathbf{E}[X^+ + Y^+ + (X+Y)^-].$$

But this follows from the fact that

$$(X+Y)^+ + X^- + Y^- = X^+ + Y^+ + (X+Y)^-,$$

which can be seen by a simple book-keeping argument. □

4.1.16 Corollary. *Let $X, Y \in L^1$ and $X \le Y$. Then*

(i) $\mathbf{E}[X] \le \mathbf{E}[Y]$,
(ii) $|\mathbf{E}[X]| \le \mathbf{E}[|X|]$.

4.2 Limits and Expectations

The next combined theorem tells us when we can change the order of limits and expectations. Its parts (i) and (ii) are called the *monotone convergence theorems*, parts (iii) and (iv) are the *Fatou's lemma*, ja part (v) is called the *Lebesgue's dominated convergence theorem.*

4.2.1 Theorem. *Let $X, Y \in L^1$ and let $(X_n)_{n\in\mathbb{N}}$ be a sequence of random variables.*

(i) *If $X_n \uparrow X$ and $\mathbf{E}[X_1] > -\infty$, then $\mathbf{E}[X_n] \uparrow \mathbf{E}[X]$.*
(ii) *If $X_n \downarrow X$ and $\mathbf{E}[X_1] < \infty$, then $\mathbf{E}[X_n] \downarrow \mathbf{E}[X]$.*
(iii) *If $X_n \geq Y$, then $\mathbf{E}[\liminf_{n\to\infty} X_n] \leq \liminf_{n\to\infty} \mathbf{E}[X_n]$.*
(iv) *If $X_n \leq Y$, then $\mathbf{E}[\limsup_{n\to\infty} X_n] \geq \limsup_{n\to\infty} \mathbf{E}[X_n]$.*
(v) *If $X_n \to X$ and $|X_n| \leq Y$ for all $n \in \mathbb{N}$, then $\lim_{n\to\infty} \mathbf{E}[X_n] = \mathbf{E}[X]$.*

For the monotone convergence theorems, i.e. parts (i) and (ii), we do not need to assume that $X \in L^1$. Then it is possible in part (i) that $\mathbf{E}[X] = \infty$, and in part(ii) that $\mathbf{E}[X] = -\infty$.

Proof. (i) Assume that X and X_n, $n \in \mathbb{N}$, are positive. By Theorem 4.1.11(ii) we can choose a sequence $(Y_{n,k})_{k\in\mathbb{N}} \subset \mathcal{E}$, such that $Y_{n,k} \uparrow X_n$ and $\mathbf{E}[Y_{n,k}] \uparrow \mathbf{E}[X_n]$. Let us then consider the maximum

$$Z_k \;:=\; \max_{n\leq k} Y_{n,k}.$$

Now $Z_k \in \mathcal{E}$. Moreover, $(Z_k)_{k\in\mathbb{N}}$ is increasing because $Y_{n,k} \leq Y_{n,k+1}$ for all $n \in \mathbb{N}$. Since $(Z_k)_{k\in\mathbb{N}}$ is increasing, there exists a limit $Z := \lim_{k\to\infty} Z_k$. Now $Z_k \leq X_k$, because

$$\begin{aligned} Z_k &= \max(Y_{1,k}, \ldots, Y_{k,k}) \\ &\leq \max(X_1, \ldots, X_k) \\ &= X_k. \end{aligned}$$

Therefore, when $n \leq k$, we have that

$$Y_{n,k} \;\leq\; Z_k \;\leq\; X_k \;\leq\; X.$$

Taking limits first with respect to k and then with respect to n, we see that $Z = X$. Also, we see that

$$\mathbf{E}[Y_{n,k}] \;\leq\; \mathbf{E}[Z_k] \;\leq\; \mathbf{E}[X_k],$$

when $n \leq k$. Keeping n fixed and letting k increase we see that

$$\mathbf{E}[X_n] \;\leq\; \mathbf{E}[Z] \;\leq\; \lim_{k\to\infty} \mathbf{E}[X_k]\,.$$

Letting then n increase, we see that

$$\lim_{n\to\infty} \mathbf{E}[X_n] \;\leq\; \mathbf{E}[Z] \;\leq\; \lim_{n\to\infty} \mathbf{E}[X_k].$$

But the claim follows from this, since $Z = X$.

Finally, we need to get rid of the assumption that X and X_n, $n \in \mathbb{N}$, are positive. We leave this as exercise.

(ii) This is the mirror image of part (i).

(iii) Let $Z_n := \inf_{k\geq n} X_k$ Then $Z_n \uparrow \liminf_{n\to\infty} X_n$. By using part (i), we obtain

$$\mathbf{E}[Z_n] \;\uparrow\; \mathbf{E}\Big[\liminf_{n\to\infty} X_n\Big]\,. \tag{4.2.2}$$

On the other hand $X_k \geq Z_n$, when $k \geq n$. Therefore $\mathbf{E}[X_k] \geq \mathbf{E}[Z_n]$, when $k \geq n$. Consequently, $\inf_{k\geq n} \mathbf{E}[X_k] \geq \mathbf{E}[Z_n]$. Taking limits with respect to n, we obtain

$$\liminf_{n\to\infty} \mathbf{E}[X_n] \;=\; \lim_{n\to\infty} \inf_{k\geq n} \mathbf{E}[X_k] \;\geq\; \lim_{n\to\infty} \mathbf{E}[Z_n]. \tag{4.2.3}$$

The claim follows by combining (4.2.2) and (4.2.3).

(iv) This is the mirror image of part (iii).

(v) The claim follows from parts (iii) and (iv). Indeed, we have

$$\begin{aligned}
\mathbf{E}[X] &= \mathbf{E}!\left[\liminf_{n\to\infty} X_n\right] \\
&\leq \liminf_{n\to\infty} \mathbf{E}[X_n] \\
&\leq \limsup_{n\to\infty} \mathbf{E}[X_n] \\
&\leq \mathbf{E}\left[\limsup_{n\to\infty} X_n\right] \\
&= \mathbf{E}[X],
\end{aligned}$$

and the claim $\liminf_{n\to\infty} \mathbf{E}[X_n] = \limsup_{n\to\infty} \mathbf{E}[X_n] = \mathbf{E}[X]$ follows from the *hamburger principle*. □

If some property A holds with probability 1, i.e. $\mathbf{P}[A] = 1$, we say that A holds *almost surely*. We use the abbreviation a.s. for almost surely.

4.2.4 Corollary. (i) *Let X_n, $n \in \mathbb{N}$, be positive random variables. Then*

$$\mathbf{E}\left[\sum_{n=1}^{\infty} X_n\right] = \sum_{n=1}^{\infty} \mathbf{E}[X_n],$$

where the equality can take the form $\infty = \infty$.

(ii) *Let $X_n, n \in \mathbb{N}$, random variables satisfying*

$$\sum_{n=1}^{\infty} \mathbf{E}[|X_n|] < \infty.$$

Then the series $\sum_{n=1}^{\infty} X_n$ converses almost surely, i.e.

$$\mathbf{P}\left[\omega \in \Omega\,;\, \sum_{n=1}^{\infty} X_n(\omega) \text{ converges}\right] = 1.$$

Moreover,

$$\mathbf{E}\left[\sum_{n=1}^{\infty} X_n\right] = \sum_{k=1}^{\infty} \mathbf{E}[X_n].$$

Proof. (i) We leave this part as Exercise 4.9.

(ii) Let $S := \sum_{n=1}^{\infty} |X_n|$. By part (i) $\mathbf{E}[S] = \sum_{n=1}^{\infty} \mathbf{E}[|X_n|] < \infty$. Thus $\mathbf{P}[S = \infty] = 0$. So, the series $\sum_{n=1}^{\infty} |X_n|$ converges almost surely. This means that the series $\sum_{n=1}^{\infty} X_n$ converges absolutely almost surely. Hence, it converges almost surely. Finally, we note that $\sum_{k=1}^{n} X_k \le S$. Hence the claim $\mathbf{E}[\sum_{n=1}^{\infty} X_n] = \sum_{n=1}^{\infty} \mathbf{E}[X_n]$ follow from the Lebesgue's dominated convergence theorem 4.2.1(v). □

4.3 Expectations of Transformations

Theorem 4.3.1 below makes it possible to calculate the expectation of a transfrom of a random variable X, $f(X)$, without calculating the distribution $\mathbf{P}_{f(X)}$. Theorem 4.3.1 also provides an easy way to argue Corollary 4.1.14.

4.3.1 Theorem. *Let X be a random variable and let $f\colon \mathbb{R} \to \mathbb{R}$ be Borel measurable. If $f(X) \in L^1$, then we have the* change-of-variables *formula*

$$\begin{aligned} \mathbf{E}[f(X)] &= \int_{-\infty}^{\infty} f(x)\, \mathbf{P}_X[\mathrm{d}x] \\ &= \int_{-\infty}^{\infty} f(x)\, \mathrm{d}F_X(x). \end{aligned} \tag{4.3.2}$$

Moreover, $f(X) \in L^1$ *if and only if*

$$\int_{-\infty}^{\infty} \big|f(x)\big| \, \mathbf{P}_X[\mathrm{d}x] \;<\; \infty. \tag{4.3.3}$$

Proof. We only show the change-of-variables formula (4.3.2). The part (4.3.3) is left as an exercise.

Stop 1: Let us first assume that f is a simple function of the form $f = \mathbf{1}_B$ for some $B \in \mathcal{B}$. Then the claim is obvious, since

$$\mathbf{E}[\mathbf{1}_B(X)] \;=\; \mathbf{P}[X \in B] \;=\; \mathbf{P}_X[B] \;=\; \int_{-\infty}^{\infty} \mathbf{1}_B(x) \, \mathbf{P}_X[\mathrm{d}x].$$

Step 2: If the function f is simple, i.e. of the form $f = \sum_{k=1}^n a_k \mathbf{1}_{B_k}$, the claim follows as in Step 1 by using linearity. Indeed,

$$\begin{aligned}
\mathbf{E}[f(X)] \;&=\; \mathbf{E}\left[\sum_{k=1}^n a_k \mathbf{1}_{B_k}(X)\right] \\
&=\; \sum_{k=1}^n a_k \mathbf{E}[\mathbf{1}_{B_k}(X)] \\
&=\; \sum_{k=1}^n a_k \int_{-\infty}^{\infty} \mathbf{1}_{B_k}(x) \, \mathbf{P}_X[\mathrm{d}x] \\
&=\; \int_{-\infty}^{\infty} \sum_{k=1}^n a_k \mathbf{1}_{B_k}(x) \, \mathbf{P}_X[\mathrm{d}x] \\
&=\; \int_{-\infty}^{\infty} f(x) \, \mathbf{P}_X[\mathrm{d}x].
\end{aligned}$$

Step 3: Finally, we approximate the functions f with simple functions f_n, $n \in \mathbb{N}$, such that $f_n \le |f|$ and the claim follows from the Lebesgue's dominated convergence theorem 4.2.1(v). □

4.3.4 Example. If X is uniformly distributed on the interval $[0,1]$ and $Y = \mathrm{e}^X$, then

$$\mathbf{E}[Y] \;=\; \mathbf{E}\big[\mathrm{e}^X\big] \;=\; \int_0^1 \mathrm{e}^x \, \mathrm{d}x \;=\; \mathrm{e} - 1.$$

So, we did not have to know the distribution of Y. We find it out, just for fun:

$$\mathbf{P}[Y \le y] \;=\; \mathbf{P}[\mathrm{e}^X \le y] \;=\; \mathbf{P}[X \le \ln y] \;=\; \ln y \mathbf{1}_{[1,\mathrm{e}]}(y) + \mathbf{1}_{(\mathrm{e},\infty)}(y),$$

We see that Y is continuously distributed with cumulative distribution function and density function given as

$$F_Y(y) \;=\; \ln y \mathbf{1}_{[1,\mathrm{e}]}(y) + \mathbf{1}_{(\mathrm{e},\infty)}(y) \quad \text{and} \quad f_Y(y) \;=\; \frac{1}{y}\mathbf{1}_{[1,\mathrm{e}]}(y).$$

4.4 Inequalities

For a function $f\colon \mathbb{R} \to \mathbb{R}$ its l_x *tangent* at point $x \in \mathbb{R}$ is any line tangential to the point $(x, f(x)) \in \mathbb{R}^2$: $l_x(y) = a(y - x) + f(x)$, where the slope $a = a(x) \in \mathbb{R}$. Tangentiality at point $(x, f(x)) \in \mathbb{R}^2$ means that $l_x(x) = f(x)$ and the graphs $\{(y, f(y)); y \in \mathbb{R}\}$ and $\{(y, l_x(y)); y \in \mathbb{R}\}$ do not cross in an open set containing the point $(x, f(x))$.

A function $f\colon \mathbb{R} \to \mathbb{R}$ is *convex*, if it is above its tangent lines: $f \geq l_x$ for all tangents l_x of f. Another way to define a convex function is to insist that

$$f\big(tx + (1-t)y\big) \;\leq\; tf(x) + (1-t)f(y)$$

for all $t \in [0, 1]$ and $x, y \in \mathbb{R}$. If the function f is smooth enough, then it is convex if and only if $f''(x) \geq 0$ for all $x \in \mathbb{R}$. Also, it can be shown that convex functions are continuous, and hence they are Borel measurable.

4.4.1 Theorem. *Let $f\colon \mathbb{R} \to \mathbb{R}$ be a convex function and let $X \in L^1$ be a random variable. Then we have the* Jensen's inequality

$$\mathbf{E}\big[f(X)\big] \;\geq\; f\big(\mathbf{E}[X]\big),$$

Proof. We first note that the claim is well-defined. Convex functions are continuous by Exercise 4.11) and continuous functions are Borel measurable. Thus $f(X)$ is a random variable.

Let the $l = l_{\mathbf{E}[X]}$ be the tangent of X at point $\mathbf{E}[X]$. Then, since f is convex

$$\begin{aligned} f(X) &\geq l(X) \\ &= a(X - \mathbf{E}[X]) + \mathbf{E}[f(X)]. \end{aligned}$$

Therefore

$$\begin{aligned} \mathbf{E}\big[f(X)\big] &\geq \mathbf{E}\left[a(X - \mathbf{E}[X]) + f(\mathbf{E}[X])\right] \\ &= a\mathbf{E}\big[X - \mathbf{E}[X]\big] + f(\mathbf{E}[X]) \\ &= f(\mathbf{E}[X]), \end{aligned}$$

which shows the claim. □

By using the Jensen's inequality, we can prove the "great inequalities" of measure theory: Hölder, Minkowski and Schwarz.

4.4.2 Corollary. *Let X and Y be random variable and let $p, q \geq 1$.*

(i) *Let* $1/p + 1/q = 1$ *and* $\mathbf{E}[|X|^p], \mathbf{E}[|Y|^q] < \infty$. *Then the following* Hölder's inequality *holds*

$$|\mathbf{E}[XY]| \;\leq\; \mathbf{E}[|XY|] \;\leq\; \mathbf{E}[|X|^p]^{\frac{1}{p}}\mathbf{E}[|Y|^q]^{\frac{1}{q}}$$

(ii) *Let* $p \geq 1$ *and* $\mathbf{E}[|X|^p], \mathbf{E}[|Y|^p] < \infty$. *Then the following* Minkowski's inequality *holds*

$$\mathbf{E}[|X+Y|^p]^{\frac{1}{p}} \;\leq\; \mathbf{E}[|X|^p]^{\frac{1}{p}} + \mathbf{E}[|Y|^p]^{\frac{1}{p}}.$$

(iii) *Let* $\mathbf{E}[X^2], \mathbf{E}[Y^2] < \infty$. *Then the following* Schwarz's inequality *holds*

$$|\mathbf{E}[XY]| \;\leq\; \mathbf{E}[|XY|] \;\leq\; \sqrt{\mathbf{E}[X^2]\mathbf{E}[Y^2]}.$$

Proof. (i) The first inequality is obvious. We can therefore assume that $X, Y \geq 0$. Morevover, we can assume that $\mathbf{E}[X^p] > 0$, since otherwise $\mathbf{P}[X = 0] = 1$ and the inequality takes the from $0 \leq 0$. The proof is based on the change of probability measure and Jensen's inequality. Let $A \in \mathcal{F}$. Define a probability measure on $(\Omega, \mathcal{F})$ as

$$\mathbf{Q}[A] \;:=\; \frac{\mathbf{E}[X^p \mathbf{1}_A]}{\mathbf{E}[X^p]}.$$

We denote by $\mathbf{E}_\mathbf{Q}$ the expectation with respect to this probability measure. Let then

$$Z \;:=\; \frac{Y}{X^{p-1}}\mathbf{1}_{\{X>0\}}.$$

Then

$$\begin{aligned}
\mathbf{E}[|XY|] &= \mathbf{E}[XY] \\
&= \mathbf{E}\big[XY\mathbf{1}_{\{X>0\}}\big] \\
&= \mathbf{E}\left[\frac{Y}{X^{p-1}}\mathbf{1}_{\{X>0\}}X^p\right] \\
&= \mathbf{E}[ZX^p] \\
&= \mathbf{E}[X^p]\mathbf{E}_\mathbf{Q}[Z]
\end{aligned}$$

Since $p \geq 1$, the mapping $x \mapsto |x|^p$ is convex. Therefore, by Jensen's inequality 4.4.1, we have that $\mathbf{E}_\mathbf{Q}[Z]^q \leq \mathbf{E}_\mathbf{Q}[Z^q]$. Consequently,

$$\begin{aligned}
\mathbf{E}[|XY|] &\leq \mathbf{E}[X^p]\mathbf{E}_\mathbf{Q}[Z^q]^{\frac{1}{q}} \\
&= \mathbf{E}[X^p]^{1-\frac{1}{q}}\mathbf{E}\left[\left(\frac{Y}{X^{p-1}}\mathbf{1}_{\{X>0\}}\right)^q X^p\right]^{\frac{1}{q}} \\
&= \mathbf{E}[X^p]^{1-\frac{1}{q}}\mathbf{E}\left[\frac{Y^q}{X^{(p-1)q}}X^p\right]^{\frac{1}{q}} \\
&= \mathbf{E}[X^p]^{\frac{1}{p}}\mathbf{E}[Y^q]^{\frac{1}{q}}.
\end{aligned}$$

In the last equality we used the fact that $1 - 1/q = 1/p$ ja $(p-1)q = p$.

(ii) If $p = 1$, the claim follows from the *triangle inequality* $|a+b| \leq |a| + |b|$. Hence, we assume that $p > 1$. The proof is based on part (i), the Hölder's inequality. Let $q = \frac{p}{p-1}$. Then

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

which shows the claim.

(iii) The Schwarz's inequality follow from the Hölder's inequality by taking $p = q = 2$. □

4.4.3 Theorem. *Let $a > 0$ and let $X \in L^1$ be positive. Then the following* Chebyshev's inequality

$$\mathbf{P}[X \geq a] \leq \frac{\mathbf{E}[X]}{a}.$$

Proof. Since $a\mathbf{1}_{\{X \geq a\}} \leq X\mathbf{1}_{\{X \geq a\}} \leq X$, we have that

$$a\mathbf{P}[X \geq a] = a\mathbf{E}\big[\mathbf{1}_{\{X \geq a\}}\big] = \mathbf{E}\big[a\mathbf{1}_{\{X \geq a\}}\big] \leq \mathbf{E}\big[X\mathbf{1}_{\{X \geq a\}}\big] \leq \mathbf{E}[X]$$

The claim follows by dividing by a the both sides. □

The Chebyshev's inequality can be improved if the random variable is "integrable enough". This is the message of Corollary 4.4.4

4.4.4 Corollary. (i) *Let $\mathbf{E}[X^2] < \infty$ and $a > 0$. Then the following* Markov's inequality *holds*

$$\mathbf{P}\big[|X - \mathbf{E}[X]| \geq a\big] \leq \frac{\mathbf{Var}[X]}{a^2},$$

where

$$\mathbf{Var}[X] := \mathbf{E}[(X - \mathbf{E}[X])^2]$$

is the variance *of the random variable X.*

(ii) *Let*

$$\Lambda(\theta) := \ln \mathbf{E}\big[\mathrm{e}^{\theta X}\big]$$

be the cumulant generating function *of the random variable X. We assume the* Cramér's condition*: $\Lambda(\theta) < \infty$ for all $\theta \in \mathbb{R}$. Then the following* Chernoff's inequality *holds*

$$\mathbf{P}[X \geq a] \leq \mathrm{e}^{-\Lambda^*(a)},$$

where

$$\Lambda^*(a) := \sup_{\theta\in\mathbb{R}}\{a\theta - \Lambda(\theta)\}$$

is the Fenchel's and Legendre's transform *of the function* Λ.

Proof. (i) The Markov's inequality follows from the Chebyshev's inequality by substituting $(X - \mathbf{E}[X])^2$ instead of X. Indeed,

$$\begin{aligned}\mathbf{P}\big[\big|X - \mathbf{E}[X]\big| \geq a\big] &= \mathbf{P}\big[(X - \mathbf{E}[X])^2 \geq a^2\big] \\ &\leq \frac{\mathbf{E}\big[(X - \mathbf{E}[X])^2\big]}{a^2} \\ &= \frac{\mathbf{Var}[X]}{a^2}.\end{aligned}$$

(ii) Let $\theta \in \mathbb{R}$. Then, by the Chebyschev's inequality,

$$\begin{aligned}\mathbf{P}[X \geq a] &= \mathbf{P}\left[\mathrm{e}^{\theta X} \geq \mathrm{e}^{\theta a}\right] \\ &\leq \frac{\mathbf{E}\big[\mathrm{e}^{\theta X}\big]}{\mathrm{e}^{\theta a}} \\ &= \mathrm{e}^{\Lambda(\theta)-\theta a} \\ &= \mathrm{e}^{-(\theta a - \Lambda(\theta))}.\end{aligned}$$

Chernoff's inequality follows from this by optimizing over the free parameter $\theta \in \mathbb{R}$. □

4.4.5 Example. In an exam there are n multiple choice questions with four choices. The student needs to get at lest 40% right. The designer of the exam does not want students who just guesses to pass the exam. How large need n be so that a pure guesser's probability of passing the exam is less than 0.1%?

