Probability Theory
FS 2020

Johanna F. Ziegel
Translated by Alexander Henzi

Supplementary literature:

J. F. Ziegel, *Masstheorie*, Skriptum, Universität Bern, 2020.

H. Bauer, *Mass- und Integrationstheorie*, de Gruyter, 1990.

H. Bauer, *Wahrscheinlichkeitstheorie*, de Gruyter, 1991.

A. Klenke, *Probability Theory: A Comprehensive Course*, Springer, 2008.

Contents

Chapter 1

Basic concepts of probability theory

1.1 Probability spaces

Definition 1.1. Let Ω be a set and $\mathfrak{F}$ a σ-field on Ω. A measure $\mathbb{P}$ on $\mathfrak{F}$ with $\mathbb{P}(\Omega) = 1$ is called *probability measure*. The triplet $(\Omega, \mathfrak{F}, \mathbb{P})$ is called a *probability space*.

The notion of probability spaces dates back to the Russian mathematician Andrey Kolmogorov, 1903–1987. By introducing probability spaces in 1933, he laid the foundation of modern probability theory.

Example 1.2 (Tossing dice)**.**

1. We would like to define a probability space which models the result of tossing a die. The sample space is $\Omega = \{1, \ldots, 6\}$. We use the σ-field $\mathfrak{F} = 2^\Omega$, i.e. the family of all subsets of Ω, and we assume that each number is rolled with equal probability, that is,
$$\mathbb{P}(\{\omega\}) = \frac{1}{6}, \quad \omega \in \