The problem can be solved explicitly by noticing that if X is the number of points of guessing student, then X is binomially distributed with parameters n and $1/4$. The designer of the exam does not want to calculate binomial probabilities explicitly, but wants just to estimate the probabilities.

Chebychev Now $\mathbf{E}[X] = n/4$. Therefore,

$$\mathbf{P}[X \geq 0.4n] \leq \frac{\mathbf{E}[X]}{0.4n} \leq \frac{n/4}{0.4n} = \frac{1}{1.6} = 0.625.$$

So, the Chebyshev's inequality does not give us n.

Markov: We use a modified Markov's inequality

$$\mathbf{P}[X \geq a] \;\leq\; \frac{\mathbf{E}[X^2]}{a^2},$$

We need to calculate $\mathbf{E}[X^2]$. We leave this as Exercise 4.14 and simply note tha $\mathbf{E}[X^2] = 3n/16 + n^2/16$. We get the estimate

$$\mathbf{P}[X \geq 0.4n] \;\leq\; \frac{\mathbf{E}[X^2]}{(0.4n)^2} \;=\; \frac{5(3+n)}{128n}.$$

So, the Markov's inequality does not give us an n either.

Chernoff: Let us calculate the cumulant generating function by using *Newton's binomial formula*

$$\sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k} \;=\; (a+b)^n.$$

We see that

$$\begin{aligned}
\Lambda(\theta) &= \ln \sum_{k=0}^{n} \mathrm{e}^{\theta k} \mathbf{P}[X=k] \\
&= \ln \sum_{k=0}^{n} \mathrm{e}^{\theta k} \binom{n}{k} \left(\frac{1}{4}\right)^k \left(1-\frac{1}{4}\right)^{n-k} \\
&= \ln \sum_{k=0}^{n} \binom{n}{k} \left(\mathrm{e}^{\theta}\frac{1}{4}\right)^k \left(\frac{3}{4}\right)^{n-k} \\
&= \ln \left(\mathrm{e}^{\theta}\frac{1}{4} + \frac{3}{4}\right)^n \\
&= n \ln \left(\mathrm{e}^{\theta}\frac{1}{4} + \frac{3}{4}\right).
\end{aligned}$$

Then we solve the optimization problem $\Lambda^*(a) = \sup_{\theta\in\mathbb{R}}\{a\theta - \Lambda(\theta)\}$ in the normal way by looking at the roots of the derivative:

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\theta}\{a\theta - \Lambda(\theta)\} &= \frac{\mathrm{d}}{\mathrm{d}\theta}\left\{a\theta - n\ln\left(\mathrm{e}^{\theta}\frac{1}{4} + \frac{3}{4}\right)\right\} \\
&= a - n\frac{\mathrm{e}^{\theta}/4}{\mathrm{e}^{\theta}/4 + 3/4}.
\end{aligned}$$

Therefore, the optimal point $\theta = \theta^*$ is given by the formula

$$a - n\frac{\mathrm{e}^{\theta^*}}{\mathrm{e}^{\theta^*}+3} \;=\; 0.$$

Consequently,

$$\theta^* = \ln \frac{3a}{n-a},$$

and

$$\begin{aligned} \Lambda^*(a) &= a\theta^* - \Lambda(\theta^*) \\ &= a\ln 3 + a\ln\frac{a}{n-a} - n\ln 3 + 2n\ln 2 - n\ln\frac{n}{n-a}. \end{aligned}$$

We obtain the estimate

$$\begin{aligned} \mathbf{P}[X \geq 0.4n] &\leq e^{-\Lambda^*(0.4n)} \\ &\leq 0.948^n \end{aligned}$$

Therefore $0.948^n \leq 0.1\%$, which yields that 130 multiple choice questions is enough.

Finally, we give the exact solution. Now

$$\begin{aligned} \mathbf{P}[X \geq 0.4n] &= \sum_{k=0.4n}^{n} \mathbf{P}[X = k] \\ &= \sum_{k=0.4n}^{n} \binom{n}{k} \left(\frac{1}{4}\right)^k \left(\frac{3}{4}\right)^{n-k}. \end{aligned}$$

By checking with a computer, we see that 92 multiple choice questions is enough.

Exercises

4.1. Let the random variible X have the cumulative distribution function

$$F_X(x) = \frac{1}{\pi}\arctan x + \frac{1}{2}.$$

Is the random variable X integrable?

4.2. Prove Lemma 4.1.10.

4.3. Show that L^1 is not an algebra, i.e. find such $X, Y \in L^1$ that $XY \notin L^1$.

4.4. Prove Corollary 4.1.14.

4.5. Show that $X \in L^1$ if and only if

$$\sum_{n=0}^{\infty} \mathbf{P}[|X| \geq n] \;<\; \infty.$$

4.6. Let X be Poisson distributed with parameter λ, i.e. X is discrete with range $\mathbb{N} \cup \{0\}$ and

$$\mathbf{P}[X = k] \;=\; \mathrm{e}^{-\lambda} \frac{\lambda^k}{k!}, \qquad k \in \mathbb{N} \cup \{0\}.$$

Calculate $\mathbf{E}[X]$ and $\mathbf{Var}[X]$.

4.7. Let X be *normally distributed* with parameters μ and σ^2, i.e. X is continuously distributed with distribution function

$$\mathbf{P}[X \in B] \;=\; \int_B \frac{1}{\sigma} \varphi\left(\frac{x-\mu}{\sigma}\right) \mathrm{d}x,$$

where φ is the standard normal density function

$$\varphi(x) \;:=\; \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}x^2}.$$

Calculate $\mathbf{E}[X]$ and $\mathbf{Var}[X]$.

4.8. Let X be *Cauchy distributed*, i.e. X is continuous with distribution function

$$\mathbf{P}[X \in B] \;=\; \frac{1}{\pi} \int_B \frac{1}{1+x^2} \,\mathrm{d}x.$$

Calculate $\mathbf{E}[X]$ and $\mathbf{Var}[X]$.

4.9. Prove the part (i) of Corollary 4.2.4.

4.10. Prove the integrability condition (4.3.3) of the Change-of-Variables Theorem 4.3.1.

4.11. Show that convex functions are continuous.

4.12. Show that twice differentiable function $f\colon \mathbb{R} \to \mathbb{R}$ is convex if and only if f'' is positive.

4.13. Let X be positive random variable for which $\mathbf{E}[X] > 0$. Let $A \in \mathcal{F}$. We set

$$\mathbf{Q}[A] \;:=\; \frac{\mathbf{E}[X \mathbf{1}_A]}{\mathbf{E}[X]}.$$

Show that $\mathbf{Q}$ is a probability measure on $(\Omega, \mathcal{F})$.

4.14. Let X be binomially distributed with parameters n and p. Calculate $\mathbf{E}[X^2]$.

Chapter 5

Independence

The idea of independence, or rather stochastic independence, is not causal independence. Indeed, in the causal world all events depend on each others. The stochastic independence means non-predictability in the following sense: the events A and B are independent, if the occurrence of B does not change our view of the probability of the occurrence of A. In other words $\mathbf{P}[A|B] = \mathbf{P}[A]$. Since $\mathbf{P}[A|B] = \mathbf{P}[A \cap B]/\mathbf{P}[B]$, we see that it is equivalent to state that $\mathbf{P}[A \cap B] = \mathbf{P}[A]\mathbf{P}[B]$, i.e. independence means multiply.

The aim of this chapter is to extend the notion of independence to σ-algebras. Since a σ-algebra can be considered information, this extension is natural.

5.1 Independent σ-algebras, Random Variables and Events

5.1.1 Definition. Let J some index set.

(a) The sub-σ-algebras $\mathcal{G}_j \subset \mathcal{F}$, $j \in J$, are *independent*, if for all finite collection of different indices $j_1, \ldots, j_n \in J$ and for all $A_k \in \mathcal{G}_{j_k}$ we have

$$\mathbf{P}\left[\bigcap_{k=1}^{n} A_k\right] \;=\; \prod_{k=1}^{n} \mathbf{P}[A_k]\,.$$

The sub-σ-algebras are *pairwise independent*, if

$$\mathbf{P}[A_k \cap A_l] \;=\; \mathbf{P}[A_k]\mathbf{P}[A_l]$$

for all different pairs $k, l \in J$ and $A_k \in \mathcal{G}_k$, $A_l \in \mathcal{G}_l$.

(b) The random variables X_j, $j \in J$, are (pairwise) independent if the σ-algebras $\sigma(X_j)$, $j \in J$, they generate are (pairwise) indepedent. Here the σ-algebra $\sigma(X)$ generated by a random variable is the smallest σ-algebra on $(\Omega, \mathcal{F})$ such that X is measurable.

(c) The events $A_j \in \mathcal{F}$, $j \in J$, are (pairwise) independent, if the corresponding indicators $\mathbf{1}_{A_j}$, $j \in J$, are (pairwise) independent.

We note that "observing" a random variable X means "observing" the σ-algebra $\sigma(X)$ it generates, i.e. for all $A \in \sigma(X)$ we know if A has occurred or not. This is the information interpretation of a σ-algebra.

5.1.2 Example. Because $\sigma(\mathbf{1}_{A_i}) = \{\varnothing, \Omega, A_i, A_i^c\}$, we see that the events A_1 and A_2 are independent if and oly if $\mathbf{P}[A_1 \cap A_2] = \mathbf{P}[A_1]\mathbf{P}[A_2]$. Indeed, it follows that $\mathbf{P}[A_1 \cap A_2^c] = \mathbf{P}[A_1]\mathbf{P}[A_2^c]$, $\mathbf{P}[A_1^c \cap A_2] = \mathbf{P}[A_1^c]\mathbf{P}[A_2]$ and $\mathbf{P}[A_1^c \cap A_2^c] = \mathbf{P}[A_1^c]\mathbf{P}[A_2^c]$ (so, if the events A_1 and A_2 are independent, the so are also the events A_1 and A_2^c, A_1^c and A_2, and A_1^c ja A_2^c). Thus, Definition 5.1.1(c) is a generalization of the well-known elementary definition of independence.

In the following

$$F_{X,Y}(x,y) \;:=\; \mathbf{P}[X \le x, Y \le y]$$

is the *joint probability distribution function* of the pair (X,Y) of random variables.

5.1.3 Theorem. *The following are equivalent:*

(i) *The random variables X and Y are independent.*

(ii) *Foa ll $A, B \in \mathcal{B}$ we have*

$$\mathbf{P}[X \in A, Y \in B] \;=\; \mathbf{P}[X \in A]\,\mathbf{P}[Y \in B].$$

(iii) *For $A, B \in \mathcal{I}$, where $\mathcal{I}$ is a π-class with $\mathcal{B} = \sigma(\mathcal{I})$ we have*

$$\mathbf{P}[X \in A, Y \in B] \;=\; \mathbf{P}[X \in A]\,\mathbf{P}[Y \in B].$$

(iv) *The joint cumulative distribution function satisfies*

$$F_{X,Y}(x,y) \;=\; F_X(x)\,F_Y(y),$$

(v) *The random variables $f(X)$ are $g(Y)$ are independent for all Borel measurable functions $f, g\colon \mathbb{R} \to \mathbb{R}$.*

(vi)

$$\mathbf{E}[f(X)g(Y)] \;=\; \mathbf{E}[f(X)]\mathbf{E}[g(Y)]$$

for all bounded Borel measurable functions $f, g\colon \mathbb{R} \to \mathbb{R}$.

Proof. (i) $\Leftrightarrow$ (ii): Part (ii) is simply a reformulation of the definition. Indeed, $\sigma(X)$ is the collection $\{X \in A\}$, $A \in \mathcal{B}$ and $\sigma(Y)$ is the collection $\{Y \in B\}$, $B \in \mathcal{B}$.

(ii) $\Rightarrow$ (iii): This implication is clear.

(ii) $\Leftarrow$ (iii): This implication follows from the Dynkin's extension theorem 2.3.8.

(ii) $\Rightarrow$ (iv): This implication is clear.

(ii) $\Leftarrow$ (iv): This implication follows from the Dynkin's extension theorem 2.3.8.

(ii) $\Leftarrow$ (v): This implication is left as Exercises 5.4.

(ii) $\Rightarrow$ (v): Note that $\sigma(f(X)) \subset \sigma(X)$ and $\sigma(g(Y)) \subset \sigma(Y)$. Indeed, $\sigma(f(X))$ is the collection $\{f(X) \in B\}$, $B \in \mathcal{B}$. But

$$\{f(X) \in B\} \;=\; \{X \in f^{-1}B\},$$

and $f^{-1}B \in \mathcal{B}$, because f is Borel measurable. Consequently, $\{f(X) \in B\} \in \sigma(X)$. In the same way we see that $\{g(Y) \in B\} \in \sigma(Y)$. The claim follows now from the fact that the independence of σ-algebras is inherited by sub-σ-algebras (Exercise 5.5).

(ii) $\Leftarrow$ (vi): This part is left as Exercise 5.6.

(ii) $\Rightarrow$ (vi): By (ii) we have

$$\mathbf{E}[f(X)g(Y)] \;=\; \mathbf{E}[f(X)]\mathbf{E}[g(Y)] \tag{5.1.4}$$

when f and g are indicators: $f = \mathbf{1}_A$ ja $g = \mathbf{1}_B$ for some $A, B \in \mathcal{B}$. But it follows from the linearity of the expectation that for all simple functions $f = \sum_{k=1}^{n} a_k \mathbf{1}_{A_k}$ and $g = \sum_{l=1}^{n'} b_l \mathbf{1}_{B_l}$ the formula (5.1.4) holds. Finally, the claim follows by approximating f and g with simple functions and by using the Lebesgue's dominated convergence theorem 4.2.1(v). We can use Theorem 4.2.1(v) since f and g are bounded; hence $f(X), g(Y) \in L^1$. □

Theorem 5.1.3 has a natural extension to the family X_j, $j \in J$, of random variables. We leave it to the reader to figure out the extension.

5.2 Borel's and Cantelli's Lemmas

Theorem 5.2.1 below, the so-called Borel's and Cantelli's lemmas, is one of the most important auxiliary results of probability theory. It is, for example, instrumental in proving later the stong law of large numbers.

5.2.1 Theorem. *Let A_n, $n \in \mathbb{N}$, be events. Then the following* Borel's and Cantelli' lemmas *hold:*

(i) *If*

$$\sum_{n=1}^{\infty} \mathbf{P}[A_n] \;<\; \infty,$$

then

$$\mathbf{P}\left[\limsup_{n\to\infty} A_n\right] \;=\; 0.$$

(ii) *If the events* A_n, $n \in \mathbb{N}$, *are independent and*

$$\sum_{n=1}^{\infty} \mathbf{P}[A_n] \;=\; \infty,$$

then

$$\mathbf{P}\left[\limsup_{n\to\infty} A_n\right] \;=\; 1.$$

Proof. (i) Set $B_n := \cup_{k=1}^{n} A_k$. Then $\limsup_{n\to\infty} A_n = \cap_{n=1}^{\infty} B_n$. Since $\cap_{n=1}^{\infty} B_n \subset B_m$ for all $m \in \mathbb{N}$, we have

$$\mathbf{P}\left[\bigcap_{n=1}^{\infty} B_n\right] \;\le\; \mathbf{P}[B_m] \;\le\; \sum_{k=m}^{\infty} \mathbf{P}[A_k].$$

But, as the remainder of a convergent series, we have $\sum_{k=m}^{\infty} \mathbf{P}[A_k] \to 0$, as $m \to \infty$. Therefore $\mathbf{P}[\limsup_{n\to\infty} A_n] = \mathbf{P}[\cap_{n=1}^{\infty} B_n] = 0$.

(ii) Because

$$\left(\limsup_{n\to\infty} A_n\right)^c \;=\; \bigcup_{m=1}^{\infty} \bigcap_{n=m}^{\infty} A_n^c \;=\; \liminf_{n\to\infty} A_n^c,$$

we have

$$\mathbf{P}\left[\left(\limsup_{n\to\infty} A_n\right)^c\right] \;=\; \lim_{m\to\infty} \mathbf{P}\left[\bigcap_{n=m}^{\infty} A_n^c\right].$$

Let then $N > m$ be fixed. Then, by independence,

$$\begin{aligned}
\mathbf{P}\left[\bigcap_{n=m}^{\infty} A_n^c\right] \;&\le\; \mathbf{P}\left[\bigcap_{n=m}^{N} A_n^c\right] \\
&=\; \prod_{n=m}^{N} \mathbf{P}[A_n^c] \\
&=\; \prod_{n=m}^{N} \big(1 - \mathbf{P}[A_n]\big).
\end{aligned}$$

Let us then use the estimate $1 - x \leq \mathrm{e}^{-x}$, for $x \geq 0$. Then

$$\mathbf{P}\left[\bigcap_{n=m}^{\infty} A_n^c\right] \;\leq\; \exp\left\{-\sum_{n=m}^{N} \mathbf{P}[A_n]\right\} \;\to\; 0,$$

as $N \to \infty$, because we have a divergent series in the exponent. Therefore we have for each $m \in \mathbb{N}$ that

$$\mathbf{P}\left[\bigcap_{n=m}^{\infty} A_n^c\right] \;=\; 0.$$

Consequently,

$$\mathbf{P}\left[\left(\limsup_{n\to\infty} A_n\right)^c\right] \;=\; \lim_{m\to\infty} \mathbf{P}\left[\bigcap_{n=m}^{\infty} A_n^c\right] \;=\; 0,$$

or $\mathbf{P}[\limsup_{n\to\infty} A_n] = 1 - \mathbf{P}[(\limsup_{n\to\infty} A_n)^c] = 1 - 0 = 1$. □

Example 5.2.2(ii) below is a famous example of monkey typing Shakespeare.

5.2.2 Example. (i) We consider infinite dice rolling. Let

$$A \;:=\; \text{“there are infinite number of sixes in the series of dice rolls”}.$$

If $A_n :=$ “n^{th} roll is 6”, then $A = \limsup_{n\to\infty} A_n$. Clearly, the events A_n, $n \in \mathbb{N}$, are independent. On the other hand $\mathbf{P}[A_n] = 1/6$. So, $\sum_{n=1}^{\infty} \mathbf{P}[A_n] = \infty$. Therefore $\mathbf{P}[A] = 1$ by Borel's and Cantelli's lemma Theorem 5.2.1(ii).

(ii) Suppose a monkey with a typewriter. The monkey types the keys completely in random. There are N letters in the Shakespeare's combined works. Let A_n be the event “The monkey types Shakespeare's combined works in the key presses $(n-1)N + 1, \ldots, nN$”. The probability

$$p \;:=\; \mathbf{P}[A_n] \;=\; \left(\frac{1}{\text{number of keys in the typewriter}}\right)^N$$

is tiny, but positive. Now the event

$$A \;=\; \text{“The monkey types Shakespeare's combined works i.o.”}$$

includes the set $\limsup_{n\to\infty} A_n$. Since he events A_n, $n \in \mathbb{N}$, are independent, and $\sum_{n=1}^{\infty} \mathbf{P}[A_n] = \sum_{n=1}^{\infty} p = \infty$, it follows from Borel's and Cantelli's lemma, Theorem 5.2.1(ii) that the monlkey will almost surely write the Shakespeare's combined works infinitely often.

Next we present the so-called *Kolmogorov's* 0–1 *law* (Theorem 5.2.4). It provides a funnier way to prove the monkey typing Shakespeare Example 5.2.2(ii). Before the theorem we need the concept of tail-σ-algebra.

5.2.3 Definition. Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of random variable. We denote $\mathcal{G}_n := \sigma(X_k; k > n)$, $n \in \mathbb{N}$. Then the σ-algebra

$$\mathcal{G}_\infty \;:=\; \bigcap_{n=1}^{\infty} \mathcal{G}_n$$

is called the *tail* σ*-algebra* generated by the sequence $(X_n)_{n\in\mathbb{N}}$.

5.2.4 Theorem. *Let* $(X_n)_{n\in\mathbb{N}}$ *be a sequence of independent random variables. Then its tail* σ*-algebra* $\mathcal{G}_\infty$ *is trivial:* $\mathbf{P}[A] = 0$ *or* $\mathbf{P}[A] = 1$ *for all* $A \in \mathcal{G}_\infty$.

Proof. Denote $\mathcal{F}_n := \sigma(X_k; k \le n)$ and $\mathcal{G}_n := \sigma(X_k; k > n)$, $n \in \mathbb{N}$. By independence, the σ-algebras $\mathcal{F}_n$ are $\mathcal{G}_n$ are independent. Therefore, if $F \in \mathcal{F}_n$ and $G \in \mathcal{G}_n$, we have that

$$\mathbf{P}[F \cap G] \;=\; \mathbf{P}[F]\mathbf{P}[G]. \tag{5.2.5}$$

But if $G \in \mathcal{G}_\infty$, then (5.2.5) holds for all $F \in \cup_{n=1}^{\infty}\mathcal{F}_n$. Therefore, by Dynkin's extension theorem 2.3.8, (5.2.5) holds for all $F \in \sigma(X_1, X_2, \ldots) =: \mathcal{F}_\infty$. But $\mathcal{G}_\infty \subset \mathcal{F}_\infty$. Therefore, it follows from (5.2.5) that $\mathbf{P}[A] = \mathbf{P}[A]\mathbf{P}[A]$ for all $A \in \mathcal{G}_\infty$. But this is possible only if $\mathbf{P}[A] = 0$ or $\mathbf{P}[A] = 1$. □

5.2.6 Corollary. *Let* $(X_n)_{n\in}$ *be a sequence of independent random variables. Then*

(i) *either the sequence* $(X_n)_{n\in\mathbb{N}}$ *converges a.s. or it diverges a.s., i.e.*

$$\mathbf{P}\left[\omega \in \Omega;\ \lim_{n\to\infty} X_n(\omega) \text{ exists}\right] \;=\; 0 \text{ or } 1,$$

(ii) *The following random variables are a.s. extended constants*

$$\limsup_{n\to\infty} X_n, \qquad \liminf_{n\to\infty} X_n,$$

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

5.3 Product Spaces and Fubini's Theorem

Product spaces corresponds repeated random trials and product measures correspond independent trials.

Let $(\Omega_1, \mathcal{F}_1)$ and $(\Omega_2, \mathcal{F}_2)$ be two measurable spaces. Let

$$\mathcal{F}_1 \times \mathcal{F}_2 \;:=\; \{A_1 \times A_2 \,;\, A_1 \in \mathcal{F}_1, A_2 \in \mathcal{F}_2\}$$

be a product set $\Omega_1 \times \Omega_2$. The problrem is that the product sets $\mathcal{F}_1 \times \mathcal{F}_2$ do not form a σ-algebra. Indeed, the complement of the set $A_1 \times A_2$ cannot be written as $B_1 \times B_2$, for some $B_1 \in \mathcal{F}_1$, $B_2 \in \mathcal{F}_2$. In fact,

$$(A_1 \times A_2)^c \;=\; (A_1 \times A_2^c) \cup (A_1^c \times A_2) \cup (A_1^c \times A_2^c).$$

5.3.1 Definition. The *product σ-algebra* on the set $\Omega_1 \times \Omega_2$ is

$$\mathcal{F}_1 \otimes \mathcal{F}_2 \;:=\; \sigma\left(A_1 \times A_2 \,;\, A_1 \in \mathcal{F}_1, A_2 \in \mathcal{F}_2\right).$$

The space $(\Omega_1 \times \Omega_2, \mathcal{F}_1 \otimes \mathcal{F}_2)$ is called the *product space*. The sets $A_1 \times A_2$, $A_1 \in \mathcal{F}_1$, $A_2 \in \mathcal{F}_2$ are called *measurable rectangles*.

5.3.2 Example. We roll a dice twice. For one dice roll $\Omega_1 = \{1, \ldots, 6\}$ and $\mathcal{F}_1 = \mathrm{pow}\,(\Omega)$. For two dice rolls $\Omega = \Omega_1^2 = \{(i,j); i,j = 1, \ldots, 6\}$ and $\mathcal{F} = \mathcal{F}_1^{\otimes 2} = \mathrm{pow}\,(\Omega_1 \times \Omega_1)$. The event "first roll is 6 and the secon roll is even" is a measurable rectangle, but the event "the sum of the rolls is even" is not a measurable rectangle (or any rectangle, for that matter).

We denote by $f(x, \cdot)$ the function $y \mapsto f(x, y)$.

5.3.3 Lemma. *Let X be a random variable from the product space $(\Omega_1 \times \Omega_2, \mathcal{F}_1 \otimes \mathcal{F}_2)$. Let $\omega_1 \in \Omega_1$ be fixed. Then the* slice *$X(\omega_1, \cdot)$ is a random variable from the space $(\Omega_2, \mathcal{F}_2)$ for each $\omega_1 \in \Omega_1$. In the same way the slice $X(\cdot, \omega_2)$ is a random variable from the space $(\Omega_1, \mathcal{F}_1)$ for all $\omega_2 \in \Omega_2$.*

Proof. It is clear that is enough to show the claim for the slice $X(\omega_1, \cdot)$. The case for the slice $X(\cdot, \omega_2)$ is the same.

We use the now standard machinery: first X is an indicator (where we argue that the set where the claim holds is a σ-algebra thus containing the minimal σ-algebra in question), then X is simple, then X is positive, and finally $X = X^+ - X^-$.

Step 1: Let first $X = \mathbf{1}_A$, where $A \in \mathcal{F}_1 \otimes \mathcal{F}_2$. We consider the set of events

$$\mathcal{H} \;:=\; \{A \in \mathcal{F}_1 \otimes \mathcal{F}_2 \,;\, \mathbf{1}_A(\omega_1, \cdot) \text{ on } \mathcal{F}_2\text{-measurable for all } \omega_1 \in \Omega_1\}.$$

If A is a measurable rectangle $A_1 \times A_2$, then $\mathbf{1}_A(\omega_1, \omega_2) = \mathbf{1}_{A_1}(\omega_1)\mathbf{1}_{A_2}(\omega_2)$. Clearly $\mathbf{1}_{A_1}(\omega_1)\mathbf{1}_{A_2}(\cdot)$ is $\mathcal{F}_2$-measurable. Hence

$$\mathcal{F}_1 \times \mathcal{F}_2 \quad \subset \quad \mathcal{H}.$$

On the other hand, $\mathcal{H}$ is a σ-algebra (Exercise 5.8). Now, $\mathcal{H}$ contains measurable rectangles. So, $\mathcal{F}_1 \otimes \mathcal{F}_2 \subset \mathcal{H}$. From the definition it is clear that $\mathcal{H} \subset \mathcal{F}_1 \otimes \mathcal{F}_2$. Thus $\mathcal{H} = \mathcal{F}_1 \otimes \mathcal{F}_2$.

We deduce that the claim follows for the indicators $\mathbf{1}_A$, $A \in \mathcal{F}_1 \otimes \mathcal{F}_2$.

Step 2: Let X be simple, i.e.,

$$X(\omega_1, \omega_2) \quad = \quad \sum_{k=1}^{n} c_k \mathbf{1}_{A_k}(\omega_1, \omega_2),$$

where $c_k \in \mathbb{R}$ snd $A_k \in \mathcal{F}_1 \otimes \mathcal{F}_2$. By Step 1 $X(\omega_1, \cdot)$ is a linear combination of $\mathcal{F}_2$-measurable functions. Therefore it is $\mathcal{F}_2$-measurable.

Askel 3: Let then X be positive. Then there exists a sequence $(X_n)_{n\in\mathbb{N}}$ of simple random variables such that $X = \lim_{n\to\infty} X_n$. Then, for a fixed $\omega_1 \in \Omega$, we have $X(\omega_1, \cdot) = \lim_{n\to\infty} X_n(\omega_1, \cdot)$. Therefore, by Step 2, $X(\omega_1, \cdot)$ is, as a limit of $\mathcal{F}_2$-measurable random variables, itself $\mathcal{F}_2$-measurable.

Step 4: Finally, let X be a genaral random variable. The claim follows from Step 3 by noticing that we can write $X(\omega_1, \cdot) = X^+(\omega_1, \cdot) - X^-(\omega_1, \cdot)$ for all $\omega_1 \in \Omega_1$. □

5.3.4 Definition. Let $(\Omega_1, \mathcal{F}_1, \mathbf{P}_1)$ and $(\Omega_2, \mathcal{F}_2, \mathbf{P}_2)$ be probability spaces. The measure $\mathbf{P}_1 \otimes \mathbf{P}_2$ on the product space $(\Omega_1 \times \Omega_1, \mathcal{F}_1 \otimes \mathcal{F}_2)$ given by

$$(\mathbf{P}_1 \otimes \mathbf{P}_2)[A_1 \times A_2] \quad = \quad \mathbf{P}_1[A_1]\,\mathbf{P}_2[A_2] \tag{5.3.5}$$

is called the *product measure*

By the extension theorems Theorem 2.3.8 and Theorem 2.3.9 there exist a unique product measure on the product σ-algebra $\mathcal{F}_1 \times \mathcal{F}_1$ satisfying (5.3.5). With respect to the product measure $\mathbf{P}_1 \otimes \mathbf{P}_2$ the σ-algebras $\mathcal{F}_1$ ja $\mathcal{F}_2$, are independent.

5.3.6 Example. (i) In Example 5.3.2 $\mathbf{P}^{\otimes 2}[A] = |A|/36$.

(ii) Let $\Omega_1 = \Omega_2 = [0,1]$, $\mathcal{F}_1 = \mathcal{F}_2 = \mathcal{B}_{[0,1]}$ the Borel's σ-algebra generated by the open sets of the interval $[0,1]$. Let $\mathbf{P}_1 = \mathbf{P}_2 = \ell$ be the Lebesgue's measure "dx" on $[0,1]$. Then the product σ-algebra is generated by the open sets of $[0,1]^2$, and the product measure $\ell^{\otimes 2}$ is the two-dimensional Lebesgue's measure "dxdy" on the rectangle $[0,1]^2$.

The following result, Theorem 5.3.7 is the celebrated *Fubini's theorem*. With the Change-of-Variables Theorem 4.3.1 oit forms the most important theorems of integral calculus.

5.3.7 Theorem. *Let X be a random variable from the product space $(\Omega_1 \times \Omega_2, \mathcal{F}_1 \otimes \mathcal{F}_2)$ and let $\mathbf{P}_1 \otimes \mathbf{P}_2$ be the product measure. Then, if $X \geq 0$ or $X \in L^1$, the mapping*

$$\int_{\Omega_2} X(\cdot, \omega_2)\, \mathbf{P}_2[\mathrm{d}\omega_2]$$

is $\mathcal{F}_1$-measurable for all $\omega_2 \in \Omega_2$. Also, the mapping

$$\int_{\Omega_1} X(\omega_1, \cdot)\, \mathbf{P}_1[\mathrm{d}\omega_1]$$

is $\mathcal{F}_2$-measurable for all $\omega_1 \in \Omega_1$. Moreover, the Fubini's formulas

$$\begin{aligned}
&\int_{\Omega_1 \times \Omega_2} X(\omega_1, \omega_2)\, (\mathbf{P}_1 \otimes \mathbf{P}_2)[\mathrm{d}\omega_1, \mathrm{d}\omega_2] \\
&= \int_{\Omega_1} \left(\int_{\Omega_2} X(\omega_1, \omega_2)\, \mathbf{P}_2[\mathrm{d}\omega_2] \right) \mathbf{P}_1[\mathrm{d}\omega_1] \\
&= \int_{\Omega_2} \left(\int_{\Omega_1} X(\omega_1, \omega_2)\, \mathbf{P}_1[\mathrm{d}\omega_1] \right) \mathbf{P}_2[\mathrm{d}\omega_2],
\end{aligned}$$

hold. Here each line exists and is finite in the same time.

Proof. The claim can be shown by using the standard machinery and Lemma 5.3.3. We leave the details as Example 5.9. □

5.3.8 Corollary. *Let $X, Y \in L^1$ be independent random variables. Then $XY \in L^1$ and*

$$\mathbf{E}[XY] \;=\; \mathbf{E}[X]\mathbf{E}[Y]$$

Proof. The claim follows from the Fubini's theorem 5.3.7 and Theorem 5.1.3(iv). Indeed, by using the cumulative distribution functions we can

write

$$\begin{aligned}
\mathbf{E}[XY] &= \iint_{\mathbb{R}^2} xy \,\mathrm{d}F_{X,Y}(x,y) \\
&= \iint_{\mathbb{R}^2} xy \,\mathrm{d}F_X(x)\mathrm{d}F_Y(y) \\
&= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} xy \,\mathrm{d}F_X(x) \right) \mathrm{d}F_Y(y) \\
&= \int_{-\infty}^{\infty} y \left(\int_{-\infty}^{\infty} x \,\mathrm{d}F_X(x) \right) \mathrm{d}F_Y(y) \\
&= \int_{-\infty}^{\infty} x \,\mathrm{d}F_X(x) \int_{-\infty}^{\infty} y \,\mathrm{d}F_Y(y) \\
&= \mathbf{E}[X]\mathbf{E}[Y],
\end{aligned}$$

which shows the claim. □

5.4 Infinite Product Spaces

Let $(\Omega_n, \mathcal{F}_n, \mathbf{P}_n)$, $n \in \mathbb{N}$, be probability spaces. In this section we consider the probability space associated with infinite random trials: $\Omega = \Omega_1 \times \Omega_2 \times \cdots = \times_{n=1}^{\infty} \Omega_n$. In Example 2.4.1 we saw that it is not always possible to have all the subsets of $\Omega = [0,1]$ measurable. This is in particular the case of infinite cartesian products $\Omega = \times_{n=1}^{\infty} \Omega_n$, where the slice space can be as simple as $\Omega_n = \{0,1\}$. In particular, we cannot set $\mathcal{F} = \mathcal{F}_1 \otimes \mathcal{F}_2 \otimes \cdots$. Instead, we must confine ourself to a smaller σ-algebra that is generated by the so-called cylinder sets.

5.4.1 Definition. The infinite product σ-algebra $\otimes_{n\in\mathbb{N}} \mathcal{F}_n$ on the product space $\times_{n=1}^{\infty} \Omega_n$ is generated by the cylinder sets. A set C is a *cylinder set*, if it is of the form

$$C = A_1 \times A_2 \times \cdots \times A_k \times \bigtimes_{j=k+1}^{\infty} \Omega_j \tag{5.4.2}$$

for some $A_i \in \mathcal{F}_i$, $i \le k$. Let $\mathcal{C}$ be the class of cylinder sets. Then

$$\bigotimes_{n=1}^{\infty} \mathcal{F}_n := \sigma(\mathcal{C}).$$

The infinite product measure $\otimes_{n=1}^{\infty} \mathbf{P}_n$ is defined via cylinder sets: Let C be given by (5.4.2). We set

$$\bigotimes_{n=1}^{\infty} \mathbf{P}_n[C] := \mathbf{P}_1[A_1] \cdots \mathbf{P}_k[A_k].$$

By the extension theorems Theorem 2.3.8 and Theorem 2.3.9 this probability measure extends uniquely to $\otimes_{n=1}^{\infty}\mathcal{F}_n$. Moreover, it is important to note that with respect to the product measure $\otimes_{n=1}^{\infty}\mathbf{P}_n$ the σ-algebras $\mathcal{F}_n$, $n \in \mathbb{N}$, are independent.

5.4.3 Definition. Let $(\Omega_n, \mathcal{F}_n, \mathbf{P}_n)$, $n \in \mathbb{N}$, be probability spaces. Their *product probability space* is the triplet $(\times_{n=1}^{\infty}\Omega_n, \otimes_{n=1}^{\infty}\mathcal{F}_n, \otimes_{n=1}^{\infty}\mathbf{P}_n)$ defined above.

5.4.4 Example. Let $\Omega_n = \{0,1\}$, $\mathcal{F}_n = \mathrm{pow}\,(\{0,1\})$ and $\mathbf{P}_n[\{0\}] = 1/2 = \mathbf{P}_n[\{1\}]$. Then the product space

$$\left(\bigtimes_{n=1}^{\infty}\Omega_n, \bigotimes_{n=1}^{\infty}\mathcal{F}_n, \bigotimes_{n=1}^{\infty}\mathbf{P}_n\right) \;=\; (\Omega_1^{\infty}, \mathcal{F}_1^{\otimes\infty}, \mathbf{P}_1^{\otimes\infty})$$

corresponds to the infinite coin tossing. In fact, if we denote

$$X_n \;:=\; \begin{cases} 1, & \text{if the } n^{\text{th}} \text{ toss is head,} \\ 0, & \text{if the } n^{\text{th}} \text{ toss is tail,} \end{cases}$$

then

$$\bigotimes_{n=1}^{\infty}\mathcal{F}_n \;=\; \sigma(X_1, X_2, \ldots).$$

Thus the product σ-algebra is the smallest σ-algebra such that the coordinate mappings $\omega \mapsto X_n(\omega) = \omega_n$, $n \in \mathbb{N}$, are measurable.

By interpreting Example 2.4.1 in a suitable way, we see that the product σ-algebra $\mathcal{F}_1^{\otimes\infty}$ strictly smaller than the full σ-algebra $\mathrm{pow}\,(\Omega_1^{\infty})$. In the same way we see that the probability measure $\mathbf{P}_1^{\otimes\infty}$ cannot be extended to the full σ-algebra $\mathrm{pow}\,(\Omega_1^{\infty})$. To show this rigorously, is left as Exercise 5.11.

5.5 Sums of Independent Random Variables

Let $X_1, \ldots, X_n$ be independent random variables. We consider the distribution $\mathbf{P}_{X_1+\ldots+X_n}$ of the sum $X_1 + \cdots + X_n$. Just for fun, we present our results in terms of the cumulative distribution functions:

$$\begin{aligned} F_{X_1+\cdots+X_n}(y) \;&:=\; \mathbf{P}_{X_1+\cdots+X_n}\big[(-\infty, y]\big] \\ &=\; \mathbf{P}[X_1 + \cdots + X_n \le y]. \end{aligned}$$

5.5.1 Definition. Let $F_1, \ldots, F_n$ be cumulative distribution functions. Their *convolution* $F_1 * \cdots * F_n$ is defined recursively as

$$\begin{aligned} F_1 * F_2(y) &:= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathbf{1}_{(-\infty,y]}(x_1+x_2)\,\mathrm{d}F_1(x_1)\mathrm{d}F_2(x_2), \\ F_1 * \cdots * F_n(y) &:= (F_1 * \cdots * F_{n-1}) * F_n(y) \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathbf{1}_{(-\infty,y]}(x_1+x_2)\,\mathrm{d}(F_1 * \cdots * F_{n-1})(x_1)\mathrm{d}F_n(x_2). \end{aligned}$$

In Definition 5.5.1 it should be noted that it if symmetric: $F_1 * \cdots * F_n = F_{\pi(1)} * \cdots * F_{\pi(n)}$, where $\pi : \{1, \ldots, n\} \to \{1, \ldots, n\}$ is any permutation of the set $\{1, \ldots, n\}$. We see this clearly from Theorem 5.5.3 below.

5.5.2 Example. Let

$$F_1(x) \;=\; F_2(x) \;=\; x\mathbf{1}_{[0,1]}(x) + \mathbf{1}_{(1,\infty)}(x).$$

Then

$$\mathrm{d}F(x) \;=\; \mathbf{1}_{[0,1]}(x)\mathrm{d}x$$

and after a short geometric interpretation we see that

$$\begin{aligned} F^{*2}(y) &:= F * F(y) \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathbf{1}_{(-\infty,y]}(x_1+x_2)\,\mathbf{1}_{[0,1]}(x_1)\mathrm{d}x_1\mathbf{1}_{[0,1]}(x_2)\mathrm{d}x_2 \\ &= \ell\Big[(x,y) \in \mathbb{R}^2\,;\, 0 \le x_1, x_2 \le 1, x_1 + x_2 \le y\Big] \\ &= \frac{y^2}{2}\mathbf{1}_{[0,1]}(y) + \left(1 - \frac{(2-y)^2}{2}\right)\mathbf{1}_{[1,2]}(y). \end{aligned}$$

The convolution of cumulative distribution functions is always a cumulative distribution function. This follows from Theorem 5.5.3 below.

5.5.3 Theorem. *Let $X_1, \ldots, X_n$ be indepedent random variables with cumulative distribution functions $F_1, \ldots, F_n$. Then the sum $X_1 + \cdots + X_n$ has cumulative distribution function $F_1 * \cdots * F_n$.*

Proof. Let us begin with the case $n = 2$. The general case follows then easily by using a standard induction argument and the recursive nature of convolution.

Let $Z = X_1 + X_2$. We want to calculate

$$F_Z(z) \;=\; \mathbf{P}[X_1 + X_2 \le z].$$

Denote

$$A := \{(x_1, x_2) \in \mathbb{R}^2 ; x_1 + x_2 \leq z\}.$$

Since X_1 and X_2 are independent, it follows from Theorem 5.1.3

$$F_Z(z) = \mathbf{P}_{X_1,X_2}(A) = \mathbf{P}_{X_1} \otimes \mathbf{P}_{X_2}(A).$$

Now the slice of A at point $x_1 \in \mathbb{R}$ is

$$A_{x_1} := \{x_2 \in \mathbb{R} ; x_2 \leq z - x_1\},$$

and

$$\mathbf{P}_{X_2}(A_{x_1}) = F_{X_2}(z - x_1).$$

Thus, it follows from the Fubini's theorem 5.3.7 that

$$\begin{aligned} F_Z(z) &= \int_{-\infty}^{\infty} \mathbf{P}_{X_2}(A_{x_1}) \, \mathrm{d}F_{X_1}(x_1) \\ &= \int_{-\infty}^{\infty} F_{X_2}(z - x_1) \, \mathrm{d}F_{X_1}(z). \end{aligned}$$

Recall the definition of a convolution, and the claim follows for $n = 2$.

Let then $n \in \mathbb{N}$ be general. Suppose the claim holds $k = 1, 2, \ldots, n$. By induction, we need to show that the claim holds for $n+1$. But for $n = 2$ we have the result for F_{X+Y}. By the induction hypothesis we have the result for $F_{X_1+\cdots+X_n}$. Therefore, taking $Y = X_{n+1}$, we have the result for $F_{X_1+\cdots+X_{n+1}}$, since

$$F_{X_1+\cdots+X_{n+1}}(z) = \int_{-\infty}^{\infty} F_{X_{n+1}}(z - x_n) \, \mathrm{d}F_{X_1+\cdots+X_n}(x_n).$$

□

In Example 5.5.2 we calculated the cumulative distribution function of a sum of two uniformy on $[0, 1]$ distributed random variables.

Definition 5.5.1 deals with cumulative distribution functions, i.e. measures. For (density) functions $f_1, \ldots, f_n : \mathbb{R} \to \mathbb{R}$ their convolution is defined by setting

$$\begin{aligned} f_1 * f_2(y) &:= \int_{-\infty}^{\infty} f_1(y - x_2) f_2(x_2) \, \mathrm{d}x_2, \\ f_1 * \cdots * f_n(y) &:= (f_1 * \cdots f_{n-1}) * f_n(y) \\ &= \int_{-\infty}^{\infty} f_1 * \cdots * f_{n-1}(y - x_n) f_n(x_n) \, \mathrm{d}x_n. \end{aligned}$$

In the same way for sequences $p_{1,\cdot}, \ldots, p_{n,\cdot}$ their convolution is

$$\begin{aligned}(p_{1,\cdot} * p_{2,\cdot})_k &:= \sum_l p_{1,k-l} p_{2,l},\\ (p_{1,\cdot} * \cdots * p_{n,\cdot})_k &:= (p_{1,\cdot} * \cdots * p_{n-1,\cdot}) * p_{n,\cdot}\\ &= \sum_l (p_{1,\cdot} * \cdots * p_{n-1,\cdot})_{k-l} p_{n,l}.\end{aligned}$$

The next Corollary 5.5.4 of Theorem 5.5.3 sheds light to the connection of different convolutions.

5.5.4 Corollary. *Let $X_1, \ldots, X_n$ be independent random variables.*

(i) *If $X_1, \ldots, X_n$ are discrete and $p_{l,k} := \mathbf{P}[X_l = k]$, then the sum $X_1 + \cdots + X_n$ is discrete and $\mathbf{P}[X_1 + \cdots + X_n = k] = p_{1,\cdot} * \cdots * p_{n,\cdot}$.*

(ii) *If $X_1, \ldots, X_n$ are continuous with densities $f_1, \ldots, f_n$, then the sum $X_1 + \cdots + X_n$ is continuous and has the density $f_1 * \cdots * f_n$.*

Proof. Exercise 5.12 □

5.5.5 Example. Let X_1 and X_2 be independent Poisson distributed random variables with parameters λ_1 and λ_2. In other words X_1 and X_2 have point probability functions

$$\begin{aligned}p_{1,k} &:= \mathbf{P}[X_1 = k] = \mathrm{e}^{-\lambda_1} \frac{\lambda_1^k}{k!},\\ p_{1,k} &:= \mathbf{P}[X_2 = k] = \mathrm{e}^{-\lambda_2} \frac{\lambda_2^k}{k!},\end{aligned}$$

ehere $k = 0, 1, \ldots$. Then, by using Newton's binomial formula, we obtain

$$\begin{aligned}\mathbf{P}[X_1 + X_2 = k] &= \sum_l p_{1,k-l} p_{2,l}\\ &= \sum_l e^{-\lambda_1} \frac{\lambda_l^{k-l}}{(k-l)!} e^{-\lambda_2} \frac{\lambda_2^l}{l!}\\ &= e^{-(\lambda_1+\lambda_2)} \sum_l \frac{1}{(k-l)!l!} \lambda_1^{k-l} \lambda_2^l\\ &= e^{-(\lambda_1+\lambda_2)} \frac{1}{k!} \sum_l \binom{n}{k} \lambda_1^{k-1} \lambda_2^k\\ &= e^{-(\lambda_1+\lambda_2)} \frac{(\lambda_1+\lambda_2)^k}{k!}.\end{aligned}$$

We see that $X_1 + X_2$ is Poisson distributed with parameter $\lambda_1 + \lambda_2$.

More generally, it can be shown by using induction that if $X_1, \ldots, X_n$ are independent Poisson distributed random variables with parameters $\lambda_1, \ldots, \lambda_n$, then the sum $X_1 + \cdots + X_n$ is Poisson distributed with parameter $\lambda_1 + \cdots + \lambda_n$.

Exercises

5.1. Give an example of events that are causally dependent but stochastically independent.

5.2. Give an example of collection $\mathcal{G}_j$, $j \in J$, of σ-algebras that are pairwise independent but not independent.

5.3. Show that a random varable X is independent of itself if and only if it is a.s. a constant.

5.4. Prove the implication (v) $\Rightarrow$ (ii) of Theorem 5.1.3.

5.5. Let $\mathcal{F}_1$ and $\mathcal{F}_2$ be independent σ-algebras. Let $\mathcal{F}_1' \subset \mathcal{F}_1$ be a sub-σ-algebra and let $\mathcal{F}_2' \subset \mathcal{F}_2$ be a sub-σ-algebra. Show that the σ-algebras $\mathcal{F}_1'$ and $\mathcal{F}_2'$ are independent.

5.6. Prove the implication (vi) $\Rightarrow$ (ii) of Theorem 5.1.3.

5.7. Argue the result of Example 5.2.2(ii) by using Kolmogorov's 0–1 law 5.2.4.

5.8. Show that in the proof of Lemma 5.3.3 the class

$$\mathcal{H} \;:=\; \{A \in \mathcal{F}_1 \otimes \mathcal{F}_2 \,;\, \mathbf{1}_A(\omega_1, \cdot) \text{ is } \mathcal{F}_2\text{-measurable for all } \omega_1 \in \Omega_1\}$$

is a σ-algebra.

5.9. Prove Fubini's theorem 5.3.7.

5.10. Let $X, Y, XY \in L^1$. Show that from the condition

$$\mathbf{E}[XY] \;=\; \mathbf{E}[X]\mathbf{E}[Y]$$

it does not follow that X and Y are independent.

5.11. LEtus consider Example 5.4.4.

(a) Show that the product-σ-algebra is a proper subset of $\mathrm{pow}\,(\{0,1\}^\infty)$.

(b) Show that the product measure $\mathbf{P}_1^{\otimes\infty}$ cannot be extended to the full σ-algebra $\mathrm{pow}(\{0,1\}^\infty)$.

5.12. Prove Corollary 5.5.4.

5.13. Let X and Y be independent $\mathbb{N}$-valued random variables and let

$$\mathbf{P}[X = n] \;=\; \mathbf{P}[Y = n] \;=\; \frac{1}{2^n}.$$

Calculate the following probabilities

(a) $\mathbf{P}[\min(X, Y) \leq n]$,
(b) $\mathbf{P}[X = Y]$,
(c) $\mathbf{P}[X > Y]$,
(d) $\mathbf{P}[X \text{ divides } Y]$.

5.14. Let $X_1, \ldots, X_n$ be independent random variables. We assume they have the same expectations and variances and denote them by $\mu := \mathbf{E}[X_k]$ and $\sigma^2 := \mathbf{Var}[X_k]$, $k \leq n$. Let

$$\bar{X} \;:=\; \frac{1}{n}\sum_{k=1}^{n} X_k \quad \text{and} \quad S^2 \;:=\; \frac{1}{n}\sum_{k=1}^{n} \left(X_k - \bar{X}\right)^2.$$

Show that

(a) $\mathbf{E}[\bar{X}] = \mu$,
(b) $\mathbf{Var}[\bar{X}] = \sigma^2/n$,
(c) $\mathbf{E}[S^2] = \sigma^2\,(n-1)/n$.

How is this related to statistical estimation?

5.15. Let X_1 and X_2 be independent normally distributed random variables with means μ_1 and μ_2 and variances σ_1 and σ_2. In other words the density functions of X_i, $i = 1, 2$, are

$$\varphi_i(x) \;=\; \frac{1}{\sqrt{2\pi}\sigma_i} \mathrm{e}^{-\frac{1}{2}\left(\frac{x-\mu_i}{\sigma_i}\right)^2}.$$

Calculate the density function of the sum $X_1 + X_2$.

5.16. Olkoot X_1 and X_2 be independent random variables with cumulative distribution functions F_{X_1} and F_{X_2}. Find the cumulative distribution functions of the following transformations:

(a) $\max(X_1, X_2)$,
(b) $\min(X_1, X_2)$,
(c) $X_1 X_2$,
(d) X_1/X_2.

5.17. Calculate the density functions of Exercise 5.16 when we assume that the random variables X_1 and X_2 are continuous with density functions f_{X_1} and f_{X_2}.

Chapter 6

Characteristic Functions

Characteristic functions are technical tools for analyzing distributions. They are particlulary powerful in analysizing sums of independent random variables. In this course we use the characteristic function mainly in proving the central limit theorem. Finally, we note that the characteristic function maybe known to you from analysis courses: There it is known as Fourier transform, or rather Fourier–Stieltjes transform.

6.1 Complex Numbers and Functions

We present here only a very short and gentle introduction to complex numbers. Basically, we present only what is needed to define the complex exponential funciton.

As a vector space (i.e. with respect to summation) and as topological space (i.e. witrh respect to norm) the complex plane $\mathbb{C}$ corresponds the real plane $\mathbb{R}^2$. As an algebra (i.e, we also consider multiplication) $\mathbb{C}$ is something completely different from $\mathbb{R}^2$. In fact, there is no natural multiplication in $\mathbb{R}$.

A *complex number* $z = (x, y) \in \mathbb{C}$ is of the form $z = x + \mathrm{i}y$, where $\mathrm{i} := \sqrt{-1}$; $x \in \mathbb{R}$ is the *real part* $x = \mathrm{Re}(z)$ of the complex number $z \in \mathbb{C}$ and $y \in \mathbb{R}$ is its *imaginary part* $y = \mathrm{Im}(z)$.

The *complex conjugate* $\bar{z}$ of a complex number $z = x + \mathrm{i}y$ is $\bar{z} := z - \mathrm{i}y$.

The *modulus* of a complex number is its length or norm

$$|z| \;:=\; \sqrt{x^2 + y^2}.$$

The *argument* of a complex number is the angle that the number makes with the positive real axis in the complex plane $\arg z \in (-\pi, \pi]$ if given by the conditions

$$\cos(\arg z) \;=\; \frac{x}{\sqrt{x^2 + y^2}} \qquad \text{and} \qquad \sin(\arg z) \;=\; \frac{y}{\sqrt{x^2 + y^2}}.$$

Then the pair $(x, y) \in \mathbb{C}$ has the *polar representation* $(|z|, \arg z) \in \mathbb{R}_+ \times (-\pi, \pi]$, i.e.

$$x = |z| \cos(\arg z) \qquad \text{and} \qquad y = |z| \sin(\arg z).$$

Later we see that we can always write

$$z = |z| \mathrm{e}^{\mathrm{i}\theta},$$

where $|z|$ is the modulus and θ is the argument.

The representation (x, y) is called the *Cartesian representation.*

The sum of complex numbers $z_1 = (x_1, y_1)$ and $z_2 = (x_2, y_2)$ is defined compoment-wise

$$z_1 + z_2 := (x_1 + x_2) + \mathrm{i}(y_1 + y_2).$$

The product of colplex numbers $z_1 = (x_1, y_1)$ is defined component-wise with the rule $\mathrm{i}^2 = -1$:

$$\begin{aligned} z_1 z_2 &= (x_1 + \mathrm{i}y_1)(x_2 + \mathrm{i}y_2) \\ &= x_1 x_2 + \mathrm{i}x_1 y_2 + \mathrm{i}y_1 x_2 + \mathrm{i}^2 y_1 y_2 \\ &= (x_1 x_2 - y_1 y_2) + \mathrm{i}(x_1 y_2 + x_2 y_1). \end{aligned}$$

The powers z^n, $n \in \mathbb{N} \cup \{0\}$, of a complex number z defined recursively as

$$\begin{aligned} z^0 &:= 1, \\ z^n &:= z^{n-1} z. \end{aligned}$$

We note that for all $n, m \in \mathbb{N} \cup \{0\}$ we have $z^n z^m = z^{n+m}$.

If $z \neq 0$, its *inverse*

$$\frac{1}{z} := z^{-1} := \frac{x}{x^2 + y^2} - i\frac{y}{x^2 + y^2}.$$

Then we have $z\frac{1}{z} = 1$ (Exercise 6.1). The division is defined by setting

$$\frac{z_1}{z_2} := z_1 \frac{1}{z_2},$$

when $z_2 \neq 0$. We set $z^{-n} := 1/z^n$. Then the formula $z^n z^m = z^{n+m}$ holds for all $n, m \in \mathbb{Z}$, kun $z \neq 0$.

Recall that the real complex function e^x, $x \in \mathbb{R}$, can be represented as the series $\mathrm{e}^x = \sum_{n=0}^{\infty} x^n/n!$. By using function theory, it can be shown that the series converges absolutely, when the real number $x \in \mathbb{R}$ is replaced with

a complex number $z \in \mathbb{C}$. This leads into the followng definition of the *complex exponential function*

$$\mathrm{e}^z := \sum_{n=0}^{\infty} \frac{z^n}{n!}. \tag{6.1.1}$$

By using the *Cauchy's product formula* and *Newton's binomial formula* we obtain

$$\begin{aligned}
\mathrm{e}^{z_1}\mathrm{e}^{z_2} &= \sum_{n=0}^{\infty} \frac{z_1^n}{n!} \sum_{m=0}^{\infty} \frac{z_2^m}{m!} \\
&= \sum_{n=0}^{\infty} \sum_{m=0}^{n} \frac{z_1^{n-m} z_2^m}{(n-m)!m!} \\
&= \sum_{n=0}^{\infty} \frac{1}{n!} \sum_{m=0}^{n} \binom{n}{m} z_1^{n-m} z_2^m \\
&= \sum_{n=0}^{\infty} \frac{(z_1+z_2)^n}{n!}.
\end{aligned}$$

We see that

$$\mathrm{e}^{z_1+z_2} = \mathrm{e}^{z_1}\mathrm{e}^{z_2}. \tag{6.1.2}$$

From (6.1.2) it follows immediately, that $\mathrm{e}^z = e^{x+\mathrm{i}y} = \mathrm{e}^x\mathrm{e}^{\mathrm{i}y}$. But

$$\begin{aligned}
\mathrm{e}^{\mathrm{i}y} &= \sum_{n=0}^{\infty} \frac{(\mathrm{i}y)^n}{n!} \\
&= \sum_{n=0}^{\infty} (-1)^n \frac{y^{2n}}{(2n)!} + \mathrm{i} \sum_{n=0}^{\infty} (-1)^n \frac{y^{2n+1}}{(2n+1)!}.
\end{aligned}$$

We notice here the series expansions of sine and cosine

$$\begin{aligned}
\cos y &= \sum_{n=0}^{\infty} (-1)^n \frac{y^{2n}}{(2n)!}, \\
\sin y &= \sum_{n=0}^{\infty} (-1)^n \frac{y^{2n+1}}{(2n+1)!}.
\end{aligned}$$

Thus, by using the formula (6.1.2) we obtain the *Euler's formula*

$$\mathrm{e}^z := \mathrm{e}^x\big(\cos x + \mathrm{i}\sin y\big).$$

From the Euler's formula it follows that the complex number $z = x + \mathrm{i}y$ has polar representation $z = re^{\mathrm{i}\theta}$, where $r = |z|$ and $\theta = \arg z$. Moreover, we see that $|\mathrm{e}^z| = \mathrm{e}^x$, $\mathrm{e}^{z+2n\pi\mathrm{i}} = \mathrm{e}^z$ for all $n \in \mathbb{Z}$ and that the mapping $z \mapsto \mathrm{e}^z$ is continuous with respect to the modulus. We also see that the reange of the mapping $z \mapsto e^z$ is $\mathbb{C} \setminus \{(0,0)\}$. This means, that unlike its real counterpart, the complex exponential function is not positive. For example we have the most beautiful formula of mathematics $\mathrm{e}^{\mathrm{i}\pi} + 1 = 0$.

The following formulas fo the imaginary complex function $x \mapsto e^{ix}$, $x \in \mathbb{R}$, will be useful when we later consider the characteristic functions.

6.1.3 Lemma. *Let $x, y, u \in \mathbb{R}$. Then*

(i) $|\mathrm{e}^{\mathrm{i}x} - 1| \leq |x|$,
(ii) $\frac{\partial}{\partial u}\mathrm{e}^{\mathrm{i}ux} = \mathrm{i}x\mathrm{e}^{\mathrm{i}ux}$.

Proof. (i) We show that $|\mathrm{e}^{\mathrm{i}x} - 1|^2 \leq x^2$. Now

$$\begin{aligned} \left|\mathrm{e}^{\mathrm{i}x} - 1\right|^2 &= \left|(\cos x - 1) + i \sin x\right|^2 \\ &= (\cos x - 1)^2 + \sin^2 x \\ &= \cos^2 x - 2\cos x + 1 + \sin^2 x \\ &= 2 - 2\cos x. \end{aligned}$$

So, we need to show that

$$1 - \cos x \leq \frac{x^2}{2}.$$

But this can be seen by differentiating both sides and by using the inequality $\sin x \leq x$, which in term can also be seen by differentiating both sides and using the fact that $\cos x \leq 1$.

(ii) This is direct calculation:

$$\begin{aligned} \frac{\partial}{\partial u}\mathrm{e}^{\mathrm{i}ux} &= \frac{\partial}{\partial u}\big(\cos(ux) + \mathrm{i}\sin(ux)\big) \\ &= \frac{\partial}{\partial u}\cos(ux) + \mathrm{i}\frac{\partial}{\partial u}\sin(ux) \\ &= -x\sin(ux) + \mathrm{i}x\cos(ux) \\ &= \mathrm{i}x\,\mathrm{i}\sin(ux) + \mathrm{i}x\cos(ux) \\ &= \mathrm{i}x\big(\cos(ux) + \mathrm{i}\sin(ux)\big) \\ &= \mathrm{i}xe^{\mathrm{i}ux}. \end{aligned}$$

□

6.2 Complex-Valued Random Variables

A complex-valued random variable is $Z = X + \mathrm{i}Y$, where X and Y are real-valued random variables.

A collection $Z_k = X_k + \mathrm{i}Y_k$, $k = 1, 2, \ldots, n$, of complex-valued random variables is independent if all the collections U_k, $k = 1, 2, \ldots, n$, are independent, where U_k is either X_k or Y_k.

The complex-valued random variable $Z = X + \mathrm{i}Y$ is integrable if both its real and imaginary parts are integrable, and hten

$$\mathbf{E}[Z] \;:=\; \mathbf{E}[X] + \mathrm{i}\mathbf{E}[Y].$$

6.2.1 Lemma. *Let Z_1 and Z_2 be integrable complex-valued random variables and let $z \in \mathbb{C}$. Then*

(i) $\mathbf{E}[zZ_1] = z\mathbf{E}[Z_1]$,
(ii) $\mathbf{E}[Z_1 + Z_2] = \mathbf{E}[Z_1] + \mathbf{E}[Z_2]$,
(iii) *if Z_1 and Z_2 are independent, then Z_1Z_2 is integrable and* $\mathbf{E}[Z_1Z_2] = \mathbf{E}[Z_1]\mathbf{E}[Z_2]$,
(iv) $|\mathbf{E}[Z_1]| \leq \mathbf{E}[|Z_1|]$.

Proof. Lauseen 4.1.15 valossa

(i) The claim follows by splitting Z_1 and z into real and imaginary parts and by using the corresponding real-valued result. Indeed,

$$\begin{aligned}
\mathbf{E}[zZ_1] &= \mathbf{E}\big[(x + \mathrm{i}y)(X_1 + iY_1)\big] \\
&= \mathbf{E}\big[xX_1 + \mathrm{i}xY_1 + \mathrm{i}yX_1 - yY_1\big] \\
&= \mathbf{E}\big[(xX_1 - yY_1) + \mathrm{i}(xY_1 + yX_1)\big] \\
&= \mathbf{E}\big[xX_1 - yY_1\big] + \mathrm{i}\mathbf{E}\big[xY_1 + yX_1\big] \\
&= x\mathbf{E}[X_1] - y\mathbf{E}[Y_1] + \mathrm{i}x\mathbf{E}[Y_1] + \mathrm{i}y\mathbf{E}[X_1] \\
&= x\mathbf{E}[X_1] + \mathrm{i}x\mathbf{E}[Y_1] + \mathrm{i}y\mathbf{E}[X_1] - y\mathbf{E}[Y_1] \\
&= (x + \mathrm{i}y)\,(\mathbf{E}[X_1] + \mathrm{i}\mathbf{E}[Y_1]) \\
&= z\mathbf{E}[Z_1].
\end{aligned}$$

(ii) This claim follows precisely the same way as part (i). Indeed,

$$\begin{aligned}
\mathbf{E}[Z_1 + Z_2] &= \mathbf{E}[(X_1 + \mathrm{i}Y_1) + (X_2 + \mathrm{i}Y_2)] \\
&= \mathbf{E}[(X_1 + X_2) + \mathrm{i}(Y_1 + Y_2)] \\
&= \mathbf{E}[X_1 + X_2] + \mathrm{i}\mathbf{E}[Y_1 + Y_2] \\
&= \big(\mathbf{E}[X_1] + \mathbf{E}[X_2]\big) + \mathrm{i}\big(\mathbf{E}[Y_1] + \mathbf{E}[Y_2]\big) \\
&= \big(\mathbf{E}[X_1] + \mathrm{i}\mathbf{E}[Y_1]\big) + \big(\mathbf{E}[X_2] + \mathrm{i}\mathbf{E}[Y_2]\big) \\
&= \mathbf{E}[Z_1] + \mathbf{E}[Z_2]
\end{aligned}$$

(iii) This part is left as Exercise 6.2.

(iv) This proof is a simple exercise of complex numbers. We can write $\mathbf{E}[Z] = |\mathbf{E}[Z]|e^{i\theta}$ for some θ. Now, multiplying Z with $e^{-i\theta}$ does not change the modulus. Hence

$$|\mathbf{E}[Z]| \;=\; \mathrm{Re}(e^{-i\theta}\mathbf{E}[Z]) \;=\; \mathbf{E}[\mathrm{Re}(e^{-i\theta}Z)].$$

But for all complex numbers we have $\mathrm{Re}(e^{-i\theta}z) \le |e^{-i\theta}z| = |z|$. Hence,

$$|\mathbf{E}[Z]| \;\le\; \mathbf{E}\left[\mathrm{Re}\left(e^{-i\theta}Z\right)\right] \;\le\; \mathbf{E}[|Z|]$$

and the claim follows. □

6.3 Characteristic Function and its Properties

Characteristic functions are useful in this course since (as we shall see later):

1. the characteristic function determines (or characterizes) the distribution uniquely,
2. the characteristic function of a sum of independent random variables is the product of the characteristic functions of the summands.

6.3.1 Definition. The *characteristic function* $\phi_X\colon \mathbb{R} \to \mathbb{C}$ of a random variable X is

$$\begin{aligned}\phi_X(u) &:= \mathbf{E}\left[e^{iuX}\right] \\ &= \mathbf{E}[\cos(uX)] + i\,\mathbf{E}[\sin(uX)]\,.\end{aligned}$$

So, the characteristic function depends on hte ranwom variable X only through its distribution $\mathbf{P}_X$: if the random variables X and Y have the same distribution, or $\mathbf{P}_X = \mathbf{P}_Y$ or equally $F_X = F_Y$, then $\phi_X = \phi_Y$. Indeed, we can write

$$\phi_X(u) \;=\; \int_{-\infty}^{\infty} e^{iux}\,\mathrm{d}F_X(x).$$

In particular, if X is continuous with density function f_X, then

$$\phi_X(u) \;=\; \int_{-\infty}^{\infty} e^{iux} f_X(x)\,\mathrm{d}x.$$

Those who know their Fourier analysis recognize the (non-normalize inverse) Fourier transform here.

Later, in Corollary 6.3.9), we see that the folloiwng inverse is true: if $\phi_X = \phi_Y$, then $\mathbf{P}_X = \mathbf{P}_Y$.

6.3.2 Example. (i) Let X be *Bernoulli distributed* with parameter p, i.e. $\mathbf{P}[X = 1] = p$ and $\mathbf{P}[X = 0] = (1 - p)$. Then

$$\begin{aligned} \phi_X(u) &= e^{iu0}(1-p) + e^{iu}p \\ &= pe^{iu} + 1 - p. \end{aligned}$$

(ii) Let X be Poisson distributed with parameter λ. Then

$$\begin{aligned} \phi_X(u) &= \mathrm{e}^{-\lambda} \sum_{k=0}^{\infty} \mathrm{e}^{\mathrm{i}uk} \frac{\lambda^k}{k!} \\ &= \mathrm{e}^{-\lambda} \sum_{k=0}^{\infty} \frac{\left(\lambda \mathrm{e}^{\mathrm{i}u}\right)^k}{k!} \\ &= \mathrm{e}^{-\lambda} \mathrm{e}^{\lambda e^{\mathrm{i}u}} \\ &= \mathrm{e}^{\lambda(\mathrm{e}^{\mathrm{i}u} - 1)}. \end{aligned}$$

(iii) Let X be normally distributed with parameters $\mu = 0$ and $\sigma^2 = 1$, i.e.

$$\begin{aligned} \mathbf{P}[X \in B] &= \int_B \varphi(x)\,\mathrm{d}x \\ &= \frac{1}{\sqrt{2\pi}} \int_B \mathrm{e}^{-\frac{1}{2}x^2}\,\mathrm{d}x. \end{aligned}$$

Then, by using the series expansion of the exponential function and the following formula, which can be seen by integrating-by-parts,

$$\begin{aligned} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^{2n} \mathrm{e}^{-\frac{1}{2}x^2}\,\mathrm{d}x &= (2n-1)!! \\ &:= (2n-1)(2n-3)\cdots 3 \cdot 1, \end{aligned}$$

we obtain

$$\begin{aligned}
\phi_X(u) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \mathrm{e}^{\mathrm{i}ux}\, \mathrm{e}^{-\frac{1}{2}x^2}\, \mathrm{d}x \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \sum_{n=0}^{\infty} \frac{(\mathrm{i}ux)^n}{n!} \mathrm{e}^{-\frac{1}{2}x^2}\, \mathrm{d}x \\
&= \sum_{n=0}^{\infty} \frac{(\mathrm{i}u)^n}{n!} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^n \mathrm{e}^{-\frac{1}{2}x^2}\, \mathrm{d}x \\
&= \sum_{n=0}^{\infty} \frac{(\mathrm{i}u)^{2n}}{(2n)!} (2n-1)!! \\
&= \sum_{n=0}^{\infty} \frac{(\mathrm{i}u)^{2n}}{(2n)!} \frac{(2n)!}{2^n n!} \\
&= \sum_{n=0}^{\infty} \left(-\frac{1}{2}u^2\right)^n \frac{1}{n!} \\
&= \mathrm{e}^{-\frac{1}{2}u^2}.
\end{aligned}$$

(iv) Let X be uniformly distributed on the interval $[0,1]$, i.e.

$$\mathbf{P}_X[B] = \int_B \mathbf{1}_{[0,1]}(x)\, \mathrm{d}x.$$

Then

$$\begin{aligned}
\phi_X(u) &= \int_0^1 \mathrm{e}^{\mathrm{i}ux}\, \mathrm{d}x \\
&= \left[\frac{\mathrm{e}^{\mathrm{i}ux}}{\mathrm{i}u}\right]_{x=0}^{1} \\
&= \frac{\mathrm{e}^{\mathrm{i}u} - 1}{\mathrm{i}u},
\end{aligned}$$

if $u \neq 0$ and $\phi_X(0) = 1$.

6.3.3 Lemma. *Let X be a random variable, and let $a, b \in \mathbb{R}$. Then*

$$\phi_{aX+b}(u) = \mathrm{e}^{\mathrm{i}ub} \phi_X(au).$$

Proof. This is a direct calculation:

$$\begin{aligned}
\phi_{aX+b}(u) &= \mathbf{E}\left[\mathrm{e}^{\mathrm{i}u(aX+b)}\right] \\
&= \mathbf{E}\left[\mathrm{e}^{\mathrm{i}uaX} \mathrm{e}^{\mathrm{i}ub}\right] \\
&= \mathrm{e}^{\mathrm{i}ub} \mathbf{E}\left[\mathrm{e}^{\mathrm{i}auX}\right] \\
&= \mathrm{e}^{\mathrm{i}ub} \phi_X(au).
\end{aligned}$$

□

6.3.4 Example. Let X be normally distributed with parameters μ ja σ^2. Then $Z := (X - \mu)/\sigma$ is normally distributed with parameter $\mu = 1$ and $\sigma^2 = 1$. By Example 6.3.2(iii)

$$\phi_Z(u) \;=\; e^{-\frac{1}{2}u^2}.$$

Therefore, by Lemma 6.3.3,

$$\begin{aligned}\phi_X(u) &= \phi_{\sigma Z+\mu}(u)\\ &= \mathrm{e}^{\mathrm{i}u\mu}\phi_Z(\sigma u)\\ &= \mathrm{e}^{\mathrm{i}u\mu-\frac{1}{2}\sigma^2u^2}.\end{aligned}$$

6.3.5 Lemma. *The characteristic function* $\phi_X\colon \mathbb{R} \to \mathbb{C}$ *is*

(i) *bounded,*
(ii) *uniformly continuous,*
(iii) *and* $\phi_X(0) = 1$.

Proof. (i) We note first that $|\mathrm{e}^{\mathrm{i}uX}| \le 1$. Thus, by the inequality 6.2.1(iv), we have $|\phi_X(u)| \le \mathbf{E}\big[\big|\mathrm{e}^{\mathrm{i}uX}\big|\big] \le 1$. This means that ϕ_X is bounded.

(ii) Our tools are the estimate $|\mathrm{e}^{\mathrm{i}vX}| \le 1$, the inequality of Lemma 6.2.1(iv) and the estimate of Lemma 6.1.3(i). Let $u, v \in \mathbb{R}$. Now

$$\begin{aligned}|\phi_X(u) - \phi_X(v)| &= \big|\mathbf{E}\big[\mathrm{e}^{\mathrm{i}uX}\big] - \mathbf{E}\big[\mathrm{e}^{\mathrm{i}vX}\big]\big|\\ &= \big|\mathbf{E}\big[\mathrm{e}^{\mathrm{i}uX} - \mathrm{e}^{\mathrm{i}vX}\big]\big|\\ &\le \mathbf{E}\big[\big|\mathrm{e}^{\mathrm{i}uX} - \mathrm{e}^{\mathrm{i}vX}\big|\big]\\ &= \mathbf{E}\big[\big|\mathrm{e}^{\mathrm{i}vX}\left(\mathrm{e}^{\mathrm{i}(u-v)X} - 1\right)\big|\big]\\ &\le \mathbf{E}\big[\big|\mathrm{e}^{\mathrm{i}vX}\big|\,\big|\mathrm{e}^{\mathrm{i}(u-v)X} - 1\big|\big]\\ &\le \mathbf{E}\big[\big|\mathrm{e}^{\mathrm{i}(u-v)X} - 1\big|\big].\end{aligned}$$

Since

$$\big|\mathrm{e}^{\mathrm{i}(u-v)X} - 1\big| \;\le\; 2,$$

we can use the Lebesgue's dominated convergence theorem. Since

$$\lim_{h\to 0}\big|\mathrm{e}^{\mathrm{i}hX} - 1\big| \;=\; 0,$$

we have that

$$\lim_{h\to 0}\sup_{u\in\mathbb{R}}|\phi_X(u+h) - \phi_X(u)| \;=\; 0,$$

which means that ϕ_X is uniformly continuous.

(iii) This claim is trivial. □

Theorem 6.3.6 below is instrumental in proving the central limit theorem later.

6.3.6 Theorem. *Let X be a randon variable such that $\mathbf{E}[|X|^n] < \infty$ for some $n \in \mathbb{N}$. Then*

(i) *the characteristic function ϕ_X has the kth order derivative for all $k \leq n$, and*

$$\frac{\mathrm{d}^k}{\mathrm{d}u^k}\phi_X(u) \;=\; \mathrm{i}^k\mathbf{E}\big[X^k e^{\mathrm{i}uX}\big]\,,$$

(ii) *ϕ_X admits the nth order Taylor expansion in 0:*

$$\phi_X(u) \;=\; 1+\sum_{k=1}^{n}\frac{1}{k!}\mathbf{E}[X^k](\mathrm{i}u)^k + u^n\varepsilon(u),$$

where $\lim_{u\to 0}\varepsilon(u) = 0$.

Proof. (i) We want to change the order of integration and differentiation in

$$\frac{\mathrm{d}^k}{\mathrm{d}u^k}\mathbf{E}\big[\mathrm{e}^{\mathrm{i}uX}\big]\,.$$

We only consider tha case $k = 1$. The general case follows with similar arguments by using induction.

The shange of order will follow from Lebesgue's dominated convergence theorem 4.2.1(v). To use that, we use the estimate 6.1.3(i):

$$\big|\mathrm{e}^{\mathrm{i}x} - 1\big| \;\leq\; x.$$

From this, we see immediately that

$$\left|\frac{\mathrm{e}^{\mathrm{i}(u+h)X} - \mathrm{e}^{\mathrm{i}uX}}{h}\right| \;\leq\; |X|.$$

Let $h \to 0$ along some sequence. We obtain from the Lebesgue's dominated convergence theorem 4.2.1(v), that

$$\begin{aligned}
\phi'_X(u) &= \lim_{h\to 0}\frac{\phi_X(u+h)-\phi_X(u)}{h}\\
&= \lim_{h\to 0}\frac{\mathbf{E}\big[\mathrm{e}^{\mathrm{i}(u+h)X}\big] - \mathbf{E}\big[\mathrm{e}^{\mathrm{i}uX}\big]}{h}\\
&= \lim_{h\to 0}\mathbf{E}\left[\frac{\mathrm{e}^{\mathrm{i}(u+h)X} - \mathrm{e}^{\mathrm{i}uX}}{h}\right]\\
&= \mathbf{E}\left[\lim_{h\to 0}\frac{\mathrm{e}^{\mathrm{i}(u+h)X} - e^{\mathrm{i}uX}}{h}\right]\\
&= \mathbf{E}\big[\mathrm{i}uX\mathrm{e}^{\mathrm{i}uX}\big]\,.
\end{aligned}$$

Thus, changing the order of differentiation and expectation is allowed. We obtain

$$\begin{aligned}\frac{\mathrm{d}^k}{\mathrm{d}u^k}\phi_X(u) &= \frac{\mathrm{d}^k}{\mathrm{d}u^k}\mathbf{E}\left[\mathrm{e}^{\mathrm{i}u\cdot X}\right] \\ &= \mathbf{E}\left[\frac{\mathrm{d}^k}{\mathrm{d}u^k}\mathrm{e}^{(\mathrm{i}X)u}\right] \\ &= \mathrm{i}^k\mathbf{E}\left[X^k\mathrm{e}^{\mathrm{i}uX}\right].\end{aligned}$$

Part (i) follows.

(ii) This part follows from part (i) by applying the nth order Taylor's expansion to the function ϕ_X at the origin. □

The following Lévy's inversion formula tells us how the distribution of X can be recovered from its characteristic function ϕ_X.

6.3.7 Theorem. *Let ϕ_X and F_X be the characteristic function and the cumulative distribution function of a random variable X. Let $a, b \in \mathbb{R}$ be the continuity points of F_X with $a < b$. Then the following* Lévy's inversion formula *holds*

$$F_X(b) - F_X(a) = \frac{1}{2\pi}\lim_{T\to\infty}\int_{-T}^{T}\frac{\mathrm{e}^{-\mathrm{i}au} - \mathrm{e}^{-\mathrm{i}bu}}{\mathrm{i}u}\,\phi_X(u)\,\mathrm{d}u.$$

Before the formal proof of the inversion formula of Theorem 6.3.7 we show heuristically why it is true.

Let $\hat{\mu}$ be the characteristic function of a measure μ:

$$\hat{\mu}(u) := \int_{-\infty}^{\infty}\mathrm{e}^{\mathrm{i}ux}\,\mu[\mathrm{d}x].$$

We set a bilinear relation between a characteristic function and its measure as

$$\langle\hat{\nu},\mu\rangle := \int_{-\infty}^{\infty}\hat{\nu}(x)\,\mu[\mathrm{d}x].$$

We note that the relation $\langle\cdot,\cdot\rangle$ satisfies the integration-by-parts

$$\langle\hat{\nu},\mu\rangle = \langle\hat{\mu},\nu\rangle.$$

Indeed, by the Fubini's theorem 5.3.7

$$\begin{aligned}
\langle \hat{\nu}, \mu \rangle &= \int_{-\infty}^{\infty} \hat{\nu}(x)\, \mu[\mathrm{d}x] \\
&= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} \mathrm{e}^{\mathrm{i}ux}\, \nu[\mathrm{d}u] \right) \mu[\mathrm{d}x] \\
&= \iint_{\mathbb{R}^2} \mathrm{e}^{\mathrm{i}ux}\, \mu[\mathrm{d}x]\nu[\mathrm{d}u] \\
&= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} \mathrm{e}^{\mathrm{i}ux}\, \mu[\mathrm{d}x] \right) \nu[\mathrm{d}u] \\
&= \int_{-\infty}^{\infty} \hat{\mu}(u)\, \nu[\mathrm{d}u] \\
&= \langle \hat{\mu}, \nu \rangle .
\end{aligned}$$

Let then $B \in \mathcal{B}$ and let $\check{\mathbf{1}}_B$ be the probability measure whose characteristic function is the indicator $\mathbf{1}_B$ (we have to assume the existence of such a probability measure). Then, if μ is the distribution of a random variable X, we have

$$\begin{aligned}
\mathbf{P}[X \in B] &= \int_{-\infty}^{\infty} \mathbf{1}_B(x) \mu[\mathrm{d}x] \\
&= \langle \mathbf{1}_B, \mu \rangle \\
&= \left\langle \hat{\mu}, \check{\mathbf{1}}_B \right\rangle \\
&= \int_{-\infty}^{\infty} \hat{\mu}(u)\, \check{\mathbf{1}}_B[\mathrm{d}u].
\end{aligned}$$

Then the Lévy's inversion formula 6.3.7 states heuristically that

$$\check{\mathbf{1}}_{[a,b]}[\mathrm{d}u] = \frac{1}{2\pi} \frac{\mathrm{e}^{-\mathrm{i}au} - \mathrm{e}^{-\mathrm{i}bu}}{iu}\, \mathrm{d}u.$$

Let us then prove the Lévy's inversion formula 6.3.7 rigorously. We need the following lemma. Its proof is left as Exercise 6.6.

6.3.8 Lemma. *Denote*

$$\operatorname{sgn}(x) := \begin{cases} -1, & kun\ x < 0, \\ 0, & kun\ x = 0, \\ 1, & kun\ x > 0. \end{cases}$$

Then

$$\lim_{T \to \infty} \int_{-T}^{T} \frac{\sin(xu)}{u}\, \mathrm{d}u = \operatorname{sgn}(x)\pi$$

and

$$\left| \int_{-T}^{T} \frac{\sin(ux)}{u} \, \mathrm{d}u \right| \leq 2\pi.$$

Proof of Theorem 6.3.7. Let $a \leq b$. We denote

$$I_T := \frac{1}{2\pi} \int_{-T}^{T} \frac{\mathrm{e}^{-\mathrm{i}ua} - \mathrm{e}^{-\mathrm{i}ub}}{\mathrm{i}u} \phi_X(u) \, \mathrm{d}u.$$

Now it follows from the Fubini's theorem 5.3.7 that

$$\begin{aligned} I_T &= \frac{1}{2\pi} \int_{-T}^{T} \frac{\mathrm{e}^{-\mathrm{i}ua} - \mathrm{e}^{-\mathrm{i}ub}}{\mathrm{i}u} \left(\int_{-\infty}^{\infty} \mathrm{e}^{\mathrm{i}ux} \, \mathrm{d}F_X(x) \right) \mathrm{d}u \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \left(\int_{-T}^{T} \frac{\mathrm{e}^{-\mathrm{i}ua} - \mathrm{e}^{-\mathrm{i}ub}}{\mathrm{i}u} \mathrm{e}^{\mathrm{i}ux} \mathrm{d}u \right) \mathrm{d}F_X(x) \\ &=: \int_{-\infty}^{\infty} \Theta_T(x) \, \mathrm{d}F_X(x), \end{aligned}$$

where we denoted

$$\Theta_T(x) := \frac{1}{2\pi} \int_{-T}^{T} \frac{\mathrm{e}^{-\mathrm{i}ua} - \mathrm{e}^{-\mathrm{i}ub}}{\mathrm{i}u} \mathrm{e}^{\mathrm{i}ux} \mathrm{d}u.$$

The use of Fubini was justified because

$$\left| \frac{\mathrm{e}^{-\mathrm{i}ua} - \mathrm{e}^{-\mathrm{i}ub}}{\mathrm{i}u} \mathrm{e}^{\mathrm{i}ux} \right| \leq |b - a|.$$

Let us the consider the inner integral $\Theta_T(x)$. After a little bit of compex geometry, we see that

$$\Theta_T(x) = \frac{1}{2\pi} \int_{-T}^{T} \frac{\sin(u(x-a)) - \sin(u(x-b))}{u} \, \mathrm{d}u.$$

Now we use Lemma 6.3.8. We obtain

$$\begin{aligned} \lim_{T \to \infty} I_T &= \lim_{T \to \infty} \int_{-\infty}^{\infty} \Theta_T(x) \, \mathrm{d}F_X(x) \\ &= \int_{-\infty}^{\infty} \lim_{T \to \infty} \Theta_T(x) \, \mathrm{d}F_X(x) \\ &= \frac{1}{2} \int_{-\infty}^{\infty} (\mathrm{sgn}(x-a) - \mathrm{sgn}(y-b)) \, \mathrm{d}F_X(x) \\ &= \frac{1}{2} \left(-\mathbf{P}[X < a] + \mathbf{P}[X > a] + \mathbf{P}[X < b] - \mathbf{P}[X > b] \right) \\ &= \mathbf{P}[a < X \leq b] \\ &= F_X(b) - F_X(a). \end{aligned}$$

This completes the proof. □

The *joint characterisic function* of the random vector $(X_1, X_2, \ldots, X_n)$ is

$$\phi_{X_1,X_2,\ldots,X_n}(u_1, u_2, \ldots, u_n) \;:=\; \mathbf{E}\left[\mathrm{e}^{\mathrm{i}(u_1X_1+u_2X_2+\cdots+u_nX_n)}\right].$$

It can be shown that the joint characteristic function determines uniquely the (cumulative) distribution

$$F_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) := \mathbf{P}[X_1 \le x_1, X_2 \le x_2, \ldots, X_n \le x_n].$$

Indeed, a natural extension of the Lévy's inversion theorem holds. We will not consider that extension in this course.

6.3.9 Corollary. (i) *The characteristic function ϕ_X determines the distribution $\mathbf{P}_X$ uniquely.*

(ii) *X ja Y are independent if and only if*

$$\phi_{X,Y}(u, v) \;=\; \phi_X(u)\,\phi_Y(v).$$

(iii) *If*

$$\int_{-\infty}^{\infty} |\phi_X(u)|\,\mathrm{d}u \;<\; \infty,$$

then X is continuous and its density is given by

$$f_X(x) \;=\; \frac{1}{2\pi}\int_{-\infty}^{\infty} \mathrm{e}^{-\mathrm{i}ux}\phi_X(u)\,\mathrm{d}u.$$

(iv) *If X is discrete, then*

$$\mathbf{P}[X = x_n] \;=\; \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathrm{e}^{-\mathrm{i}ux_n}\phi_X(u)\,\mathrm{d}u.$$

Proof. (i) Exercise 6.7.
(ii) Exercise 6.8.
(iii) Exercise 6.9.
(iv) Exercise 6.10. □

Next Theorem 6.3.10 will be extremely useful in proving the central limit theorem 8.3.1.

6.3.10 Theorem. *Let $X_1, X_2, \ldots X_n$ be independent random variables. Then*

$$\phi_{\sum_{k=1}^n X_k}(u) \;=\; \prod_{k=1}^{n} \phi_{X_k}(u).$$

Proof. Exercise 6.11. □

What Theorem 6.3.10 actually says, is that the characteristic functions transform convolutions into sums.

Exercises

6.1. Let $z = (x, y) \in \mathbb{C} \setminus \{(0, 0)\}$. We set

$$\frac{1}{z} \;:=\; \frac{x}{x^2 + y^2} - \mathrm{i}\frac{y}{x^2 + y^2}.$$

Show that

$$z \cdot \frac{1}{z} \;=\; 1.$$

6.2. Prove Lemma 6.2.1(iii).

6.3. Let X be degerated at a point $x_0 \in \mathbb{R}^d$, i.e.

$$\mathbf{P}[X \in B] \;=\; \delta_{x_0}[B].$$

Calculate ϕ_X.

6.4. Let X be a simple random varaible that gets the values $x_1, \ldots, x_n$ with probabilities $p_1, \ldots, p_n$. Calculate ϕ_X.

6.5. Let X be *exponentially distributied* with parameter $\mu > 0$, i.e. it has the density function

$$f_X(x) \;=\; \mu \mathrm{e}^{-\mu x} \mathbf{1}_{[0,\infty)}(x).$$

Calculate ϕ_X.

6.6. Prove Lemma 6.3.8.

6.7. Prove Corollary 6.3.9(i).

6.8. Prove Corollary 6.3.9(ii).

6.9. Prove Corollary 6.3.9(iii).

6.10. Prove Corollary 6.3.9(iv).

6.11. Prove Theorem 6.3.10.

Chapter 7

Modes of Convergence

Of all the modes of convergence in probability theory, the almost sure (a.s.) convergence is probably the most natural. Other modes of converge is needed mainly for the following two reasons:

1. *Almost sure convergence is a very strong mode of convergence. For example in the cental limit theorem it does not hold.*
2. *If a sequence $(X_n)_{n\in\mathbb{N}}$ or random varaibles converses almost surely to the random variable X, we cannot deduce that $\mathbf{E}[X_n] \to \mathbf{E}[X]$, i.e. the expectation $\mathbf{E}$ is not continuous with respect to almost sure convergence*

Reason 1 leads to the concept of convergence in probability and convergence in distribution (or weak convergence). Reason 2 leads to the concept of L^p convergence.

7.1 Almost Sure Convergence

The most simple form of convergence would the the pointwise convergence: $X_n \to X$ if $X_n(\omega) \to X(\omega)$ for all $\omega \in \Omega$. Unfortunately, this is too strong even for the strong law law of large numbers for the coint tossing. So, instead we consider the following alsmost as good mode of convergence

7.1.1 Definition. The sequence $(X_n)_{n\in\mathbb{N}}$ of random variables converge *almost surely* to the random variable X, if

$$\mathbf{P}\big[\omega \in \Omega\,;\, \lim_{n\to\infty} X_n(\omega) = X(\omega)\big] \;=\; 1.$$

In this case we denote $X_n \overset{a.s.}{\to} X$.

Let us consider briefly the concept of "almost sure". A property is said to hold almost surely, if it holds with probability 1. For example the notation $X \overset{a.s.}{=} Y$ means that $\mathbf{P}[X = Y] = 1$. It is easy to see that the relation $\overset{a.s.}{=}$ is an equivalence:

- $X \overset{a.s.}{=} X$,
- if $X \overset{a.s.}{=} Y$, then $Y \overset{a.s.}{=} X$,
- if $X \overset{a.s.}{=} Y$ and $Y \overset{a.s.}{=} Z$, then $X \overset{a.s.}{=} Z$.

Moreover, $\mathbf{E}[X] = \mathbf{E}[Y]$, if $X \overset{a.s.}{=} Y$. In fact $\mathbf{E}[|X - Y|] = 0$ if and only if $X \overset{a.s.}{=} Y$.

It is natural to identify random variables that are almost sure the same. Indeed, in what follows we shall do precisely that.

Theorem 4.2.1 extends to a.s. convergence. We only state and show the Theorem 4.2.1(i).

7.1.2 Theorem. *Let $(X_n)_{n\in\mathbb{N}}$ be an a.s. increasing sequence of random variables such that $X_n \overset{a.s}{\to} X$. if $\mathbf{E}[X_1] > -\infty$, then $\mathbf{E}[X_n] \uparrow \mathbf{E}[X]$.*

Proof. Let $\Omega_n \subset \Omega$, $n \in \mathbb{N}$, be the subset for which $X_n \leq X_{n+1}$ holds. Let $\Omega_0 \subset \Omega$ be the subset for which $X_n \to X$ holds. Now, of the set

$$\tilde{\Omega} \;:=\; \Omega_0 \cap \bigcap_{n=1}^{\infty} \Omega_n$$

holds: $(X_n)_{n\in\mathbb{N}}$ is increasing and $X_n \to X$. Therefore $(X_n \mathbf{1}_{\tilde{\Omega}})_{n\in\mathbb{N}}$ is increasing and $\mathbf{E}[X_n \mathbf{1}_{\tilde{\Omega}}] \to \mathbf{E}[X \mathbf{1}_{\tilde{\Omega}}]$. But on the other hand

$$\begin{aligned}
\mathbf{P}[\tilde{\Omega}^c] &= \mathbf{P}\left[\left(\bigcap_{n=1}^{\infty}(\Omega_n \cap \Omega_0)\right)^c\right] \\
&= \mathbf{P}\left[\bigcup_{n\in\infty}(\Omega_n \cap \Omega_0)^c\right] \\
&\leq \sum_{n=1}^{\infty} \mathbf{P}[\Omega_n^c \cup \Omega_0^c] \\
&\leq \sum_{n=1}^{\infty} (\mathbf{P}[\Omega_n^c] + \mathbf{P}[\Omega_0^c]) \\
&= 0.
\end{aligned}$$

Thus $\mathbf{P}[\tilde{\Omega}] = 1$. We notice that $\mathbf{E}[X] = \mathbf{E}[Y]$, when $X \overset{a.s.}{=} Y$. Thus $\mathbf{E}[X_n \mathbf{1}_{\tilde{\Omega}}] = \mathbf{E}[X_n]$ and $\mathbf{E}[X \mathbf{1}_{\tilde{\Omega}}] = \mathbf{E}[X]$. □

By the proof above, we can identify the random sequences $X = (X_n)_{n\in\mathbb{N}}$ ja $\tilde{X} = (\tilde{X}_n)_{n\in\mathbb{N}}$, if

$$\mathbf{P}\Big[X_n = \tilde{X}_n\Big] \;=\; 1 \qquad \text{for all } n \in \mathbb{N}.$$

Indeed, we have

$$\begin{aligned}\mathbf{P}\Big[\lim_{n\to\infty} X_n \neq \lim_{n\to\infty} \tilde{X}_n\Big] \;&\leq\; \mathbf{P}\Big[X_n \neq \tilde{X}_n \text{ for some } n \in \mathbb{N}\Big]\\ &\leq\; \sum_{n=1}^{\infty} \mathbf{P}\Big[X_n \neq \tilde{X}_n\Big]\\ &=\; 0.\end{aligned}$$

The following result follows from Borel's and Cantelli's lemma 5.2.1.

7.1.3 Lemma. *Let X be a random variable and let $(X_n)_{n\in\mathbb{N}}$ e a sequence of random variables. If for all $\varepsilon > 0$ we have*

$$\sum_{n=1}^{\infty} \mathbf{P}[|X_n - X| > \varepsilon] \;<\; \infty,$$

then $X_n \overset{a.s.}{\to} X$.

Proof. Exercise 7.5. □

7.1.4 Lemma. *A sufficient condition for the a.s. convergence of the random series $\sum_{n=1}^{\infty} X_n$ is the existence of the series of positive real numbers $\sum_{n=1}^{\infty} \varepsilon_n$ such that*

(i) $\sum_{n=1}^{\infty} \varepsilon_n \;<\; \infty$,
(ii) $\sum_{n=1}^{\infty} \mathbf{P}[|X_n| > \varepsilon_n] \;<\; \infty$.

Proof. Denote

$$A \;:=\; \{|X_n| > \varepsilon_n \text{ ä.u.}\}.$$

Condition (ii) implies, by Borel's and Cantelli's lemma 5.2.1(i), that $\mathbf{P}[A] = 0$. Let then $\omega \in A^c = \{|X_n| \leq \varepsilon_n \text{ ev.}\}$. Then there exists a random constant $n_0(\omega)$ such that $|X_n(\omega)| < \varepsilon_n$, when $n \geq n_0(\omega)$. Tne, by using the majorant principle

$$\begin{aligned}\sum_{n=1}^{\infty} |X_n(\omega)| \;&<\; \sum_{n=1}^{n_0(\omega)} |X_n(\omega)| + \sum_{n=n_0(\omega)}^{\infty} \varepsilon_n\\ &\leq\; \sum_{n=1}^{n_0(\omega)} |X_n(\omega)| + \sum_{n=1}^{\infty} \varepsilon_n\\ &<\; \infty.\end{aligned}$$

So, the series $\sum_{n=1}^{\infty} X_n(\omega)$ converges absolutely for all $\omega \in A^c$. Since $\mathbf{P}[A^c] = 1$, the series $\sum_{n=1}^{\infty} X_n$ (absolutely) converges a.s. □

7.2 L^p Convergence

Without monotonicity or dominance one cannot deduce from $X_n \overset{a.s.}{\to} X$ that $\mathbf{E}[X_n] \to \mathbf{E}[X]$. In this section we present a mode of convergence, where $\mathbf{E}[X_n] \to \mathbf{E}[X]$ is true with out additional assumptions.

Let $p \geq 1$. We denote by L^p the space of random variables X satisfying $\mathbf{E}[|X|^p] < \infty$.

7.2.1 Definition. Let $p \geq 1$. A sequence $(X_n)_{n\in\mathbb{N}}$ converges in L^p to the random variable X, if

$$\mathbf{E}\big[|X_n - X|^p\big] \;\to\; 0.$$

In this case we denote $X_n \overset{L^p}{\to} X$.

7.2.2 Lemma. *Let $p \geq q \geq 1$.*

(i) *If $X \in L^p$, then $X \in L^q$.*
(ii) *If $X_n \overset{L^p}{\to} X$, then $X_n \overset{L^q}{\to} X$.*

Proof. Exercise 7.7. □

7.2.3 Theorem. *If $X_n \overset{L^p}{\to} X$ for some $p \geq 1$, then $\mathbf{E}[X_n] \to \mathbf{E}[X]$.*

Proof. By Lemma 7.2.2 we have $X_n \overset{L^1}{\to} X$. But

$$\begin{aligned} \big|\mathbf{E}[X_n] - \mathbf{E}[X]\big| &= \big|\mathbf{E}[X_n - X]\big| \\ &\leq \mathbf{E}\big[|X_n - X|\big] \\ &\to 0. \end{aligned}$$

This shows the claim. □

The net example shows that L^p convergence does not follow from a.s. convergence, or vice versa.

7.2.4 Example. Let $(\Omega, \mathcal{F}, \mathbf{P}) := ([0,1], \mathcal{B}, \ell)$. This means that we consider the uniform probability measure on the real interval $[0,1]$.

(i) Let

$$X_n(\omega) \;:=\; n\mathbf{1}_{\left(0,\frac{1}{n}\right)}(\omega).$$

Clearly $X_n \overset{a.s.}{\to} 0$. But

$$\begin{aligned}
\mathbf{E}\big[|X_n - 0|^p\big] &= \mathbf{E}[|X_n|^p] \\
&= \int_0^1 |X(\omega)|^p \,\mathrm{d}\omega \\
&= \int_0^1 \left|n\mathbf{1}_{\left(0,\frac{1}{n}\right)}(\omega)\right|^p \,\mathrm{d}\omega \\
&= n^p \int_0^1 \mathbf{1}_{\left(0,\frac{1}{n}\right)}(\omega)\,\mathrm{d}\omega \\
&= n^{p-1} \\
&\not\to 0,
\end{aligned}$$

since $p \geq 1$.

(ii) Let $m \in \mathbb{N}$ and $\Delta_{m,k}$, $k \leq 2^m$ be the equidistant partition of the interval $[0,1]$. Note that $|\Delta_{m,k}| = 2^{-m}$. We then define a sequence of ransom variables $(X_n)_{n\in\mathbb{N}}$ as

$$X_1 := \mathbf{1}_{\Delta_{1,1}},\;\; X_2 := \mathbf{1}_{\Delta_{1,2}},\;\; \ldots\;\; , X_{2^m} := \mathbf{1}_{\Delta_{1,2^m}},$$

$$X_{2^m+1} := \mathbf{1}_{\Delta_{2,1}},\;\; \ldots,\;\; X_{k2^m+1} := \mathbf{1}_{\Delta_{k,1}}, \ldots$$

Then $(X_n)_{n\in\mathbb{N}}$ cannot converge a.s., since $\limsup_{n\to\infty} X_n = 1$ and $\liminf_{n\to\infty} X_n = 0$. But $X_n \overset{L^p}{\to} 0$, since

$$\lim_{n\to\infty} \mathbf{E}[|X_n|^p] \;=\; \lim_{n\to\infty} \mathbf{P}[\Delta_{n,1}] \;=\; 0.$$

7.3 Convergence in Probability

The concept of convergence in probability is used very often in statistics. For example, an estimator is called consistent if it converges in probability to the quantity being estimated. Convergence in probability is also the mpde of convergence established by the weak law of large numbers.

The basic idea behind this type of convergence is that the probability of an "unusual" outcome becomes smaller and smaller as the sequence progresses.

7.3.1 Definition. A sequence $(X_n)_{n\in\mathbb{N}}$ of random variavles converge *in probability* to a random variable X, if for all $\varepsilon > 0$ holds

$$\mathbf{P}\big[|X_n - X| > \varepsilon\big] \;\to\; 0.$$

In this case we denote $X_n \overset{\mathbf{P}}{\to} X$.

7.3.2 Theorem. *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of random variables.*

(i) *If $X_n \xrightarrow{a.s.} X$, then $X_n \xrightarrow{\mathbf{P}} X$.*
(ii) *if $X_n \xrightarrow{L^p} X$ for some $p \geq 1$, then $X_n \xrightarrow{\mathbf{P}} X$.*

Proof. (i) Exercise 7.11.

(ii) By Lemma 7.2.2 it is enough to show L^1-convergence. But the claim follows by using Chebyshe'v inequality 4.4.3. Indeed,

$$\begin{aligned}\mathbf{P}[|X_n - X| > \varepsilon] &\leq \frac{1}{\varepsilon}\mathbf{E}[|X_n - X|] \\ &\to 0\end{aligned}$$

for all $\varepsilon > 0$. □

Next example shows that convergence in probability is strictly weaker than a.s. or L^p convergence.

7.3.3 Example. Let $(\Omega, \mathcal{F}, \mathbf{P})$ ja $\Delta_{m,k}$ be as in Example 7.2.4. Let also $X_n = \mathbf{1}_{\Delta_{k,l}}$, $n = m2^k + l$, be as in Example 7.2.4.

(i) The sequence $(X_n)_{n\in\mathbb{N}}$ still does not converge a.s., but for all $\varepsilon > 0$ we have
$$\lim_{n\to\infty} \mathbf{P}[|X_n| > \varepsilon] = \lim_{n\to\infty} \mathbf{P}[\Delta_{n,1}] = 0.$$
Consequently $X_n \xrightarrow{\mathbf{P}} 0$.
(ii) Let then $Y_n := a_n X_n$, where
$$a_n = \frac{n}{\mathbf{P}[\Delta_{k,l}]},$$
when $n = m2^k + l$. Then $(Y_n)_{n\in\mathbb{N}}$ does not converge in L^p, because
$$\mathbf{E}[|Y_n|^p] = n^p \to \infty.$$
But
$$\lim_{n\to\infty} \mathbf{P}[|Y_n| > \varepsilon] = \lim_{n\to\infty} \mathbf{P}[\Delta_{n,1}] = 0.$$
Consequently $X_n \xrightarrow{\mathbf{P}} 0$.

7.3.4 Theorem. *If $X_n \xrightarrow{\mathbf{P}} X$, then there exists a subsequence $(n_k)_{k\in\mathbb{N}}$ such that $X_{n_k} \xrightarrow{a.s.} X$.*

Proof. Let us construct the desired subsequence $(n_k)_{k\in\mathbb{N}}$. Let $n_1 = 1$. We find the index n_k inductively: n_k is the smallest index that is greater than n_{k-1} and satisfies

$$\mathbf{P}\big[|X_{n_k} - X| > 2^{-k}\big] \;<\; 2^{-k}.$$

Since $X_n \xrightarrow{\mathbf{P}} X$, such index n_k can be found for all $k \in \mathbb{N}$. Thus we have found a subsequence $(n_k)_{k\in\mathbb{N}}$.

Now

$$\sum_{k=1}^{\infty} \mathbf{P}\big[|X_{n_k} - X| > 2^{-k}\big] \;<\; \sum_{k=1}^{\infty} 2^{-k} \;<\; \infty.$$

Therefore, by Borel's and Cantelli's lemma 5.2.1(i), we have

$$\mathbf{P}\big[|X_{n_k} - X| > 2^{-k} \text{ i.o.}\big] \;=\; 0.$$

But this means that $X_{n_k} \xrightarrow{a.s.} X$. Indeed, for all $\varepsilon > 0$ we have

$$\begin{aligned} 0 \;&=\; \mathbf{P}\big[|X_{n_k} - X| > 2^{-k} \text{ i.o.}\big] \\ &\geq\; \mathbf{P}[|X_{n_k} - X| > \varepsilon \text{ i.o.}]\,. \end{aligned}$$

Since $\varepsilon > 0$ was arbitrary, this means that $\mathbf{P}[X_{n_k} \not\to X] = 0$. □

7.4 Convergence in Distribution

Not even the convergence in probability is weak enough for the central limit theorem. We need yet a weaker mode of convergence that we call the convergence in distribution.

7.4.1 Definition. A sequence $(X_n)_{n\in\mathbb{N}}$ of random variables converge *in distribution* to a random variable X if the corresponding cumulative distribution functions converge

$$F_{X_n} \to F_X$$

in all the continuity points of F_X. In this case we denote $X_n \xrightarrow{d} X$.

The next example shows that we indeed need the condition only for the continuity points of the limiting cumulative distribution function.

7.4.2 Example. Let $(\Omega, \mathcal{F}, \mathbf{P}) := ([0,1], \mathcal{B}, \ell)$. So, we consider yet again the uniform distribution on the inverval $[0,1]$. We set

$$X_n(\omega) \;:=\; \omega^n.$$

Then it is eqasy to see that $X_n \to 0$ a.s, in L^p, and in probability. (We show later in Theorem 7.4.4, that then $X_n \xrightarrow{d} 0$.) Now the cumulative distribution function of the "random" variable $X = 0$ is $F_X(x) = \mathbf{1}_{[0,\infty)}(x)$. But

$$\begin{aligned} F_{X_n}(x) &= \mathbf{P}[X_n \le x] \\ &= \ell[\omega \in [0,1]\,;\, \omega^n \le x] \\ &= \ell\big[\omega \in [0,1]\,;\, \omega \le x^{1/n}\big] \\ &= x^{1/n}\mathbf{1}_{[0,1]}(x) + \mathbf{1}_{(1,\infty)}(x) \\ &\to \mathbf{1}_{(0,\infty)}(x). \end{aligned}$$

Thus $0 = F_{X_n}(0) \not\to F_X(0) = 1$. We see that the cumulative distribution functions F_n, $n \in \mathbb{N}$, do not converge to the limiting distribution function F_X in its point of discontinuity $x = 0$.

Convergence in distribution can be characterized many ways as the following Portmanteau theorem states

7.4.3 Theorem. *The following are equivalent:*

(i) $X_n \xrightarrow{d} X$.
(ii) $\lim_{n\to\infty} F_{X_n} = F_X$ *in all the continuity points of* F_X.
(iii) $\lim \mathbf{P}[X_n \in B] = \mathbf{P}[X \in B]$ *for all* $B \in \mathcal{B}$, *such that* $\mathbf{P}[X \in \partial B] = 0$. *Notation* ∂B *means the boundary of* B.
(iv) $\limsup_{n\to\infty} \mathbf{P}[X_n \in F] \le \mathbf{P}[X \in F]$ *for all closed* $F \subset \mathbb{R}$.
(v) $\liminf_{n\to\infty} \mathbf{P}[X_n \in G] \ge \mathbf{P}[X \in G]$ *for all open* $G \subset \mathbb{R}$.
(vi) $\lim_{n\to\infty} \mathbf{E}[g(X_n)] = \mathbf{E}[g(X)]$ *for all bounded and continuous functions* $g\colon \mathbb{R} \to \mathbb{R}$.
(vii) $\lim_{n\to\infty} \phi_{X_n} = \phi_X$.

Proof. This is a good exercise for a master's thesis. □

By Theorem 7.4.4 below, the convergence in distribution is the weakest mode of convergence.

7.4.4 Theorem. *If* $X_n \xrightarrow{\mathbf{P}} X$, *then* $X_n \xrightarrow{d} X$.

Proof. Exercise 7.15. □

Example 7.4.5 below shows that the convergence in distribution is strictly weaker than the convergence in probability.

7.4.5 Example. Let X and $\tilde{X}$ be independent standard normally distributed, i.e. with parameters $\mu = 0$ and $\sigma^2 = 1$. Let $X_n = X$, when n is odd and $X_n = \tilde{X}$, when n is even. Now

$$F_{X_n}(x) \;=\; \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \mathrm{e}^{-\frac{1}{2}y^2}\,\mathrm{d}y$$

for all $n \in \mathbb{N}$, so it is evident that $(X_n)_{n\in\mathbb{N}}$ converges in distribution. On the other hand $X - \tilde{X}$ is "genuinely random". In fact, it can be shown that $X - \tilde{X}$ is normally distributed with parameters $\mu = 0$ and $\sigma^2 = 2$. Therefore the probability $\mathbf{P}[|X_n - X_{n+1}| > \varepsilon]$ is a positive number depending on $\varepsilon > 0$, but independent of the index n. Therefore $(X_n)_{n\in\mathbb{N}}$ cannot converge in probability.

Despite of Example 7.4.5 we have the following partial inverse.

7.4.6 Theorem. *If $X_n \xrightarrow{d} X$ and X is deterministic, then $X_n \xrightarrow{\mathbf{P}} X$.*

Proof. Let $c := X \in \mathbb{R}$ be constant and let $\varepsilon > 0$ be arbitrary. The proof is based on the observation that the set $\{x \in \mathbb{R}; |x - c| < \varepsilon\}$ is open. Indeed, by the Portmanteau theorem 7.4.3(v) we have that

$$\liminf_{n\to\infty} \mathbf{P}[|X_n - c| < \varepsilon] \;\geq\; \mathbf{P}[|X - c| < \varepsilon] \;=\; 1.$$

Thus $\lim_{n\to\infty} \mathbf{P}[|X_n - c| < \varepsilon] = 1$, which means that $X_n \xrightarrow{\mathbf{P}} c = X$. □

Theorem 7.4.7 below is the celebrated *Lévy's continuity theorem*. It characterizes the convergence in distribution in tems of characteristic functions.

7.4.7 Theorem. *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of random variables.*

(i) *If $X_n \xrightarrow{d} X$, then $\phi_{X_n} \to \phi_X$.*
(ii) *If $\phi_{X_n} \to \phi$ and ϕ is continuous at the origin, then there exist a random variable X, such that ϕ is its chartacteristic function and $X_n \xrightarrow{d} X$.*

Proof. This is a good exercise for a master's thesis. □

7.5 Relationships between the Modes of Convergences

Theorem 7.5.1 below is a compilation theorem where we have collected together the interrelations of different modes of convergece. The proof of the complication theorem is left as Exercise 7.17.

7.5.1 Theorem. *The following implications hold true and none of them can be reversed*

(i) *If* $X_n \xrightarrow{a.s.} X$, *then* $X_n \xrightarrow{\mathbf{P}} X$.
(ii) *If* $X_n \xrightarrow{L^p} X$ *for some* $p \geq 1$, *then* $X_n \xrightarrow{\mathbf{P}} X$.
(iii) *Let* $p \geq q \geq 1$. *If* $X_n \xrightarrow{L^p} X$, *then* $X_n \xrightarrow{L^q} X$.
(iv) *If* $X_n \xrightarrow{\mathbf{P}} X$, *then there exists a subsequence* $(n_k)_{k\in\mathbb{N}}$ *such that* $X_{n_k} \xrightarrow{a.s.} X$.
(v) *Id* $X_n \xrightarrow{\mathbf{P}} X$, *then* $X_n \xrightarrow{d} X$.

Exercises

7.1. Show the dominated condergence theorem 4.2.1(v), when we only assume that $X_n \xrightarrow{a.s.} X$ and $|X_n| \leq Y$ a.s. for some $Y \in L^1$.

7.2. Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of random variables. Show that $X_n \to X$ precisely in the set

$$\bigcap_{m=1}^{\infty} \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} \left\{ |X_k - X| \leq \frac{1}{m} \right\}.$$

7.3. Let $X_n \xrightarrow{a.s.} X$ and let $f\colon \mathbb{R} \to \mathbb{R}$ be continuous. Show that $f(X_n) \xrightarrow{a.s.} f(X)$.

7.4. Let $X_n \xrightarrow{a.s.} X$, $Y_n \xrightarrow{a.s.} Y$ and $a, b \in \mathbb{R}^d$. Show that

$$aX_n + bY_n \xrightarrow{a.s.} aX + bY.$$

7.5. Prove Lemma 7.1.3.

7.6. Show that the condition $\sum_{n=1}^{\infty} \mathbf{P}[|X_n - X| > \varepsilon] < \infty$ of Lemma 7.1.3 is not necessary for the a.s. convergence.

7.7. Prove Lemma 7.2.2.

7.8. Let $X_n \xrightarrow{L^p} X$, $Y_n \xrightarrow{L^p} Y$ and $a, b \in \mathbb{R}$. Show that

$$aX_n + bY_n \xrightarrow{L^p} aX + bY.$$

7.9. Let $X_n \xrightarrow{L^p} X$ and let $f\colon \mathbb{R} \to \mathbb{R}$ be continuous. Is it true that then $f(X_n) \xrightarrow{L^p} f(X)$?

7.10. Shoe that Lemma 7.2.2 does not hold if instead of probability measures we consider general unbounded measures.

7.11. Prove Theorem 7.3.2(i).

7.12. Show that $X_n \xrightarrow{\mathbf{P}} X$ if and only if

$$\mathbf{E}\left[\frac{|X_n - X|}{1 + |X_n - X|}\right] \;\to\; 0.$$

7.13. Let $X_n \xrightarrow{\mathbf{P}} X$ and let $f\colon \mathbb{R} \to \mathbb{R}$ continuous. Does it hold then that $f(X_n) \xrightarrow{\mathbf{P}} f(X)$?

7.14. Let $X_n \xrightarrow{\mathbf{P}} X$, $Y_n \xrightarrow{\mathbf{P}} Y$ and $a, b \in \mathbb{R}^d$. Does it then hold that

$$aX_n + bY_n \;\xrightarrow{\mathbf{P}}\; aX + bY?$$

7.15. Prove Theorem 7.4.4.

7.16. Let $X_n \xrightarrow{d} X$ and $f\colon \mathbb{R} \to \mathbb{R}$ be continuous on all the continuity points pf F_X. Show that then $f(X_n) \xrightarrow{d} f(X)$.

7.17. Collect the proof of Theorem 7.5.1.

7.18. Are the modes of convergence consistent? In other words, if $X_n \to X$ in some sense and $X_n \to Y$ in another sense, do we necessarily have $X = Y$?

Chapter 8

Limit Theorems

Now we arrive at our grande finale: the great limit theorems of probability theory: the laws of large numbers, the large deviations principles, and the central limit theorems. These limit theorems tell us, that while random events are unpredictable, their sums are regular, in sense that the randomness vanishes in the limit as the number of summands increase and we take averages. Thus, these limit theorems bring order out of chaos.

8.1 Laws of Large Numbers

We first recall the notion of correlation.

If $X \in L^2$, we say that X is *square integrable*.

Recall the *covariance*

$$\mathbf{Cov}[X,Y] \;:=\; \mathbf{E}\left[(X-\mathbf{E}[X])(Y-\mathbf{E}[Y])\right].$$

Square integrable random variables X_n, $n \in \mathbb{N}$, are *uncorrelated* if for all $n \neq m$ we have

$$\mathbf{E}[X_n X_m] \;=\; \mathbf{E}[X_n]\mathbf{E}[X_m].$$

The name "uncorrelated" comes from the following fact. The *correlation* between the random variables X_n and X_m is then

$$\mathbf{Cor}[X_n,X_m] \;:=\; \frac{\mathbf{Cov}[X_n,X_m]}{\sqrt{\mathbf{Var}[X_n]\mathbf{Var}[X_m]}} \;=\; 0.$$

Now, it is easy to see that for uncorrelated random variables we have

$$\mathbf{Var}\left[\sum_{k=1}^{n} X_k\right] \;=\; \sum_{k=1}^{n} \mathbf{Var}[X_k]$$

(Exercise 8.1).

The following Theorem 8.1.1 is the *weak law of large numbers* for uncorrelated square integrable random variables.

8.1.1 Theorem. *Let $(X_n)_{n\in\mathbb{N}} \subset L^2$ be a sequence of uncorrelated random variables such that*

$$\begin{aligned} \mathbf{E}[X_n] &= \mu, \\ \frac{1}{n^2}\sum_{k=1}^{n} \mathbf{Var}[X_k] &\to 0. \end{aligned}$$

Then

$$\frac{1}{n}\sum_{k=1}^{n} X_k \xrightarrow{\mathbf{P}} \mu.$$

Proof. Denote

$$S_n := \sum_{k=1}^{n} X_k.$$

Then

$$\begin{aligned} \mathbf{E}[S_n] &= \mathbf{E}\left[\sum_{k=1}^{n} X_k\right] \\ &= \sum_{k=1}^{n} \mathbf{E}[X_k] \\ &= \sum_{k=1}^{n} \mu \\ &= n\mu \end{aligned}$$

and

$$\begin{aligned} \mathbf{Var}[S_n] &= \mathbf{Var}\left[\sum_{k=1}^{n} X_k\right] \\ &= \sum_{k=1}^{n} \mathbf{Var}[X_k]. \end{aligned}$$

Now, by Markov's inequality 4.4.4(i), we have

$$\begin{aligned}
\mathbf{P}\left[\left|\frac{1}{n}\sum_{k=1}^{n} X_k - \mu\right| > \varepsilon\right] &= \mathbf{P}\Big[\big|S_n - n\mu\big| > n\varepsilon\Big] \\
&= \mathbf{P}\Big[\big|S_n - \mathbf{E}[S_n]\big| > n\varepsilon\Big] \\
&\leq \frac{\mathbf{Var}[S_n]}{n^2\varepsilon^2} \\
&= \frac{1}{\varepsilon^2}\frac{1}{n^2}\sum_{k=1}^{n}\mathbf{Var}[X_k] \\
&\to 0,
\end{aligned}$$

which shows the claim. □

8.1.2 Corollary. *Let $(X_n)_{n\in\mathbb{N}} \subset L^2$ be a sequence of uncorrelated identically distributed random variables. Then*

$$\frac{1}{n}\sum_{k=1}^{n} X_k \xrightarrow{\mathbf{P}} \mathbf{E}[X_1].$$

Let us note that the weak law of large numbers Theorem 8.1.1 holds for correlated random variables also. We just need to assume that

$$\frac{1}{n^2}\mathbf{Var}\left[\sum_{k=1}^{n} X_k\right] \to 0.$$

8.1.3 Example. Let us consider an application of the weak law of large numbers, the so-called *Weierstraß's approximation theorem*:

> Let $f\colon [0,1] \to \mathbb{R}$ be a continuous function. Then for all $\varepsilon > 0$ there exists a polynomial P such that
>
> $$\sup_{x\in[0,1]} |f(x) - P(x)| < \varepsilon.$$

In the approximation above we use *Bernstein's polynomials*

$$B_n(x) := \sum_{k=0}^{n} f\left(\frac{k}{n}\right)\binom{n}{k}x^k(1-x)^{n-k}.$$

Let S_n be binomially distributed with parameters n ja x, i.e.

$$S_n = \sum_{k=1}^{n} X_k,$$

where X_k's are independent binomially distributed with parameters1 ja x. Now

$$\mathbf{P}[S_n = k] \;=\; \binom{n}{k} x^k (1-x)^{n-k}.$$

So

$$\mathbf{E}\left[f\left(\frac{1}{n}S_n\right)\right] \;=\; \sum_{k=0}^{n} f\left(\frac{k}{n}\right)\binom{n}{k} x^k (1-x)^{n-k} \;=\; B_n(x).$$

Let then $\varepsilon > 0$. Since f is a continuous function on a closed interval, it is *uniformly continuous*. In other words, for every $\varepsilon > 0$ there exist a $\delta = \delta(\varepsilon)$, such that

$$\sup_{|x-y|\le\delta} |f(x) - f(y)| \;<\; \frac{1}{2}\varepsilon.$$

Now, for every $x \in [0,1]$ we have

$$\left|B_n(x) - f(x)\right| \;=\; \left|\mathbf{E}\left[f\left(\frac{1}{n}S_n\right) - f(x)\right]\right|.$$

Let us then denote

$$\begin{aligned} Y_n &:= \left|f\left(\frac{1}{n}S_n\right) - f(x)\right|, \\ Z_n &:= \left|\frac{1}{n}S_n - x\right|. \end{aligned}$$

Now $Y_n < \frac{1}{2}\varepsilon$, when $Z_n < \delta$. Thus

$$\begin{aligned} \left|B_n(x) - f(x)\right| &\le \mathbf{E}[Y_n] \\ &= \mathbf{E}\big[Y_n \mathbf{1}_{\{Z_n \le \delta\}}\big] + \mathbf{E}\big[Y_n \mathbf{1}_{\{Z_n > \delta\}}\big] \\ &\le \frac{1}{2}\varepsilon \mathbf{P}[Z_n \le \delta] + 2 \max_{y\in[0,1]} |f(y)| \mathbf{P}[Z_n > \delta] \\ &\le \frac{1}{2}\varepsilon + 2 \max_{y\in[0,1]} |f(y)| \, \frac{1}{4n\delta^2}. \end{aligned}$$

Let us then choose n big enough so, that

$$2 \max_{y\in[0,1]} |f(y)| \, \frac{1}{4n\delta^2} \;<\; \frac{1}{2}\varepsilon.$$

Then we have $|B_n(x) - f(x)| < \varepsilon$. Since the choice of δ did not depend on the point x, we have proven the claim.

Let us then consider the strong law of large numbers. We begin the the so-called *Skorohod's lemma* or *Skorohod's maximal inequality* 8.1.4 below.

8.1.4 Lemma. *Let $X_1, \ldots, X_n$ be independent random variables and let $S_n := \sum_{k=1}^{n} X_k$. Suppose*

$$c_n \;:=\; \max_{k\le n} \mathbf{P}\big[\big|S_n - S_k\big| > s\big] \;<\; 1.$$

Then

$$\mathbf{P}\left[\max_{k\le n}\big|S_k\big| > 2s\right] \;\le\; \frac{1}{1-c_n}\mathbf{P}\big[\big|S_n\big| > s\big]\,.$$

Proof. Exercise 8.4. □

The next interesting result shows that in the independent case the convergence of series in probability is in fact almost sure convergence of the said series.

8.1.5 Theorem. *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent random variables. Then the series $\sum_{n=1}^{\infty} X_n$ converges almost surely if and only if it converges in probability.*

Proof. Clearly it is enough to show the "if" part

We proceed with a proof by contradiction.

If a sequence $(s_n)_{n\in\mathbb{N}}$ of real numbers does not converge, then there exists an $\varepsilon > 0$ such that for all $m \in \mathbb{N}$ we have

$$\sup_{n;n>m} |s_n - s_m| \;>\; \varepsilon.$$

Now, by Corollary 5.2.6, the series $\sum_{n=1}^{\infty} X_n$ either converges or diverges a.s.

Let us assume that the series $\sum_{n=1}^{\infty} X_n$ diverges (with a positive probability). This means that there exist an $\varepsilon > 0$ and $\delta > 0$ such that for all $m \in \mathbb{N}$ we have

$$\mathbf{P}\left[\sup_{n\ge m}\left|\sum_{k=m+1}^{n} X_k\right| > \varepsilon\right] \;\ge\; \delta. \tag{8.1.6}$$

Let then $N \in \mathbb{N}$. By Skorohod's lemma 8.1.4 we have that

$$\mathbf{P}\left[\max_{m<n\le N}\left|\sum_{k=m+1}^{n} X_k\right| > \varepsilon\right] \;\le\; \frac{1}{1-C_{m,N}}\mathbf{P}\left[\left|\sum_{k=m+1}^{N} X_k\right| > \frac{\varepsilon}{2}\right],$$

where

$$C_{m,N} \;:=\; \max_{m<n\le N} \mathbf{P}\left[\left|\sum_{k=n+1}^{N} X_k\right| > \frac{\varepsilon}{2}\right].$$

Let us then assume that the series $\sum_{n=1}^{\infty} X_n$ converges in probability. Then

$$\sum_{k=m+1}^{N} X_k \;\xrightarrow{\mathbf{P}}\; 0,$$

as $m, N \to \infty$. In other words, when $m, N \to \infty$, we have

$$\mathbf{P}\left[\left|\sum_{k=m+1}^{N} X_k\right| > \frac{\varepsilon}{2}\right] \;\to\; 0.$$

Moreover,

$$C_{m,N} \;\to\; 0.$$

Let then first $N \to \infty$. We see that

$$\lim_{m\to\infty} \mathbf{P}\left[\sup_{n>m}\left|\sum_{k=m+1}^{n} X_k\right| > \varepsilon\right] \;=\; 0.$$

But this is in contradiction with the positive divergence (8.1.6) of the series. Hence, the series converges. □

8.1.7 Corollary. *Let $(X_n)_{n\in\mathbb{N}} \subset L^2$ be a sequence of independent random variables with $\mathbf{E}[X_n] = 0$. If the series $\sum_{n=1}^{\infty} \mathbf{E}[X_n^2]$ converges, then the series $\sum_{n=1}^{\infty} X_n$ converges a.s.*

Next we prove the *strong law of large numbers* for centered square integrable random variables. Before that we still need the following so-called *Kronecker's lemma* 8.1.8.

8.1.8 Lemma. *Let $\sum_{n=1}^{\infty} x_n$ be a convergent series. Let $b_n \uparrow \infty$. Then*

$$\frac{1}{b_n}\sum_{k=1}^{n} b_k x_k \;\to\; 0.$$

Proof. Exercise 8.6. □

8.1.9 Theorem. *Let $(X_n)_{n\in\mathbb{N}} \subset L^2$ be a sequence of independent random variables with $\mathbf{E}[X_n] = 0$. Let $b_n \uparrow \infty$ be such a sequence of real numbers that*

$$\sum_{n=1}^{\infty} \frac{\mathbf{E}[X_n^2]}{b_n^2} \;<\; \infty.$$

Then

$$\frac{1}{b_n}\sum_{k=1}^{n} X_k \quad \overset{a.s.}{\to} \quad 0.$$

Proof. By Kronecker's lemma 8.1.8 the claim follows fron the a.s. convergence of the series $\sum_{n=1}^{\infty} X_n/b_n$. But this follows from Corollary 8.1.7, since the series

$$\sum_{n=1}^{\infty} \frac{\mathbf{E}[X_n^2]}{b_n^2} \quad = \quad \sum_{n=1}^{\infty} \mathbf{E}\left[\left(\frac{X_n}{b_n}\right)^2\right]$$

converges. □

In Theorem 8.1.1 and Theorem 8.1.9 we had to assume that the random variables are square integrable. We can relax this assumption to integrable random variables. The integrability, i.e. $(X_n)_{n\in\mathbb{N}} \subset L^1$ we have to assume, since the expected limit is $\mu = \mathbf{E}[X_n]$.

The next result is the celebrated *Kolmogorov's strong law of large numbers.*

8.1.10 Theorem. *Let $(X_n)_{n=1}^{\infty} \subset L^1$ be a sequence of independent identically distributed randdom variables. Denote $\mu := \mathbf{E}[X_1]$. Then*

$$\frac{1}{n}\sum_{k=1}^{n} X_k \quad \overset{a.s.}{\to} \quad \mu.$$

If $(X_n)_{n=1}^{\infty} \not\subset L^1$, i.e. $\mathbf{E}[|X_1|] = \infty$, the series above diverges a.s.

Proof. We only consider the case $\mathbf{E}[|X_1|] < \infty$. The case $\mathbf{E}[|X_1|] = \infty$ is left as Exercise 8.7.

The proof is based on our previous law of large numbers for squeare integrable random variables Theorem 8.1.9, and a cutting argument. Indeed, let us define the cut random variables

$$\tilde{X}_n \quad := \quad X_n \mathbf{1}_{\left\{|X_n|\leq n\right\}}.$$

Now

$$\begin{aligned}
\sum_{n=1}^{\infty} \mathbf{P}\left[X_n \neq \tilde{X}_n\right] \quad &= \quad \sum_{n=1}^{\infty} \mathbf{P}[|X_n| > n] \\
&\leq \quad \sum_{n=0}^{\infty} \mathbf{P}[|X_1| > n] \\
&\leq \quad \mathbf{E}[|X_1|] \\
&< \quad \infty.
\end{aligned}$$

By using Borel's and Cantelli's lemma 5.2.1(i) we see from this that

$$\mathbf{P}\Big[X_n \neq \tilde{X}_n \text{ i.o.}\Big] \;=\; 0.$$

Therefore, it is enough to show that

$$\frac{1}{n}\sum_{k=1}^{n} \tilde{X}_k \;\overset{a.s.}{\to}\; \mu.$$

Indeed, if $n_0(\omega) \in \mathbb{N}$ is such that $\tilde{X}_n(\omega) = X_n(\omega)$, when $n \geq n_0(\omega)$, we have

$$\begin{aligned}
\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} X_k(\omega) &= \lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n_0(\omega)} X_k(\omega) + \lim_{n\to\infty} \frac{1}{n}\sum_{k=n_0(\omega)+1}^{n} X_k(\omega) \\
&= \lim_{n\to\infty} \frac{1}{n}\sum_{k=n_0(\omega)+1}^{k} \tilde{X}_k(\omega) \\
&= \lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} \tilde{X}_k(\omega).
\end{aligned}$$

Now $\mathbf{E}[\tilde{X}_n] \to \mathbf{E}[X_1] = \mu$. So, the claim follows if

$$\frac{1}{n}\sum_{k=1}^{n} \Big(\tilde{X}_k - \mathbf{E}[\tilde{X}_k]\Big) \;\overset{a.s.}{\to}\; 0.$$

We argue this by using the square integrable law of large numbers 8.1.9. Because of the cutting, the random varaibles $\tilde{X}_n$, $n \in \mathbb{N}$, are square integrable. Since

$$\mathbf{E}\left[\Big(\tilde{X}_n - \mathbf{E}[\tilde{X}_n]\Big)^2\right] \;\leq\; \mathbf{E}[\tilde{X}_n^2],$$

it is enough to show the convergent series condition

$$\sum_{n=1}^{\infty} \frac{\mathbf{E}[\tilde{X}_n^2]}{n^2} \;<\; \infty.$$

But

$$\begin{aligned}
\sum_{n=1}^{\infty} \frac{\mathbf{E}[\tilde{X}_n^2]}{n^2} &= \sum_{n=1}^{\infty} \frac{1}{n^2} \mathbf{E}\big[X_1^2 \mathbf{1}_{\{|X_1| \le n\}}\big] \\
&= \sum_{n=1}^{\infty} \sum_{k=1}^{n} \frac{1}{n^2} \mathbf{E}\big[X^2 \mathbf{1}_{\{k-1<|X_1| \le k\}}\big] \\
&= \sum_{k=1}^{\infty} \sum_{n=k+1}^{\infty} \frac{1}{n^2} \mathbf{E}\big[X_1^2 \mathbf{1}_{\{k-1<|X_1| \le k\}}\big] \\
&= \sum_{k=1}^{\infty} \mathbf{E}\big[X_1^2 \mathbf{1}_{\{k-1<|X_1| \le k\}}\big] \sum_{n=k+1}^{\infty} \frac{1}{n^2} \\
&\le \sum_{k=1}^{\infty} \mathbf{E}\big[X_1^2 \mathbf{1}_{\{k-1<|X_1| \le k\}}\big] \int_k^{\infty} \frac{\mathrm{d}x}{x^2} \\
&= \sum_{k=1}^{\infty} \mathbf{E}\big[X_1^2 \mathbf{1}_{\{k-1<|X_1| \le k\}}\big] \frac{1}{k} \\
&= \sum_{k=1}^{\infty} \mathbf{E}\left[\frac{X_1^2}{k} \mathbf{1}_{\{k-1<|X_1| \le k\}}\right] \\
&\le \sum_{k=1}^{\infty} \mathbf{E}\left[\frac{X_1^2}{|X_1|} \mathbf{1}_{\{k-1<|X_1| \le k\}}\right] \\
&= \mathbf{E}[|X_1|] \\
&< \infty.
\end{aligned}$$

This completes the proof. □

8.1.11 Example. As an application of the strong law of large numbers we consider the so-called *Monte Carlo integration*.

Let $g\colon \mathbb{R} \to \mathbb{R}$ be Borel measurable. We want to calculate numerically the integral

$$I := \int_a^b g(x)\,\mathrm{d}x.$$

Let $U_1, U_2, \ldots$ be independent uniformy on the interval $[a, b]$ distributed random variables. Let

$$X_k := g(U_k)$$

and

$$S_n := \sum_{k=1}^{n} X_k.$$

Then

$$\mu \;:=\; \mathbf{E}[X_k] \;=\; \int_a^b g(x)\frac{1}{b-a}\,\mathrm{d}x \;=\; \frac{1}{b-a}\,I.$$

Now the strong law of large numbers tells us that

$$(b-a)\frac{1}{n}S_n \;\overset{a.s.}{\to}\; (b-a)\mu \;=\; I.$$

8.1.12 Example. Another application of the law of large numbers is the estimation of an unknown distribution. We have independent observations $X_1, X_2, \ldots$ with unknown cumulative distribution function $F(x) = \mathbf{P}[X \leq x]$. We define the *empirical cumulative distribution function* as

$$\hat{F}_n(x) \;:=\; \frac{\big|\,\{k \leq n\,;\, X_k \leq x\}\,\big|}{n} \;=\; \frac{1}{n}\sum_{k=1}^{n} \mathbf{1}_{\{X_k \leq x\}}.$$

Now the indicators $\mathbf{1}_{\{X_k \leq x\}}$, $k \in \mathbb{N}$, are independent and identically distributed. So the strong law of large numbers state that

$$\hat{F}_n(x) \;\overset{a.s.}{\to}\; \mathbf{E}\big[\mathbf{1}_{\{X_k \leq x\}}\big] \;=\; \mathbf{P}[X_k \leq x] \;=\; F(x).$$

We end this section by stating the *Kolmogorov's three series theorem* that characterizes the a.s. convergence of independent series.

8.1.13 Theorem. *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent random variables. Let $\varepsilon > 0$. Denote*

$$X_n^{\varepsilon} \;=\; X_n \mathbf{1}_{\{|X_n| \leq \varepsilon\}}.$$

Then the series $\sum_{n=1}^{\infty} X_n$ converges a.s. if and only if the series

$$\sum_{n=1}^{\infty} \mathbf{E}[X_n^{\varepsilon}], \quad \sum_{n=1}^{\infty} \mathbf{Var}[X_n^{\varepsilon}], \quad \textit{and} \quad \sum_{n=1}^{\infty} \mathbf{P}[X_n \geq \varepsilon]$$

converge for some (and thus for all) $\varepsilon > 0$.

Proof. This is a good exercise for a master's thesis. □

8.2 Large Deviations Principle

The large deviations principle investigates the rate of convergence in the (weak) law of large numbers. Indeed, let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent random variables that satisfy the weak law of large numbers

$$\frac{1}{n}S_n \;\overset{\mathbf{P}}{\to}\; \mu,$$

where $S_n := \sum_{k=1}^n X_k$. In other words

$$\mathbf{P}\left[\left|\frac{1}{n}S_n - \mu\right| \geq \varepsilon\right] \quad \to \quad 0.$$

The *large deviations principle* states that (under suitable assumptions) the convergence above is exponentially fast:

$$\mathbf{P}\left[\left|\frac{1}{n}S_n - \mu\right| \geq \varepsilon\right] \quad \simeq \quad \mathrm{e}^{-nI(\varepsilon)}.$$

Here I is the *rate function* associated with the large deviations principle. It is a positive function with minimum $I(0) = 0$.

The large deviations principle is a collection of many results. In these notes we only consider the so.-called *Cramér's theorem* where the rate function I is the *Fenchel's and Legendre's transform* Λ^* of the *cumulative generating function* Λ of the common distribution of the summands X_n, $n \in \mathbb{N}$.

8.2.1 Theorem. *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent identically distributed random variables satisfying the* Cramér's condition*:*

$$\Lambda(\theta) \quad := \quad \ln \mathbf{E}\left[\mathrm{e}^{\mathrm{i}\theta X_n}\right] \quad < \quad \infty$$

for all $\theta \in \mathbb{R}$. Denote $\mu := \mathbf{E}[X_n]$ and $S_n := \sum_{k=1}^n X_k$. Then for all $x > \mu$ we have

$$\lim_{n\to\infty} \frac{1}{n} \ln \mathbf{P}\left[\frac{1}{n}S_n \geq x\right] \quad = \quad -\Lambda^*(x).$$

and for all $x < \mu$ we have

$$\lim_{n\to\infty} \frac{1}{n} \ln \mathbf{P}\left[\frac{1}{n}S_n \leq x\right] \quad = \quad -\Lambda^*(x),$$

where

$$\Lambda^*(x) \quad := \quad \sup_{\theta\in\mathbb{R}^d} \left\{\theta x - \Lambda(\theta)\right\}.$$

Proof. This is an excellent topic for a master's thesis. □

Cramér worked as a consultant for a Swedish insurance company. Therefore it is natural to apply the Cramér's theorem 8.2.1 in an insurance setting.

8.2.2 Example. Suppose an insurance company that gains daily amount p of premiums. Each day t the insurance company receives X_t amount of claims. We assume that the claims X_t, $t \in \mathbb{N}$, are independent and identically distributed with common finite cumulant generating function

$$\Lambda(\theta) \;:=\; \ln \mathbf{E}\left[\mathrm{e}^{\theta X_t}\right].$$

The problem of the insurance company is to define the premium p so that the probability of ruin (before a large finite time T, or eventually) is as small as possible.

Now the Cramér's theorem 8.2.1 states that

$$\mathbf{P}\left[\sum_{t=1}^{T} X_t > pT\right] \;\simeq\; \mathrm{e}^{-T\Lambda^*(p)}.$$

So, we have to solve p from the equation

$$\Lambda^*(p) \;=\; -\frac{\ln \varepsilon}{T},$$

where ε is the fixed level or risk. To solve this we of course need to know the function Λ^*, but it can be estimated statistically from the realized claims.

8.3 Central Limit Theorem

Recall that a random variable G is *normally distributed* with parameters μ and σ^2 if its distribution can be written as

$$\mathbf{P}_G[B] \;=\; \int_B \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \mathrm{d}x$$

The next result is a version of the central limit theorem in classical simple form.

8.3.1 Theorem. *Let $(X_n)_{n\in\mathbb{N}} \subset L^2$ be a sequence of independent identically distributed randonm variables. Denote $\mu := \mathbf{E}[X_n]$ and $\sigma^2 := \mathbf{Var}[X_n]$. Then*

$$\frac{1}{\sqrt{n}} \sum_{k=1}^{n} \frac{(X_k - \mu)}{\sigma} \;\xrightarrow{d}\; G,$$

where G is normally distributed with parameters 0 and 1.

The proof of the central limit theorem 8.3.1 we present below is based on the following three facts:

(i) For the characteristic function ϕ_X of a random variable $X \in L^2$ we have

$$\begin{aligned} \phi_X(u) &\simeq \phi_X(0) + \frac{\mathrm{d}}{\mathrm{d}u}\phi_X(0)u + \frac{1}{2}\frac{\mathrm{d}^2}{\mathrm{d}u^2}\phi_X(0)u^2 \\ &= 1 + \mathrm{i}\mathbf{E}[X]u - \frac{1}{2}\mathbf{E}[X^2]u^2. \end{aligned}$$

(ii) If $X_1, \ldots, X_n$ are independent identically distributed random variables, then

$$\begin{aligned} \phi_{X_1+\cdots+X_n}(u) &= \phi_{X_1}(u)\cdots\phi_{X_n}(u) \\ &= \phi_{X_1}(u)^n. \end{aligned}$$

(iii) Let $S_n = \sum_{k=1}^n (X_k - \mu)/\sigma$. We have the converge $S_n/\sqrt{n} \xrightarrow{\mathbf{P}} G$, if $\phi_{S_n/\sqrt{n}} \to \phi_G$ pointwise. Here

$$\phi_G(u) = \mathrm{e}^{-\frac{1}{2}u^2}.$$

Proof of Theorem 8.3.1. Let ϕ be the characteristic function of the random variable $X_n - \mu$. Since $\mathbf{E}[X_n - \mu] = 0$ ans $\mathbf{E}[(X_n - \mu)^2] = \sigma^2$, we have by Theorem 6.3.6(ii) that

$$\phi(u) = 1 - \frac{\sigma^2}{2}u^2 + u^2\varepsilon(u),$$

where $\varepsilon(u) \to 0$, as $u \to 0$. Then the characteristic function of the partial sum

$$\sum_{k=1}^n (X_k - \mu)$$

is, by indenendence, ϕ^n. For the normed partial sum

$$\frac{1}{\sigma\sqrt{n}}\sum_{k=1}^n (X_k - \mu)$$

we have the characteristic function

$$\begin{aligned} \phi_n(u) &= \phi\left(\frac{u}{\sqrt{n}\sigma}\right)^n \\ &= \mathrm{e}^{n\ln\phi\left(\frac{u}{\sigma\sqrt{n}}\right)} \\ &= \mathrm{e}^{n\ln\left(1-\frac{u^2}{2n}+\frac{u^2}{n\sigma^2}\varepsilon\left(\frac{u}{\sigma\sqrt{n}}\right)\right)}. \end{aligned}$$

Now, for a fixed $u \in \mathbb{R}$ we have, by Exercise 8.8, that

$$\lim_{n\to\infty} \left(n \ln \left(1 - \frac{u^2}{2n} + \frac{u^2}{n\sigma^2} \varepsilon \left(\frac{u}{\sigma\sqrt{n}} \right) \right) \right) \;=\; -\frac{1}{2}u^2.$$

Consequently,

$$\lim_{n\to\infty} \phi_n(u) \;=\; \mathrm{e}^{-\frac{1}{2}u^2}.$$

But this is the characteristic function of a standard normal distribution. Hence, the claim follows from the Lévy's continuity theorem 7.4.7(ii). □

The classical application of the central limit theorem is the so-called *normal approximation*. The next example is precisely that.

8.3.2 Example. Suppose an electronic component with mean lifespan a. Suppose the *standard deviation* (square root of the variance) of the lifespan is also a. Our task is to define how many independent electronic components are needed in order to have total lifespan S less that $8a$ with probability not exceeding 5%.

A sure inequality can be attained by using Markov's inequality. We derive an approximate inequality by using the central limit theorem. We note that the normalized random variable

$$\frac{S - na}{a\sqrt{n}}$$

is approximately standard normal distributed, i.e. normally distributed with parameters $\mu = 0$ and $\sigma^2 = 1$. Consequently,

$$\mathbf{P}[S \le 8a] \;=\; \mathbf{P}\left[\frac{S - na}{a\sqrt{n}} \le \frac{8 - n}{\sqrt{n}} \right] \;\simeq\; \Phi\left(\frac{8 - n}{\sqrt{n}} \right),$$

where

$$\Phi(x) \;:=\; \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \mathrm{e}^{-\frac{1}{2}y^2} \, \mathrm{d}y$$

is the cumulative distribution function of the standard normal distribution. We obtain

$$\Phi\left(\frac{8 - n}{\sqrt{n}} \right) \;\le\; 0.05.$$

In other words

$$\frac{8 - n}{\sqrt{n}} \;\le\; \Phi^{-1}(0.05).$$

Now by using a table (or computer) to calculate the inverse $\Phi^{-1}(0.05)$, we obtain $n \ge 15$.

We and this booklet by stating the so-called *Lindeberg's and Feller's central limit theorem*. It gives, in a certain sense, necessary and sufficient conditions for the central convergence.

8.3.3 Theorem. *Let $\xi_{n,l}$, $n \in \mathbb{N}$, $l \leq k_n$ be an array of square integrable random variables. We assume that $\xi_{n,1}, \dots, \xi_{n,k_n}$ are independent for all $n \in \mathbb{N}$, $\lim_{n\to\infty} k_n = \infty$, and*

$$\begin{aligned} \mathbf{E}[\xi_{n,l}] &= 0, \\ \sum_{l=1}^{k_n} \mathbf{Var}[\xi_{n,l}] &= 1. \end{aligned}$$

If the random variables $\xi_{n,l}$, $n \in \mathbb{N}$, $l \leq k_n$ satisfy the Lindeberg's condition

$$\lim_{n\to\infty} \sum_{l=1}^{k_n} \mathbf{E}\left[\xi_{n,l}^2 \mathbf{1}_{\{|\xi_{n,l}| \geq \varepsilon\}}\right] = 0 \tag{8.3.4}$$

for all $\varepsilon > 0$, then

$$\sum_{l=1}^{k_n} \xi_{n,l} \xrightarrow{d} G, \tag{8.3.5}$$

whwre G is standard normal distributed.

Moreover, if

$$\lim_{n\to\infty} \max_{l \leq k_n} \mathbf{P}[\xi_{n,l} \geq \varepsilon] = 0$$

for all $\varepsilon > 0$, the condition (8.3.4) is necessary for the convergence (8.3.5).

Proof. This is a good exercise for a master's thesis. □

Exercises

8.1. Let $X_1, \dots, X_n$ random variables such that $X_k^2 \in L^1$ for all $k \leq n$. Show that

(a)
$$\mathbf{Var}\left[\sum_{k=1}^{n} X_k\right] = \sum_{k=1}^{n} \mathbf{Var}[X_k] + 2\sum_{k=1}^{n} \sum_{l=k+1}^{n} \mathbf{Cov}[X_k, X_l].$$

(b) if the random variables $X_1, \ldots, X_n$ are uncorrelated, then

$$\mathbf{Var}\left[\sum_{k=1}^{n} X_k\right] = \sum_{k=1}^{n} \mathbf{Var}[X_k].$$

8.2. Let $X, Y \in L^2$. Show that

(a) $|\mathbf{Cor}[X,Y]| \leq 1$,

(b) if $\mathbf{Cor}[X,Y] = 1$, then there exist $a, b \in \mathbb{R}$ such that $Y \overset{a.s.}{=} aX + b$.

8.3. Let $(X_n)_{n\in\mathbb{N}} \subset L^2$ be a sequence of uncorrelated and identically distributed random variables. Show that

$$\frac{1}{n}\sum_{k=1}^{n} X_k \overset{a.s.}{\to} \mathbf{E}[X_1].$$

8.4. Prove the Skorohod's maximal inequality 8.1.4.

8.5. Prove Corollary 8.1.7.

8.6. Prove the Kronecker's lemma 8.1.8.

8.7. Show that in Theorem 8.1.10 the series diverges a.s. if $\mathbf{E}[|X_1|] = \infty$.

8.8. Show the claim

$$\lim_{n\to\infty}\left(n \ln\left(1 - \frac{u^2}{2n} + \frac{u^2}{n\sigma^2}\varepsilon\left(\frac{u}{\sigma\sqrt{n}}\right)\right)\right) = -\frac{1}{2}u^2,$$

which was used in the proof of the central limit theorem 8.3.1.

Index

$(\Omega, \mathcal{F}, \mathbf{P})$, probability space, 14
L^p, 81
$\mathbf{1}_A$, indicator, 7
$\mathcal{B}$, $\mathbb{R}$'s Borel's σ-algebra, 13
Cor, correlation, 89
$\mathcal{E}$, simple random variables, 27
E, expectation, 30
L^1, integrable random variables, 30
L^p convergence, 81
Ω, sample space, 5
$\mathbf{P}_1 \otimes \mathbf{P}_2$, product measure, 53
Φ, cumulative distribution function of standard normal law, 102
$\bar{\mathbb{R}}$, extended reals, 6
Var, variance, 41
δ_{x_0}, Dirac's point mass at x_0, 25
ℓ, Lebesgue's measure, 19
lim inf, 6
lim sup, 6
$\mathcal{C}$, cylinder sets, 55
π-class, 16
$\pi(\mathcal{C})$, 17
pow (E), set of all subsets of E, 5
σ-additivity, 15
$\sigma(X)$, σ-algebra generated by X, 46
$\sigma(\mathcal{C})$, σ-algebra generated by $\mathcal{C}$, 12
$\xrightarrow{L^p}$, L^p convergence, 81
$\xrightarrow{\mathbf{P}}$, convergence in probability, 82
$\xrightarrow{a.s.}$, almost sure convergence, 78
$\xrightarrow{d}$, convergence in distribution, 84
φ, density function of standard normal law, 45
$a(\mathcal{C})$, 17
d-class, 16
$d(\mathcal{C})$, 17

a.s., 36
additivity, 15
algebra, 16
almost sure, 36
almost sure convergence, 78
argument, 63

Bernoulli distribution, 69
Bernstein's polynomial, 91
Borel measurable, 23
Borel's σ-algebra, 13
Borel's and Cantelli's lemma, 48
Borel's set, 13

Carathéodory's extension theorem, 19
Cartesian product, 6
Cartesian representation, 64
Cauchy distribution, 45
Cauchy's product formula, 65
central limit theorem, 8, 100
change-of-variables, 37
characteristic function, 68
Chebyshev's inequality, 41
Chernoff's inequality, 41
complement, 5
completion of a probability space, 19
complex conjugate, 63
complex exponential function, 65
complex number, 63

continuous random variable, 26
convergence in distribution, 84
convergence in probability, 82
convex function, 39
convolution, 56
correlation, 89
countable, 5
covariance, 89
Cramér's condition, 41, 99
Cramér's theorem, 99
cumulant generating function, 41
cumulative distribution function, 26
cumulative generating function, 99
cylinder set, 14, 55

De Morgan's laws, 6
Dedekind completeness, 5
density function, 26
Dirac's point mass, 25
discrete random variable, 26
distribution, 25
Dynkin's extension theorem, 18
Dynkin's lemma, 17

empirical cumulative distribution function, 98
Euler's formula, 65
ev., 7
event, 13
expectation, 28, 30
exponential distribution, 77

Fatou's lemma, 35
Fenchel's and Legendre's transform, 42, 99
Fubini's theorem, 54

Hölders inequality, 40
hamburger principle, 36

i.o., 7
imaginary part, 63
independence, 46
indicator, 7
infimum, 5
integrability, 30
inverse, 64

Jensen's inequality, 39
joint characteristic function, 76
joint probability distribution function, 47

Kolmogorov's 0–1 law, 51
Kolmogorov's axioms, 14
Kolmogorov's strong law of large numbers, 95
Kolmogorov's three series theorem, 98
Kronecker's lemma, 94

Lévy's continuity theorem, 86
large deviations limit, 8
large deviations principle, 99
law of large numbers, 8
Lebesgue's dominated convergence theorem, 35
Lebesgue's measure, 19
Lévy's inversion formula, 73
Lindeberg's and Feller's central limit theorem, 103
Lindeberg's condition, 103

Markov's inequality, 41
measurable function, 22
measurable rectrangle, 52
measurable set, 13
measurable space, 13
measure, 14
Minkowski's inequality, 40
modulus, 63
monotone class theorem, 17
monotone cleass, 16
monotone convergence theorem, 35

Monte Carlo integration, 97
most beautiful formula of mathematics, 66

Newton's binomial formula, 43
normal approximation, 102
normal distribution, 45, 100
null set, 19

open set, 13

pairwise independence, 46
Poisson distribution, 45
polar reprentation, 64
probability measure, 14
probability space, 14
product σ-algebra, 52
product measure, 53
product probability space, 56
product space, 52

random variable, 22
range, 28
rate function, 99
real part, 63
relative frequency, 7

Schwarz's inequality, 40
σ-algebra, 11
simple random variable, 27
Skorohod's maximal inequality, 93
slice, 52
square integrable random variable, 89
standard deviation, 102
Stirling's formula, 8
strong law of large numbers, 9, 94
subtraction, 5
supremum, 5

tail σ-algebra, 51
tangent, 39
triangle inequality, 41

uncorrelated random variables, 89
uncountable, 5
uniform continuity, 92

variance, 41
vector algebra, 28

weak law of large numbers, 9, 90
Weierstraß's approximation theorem, 91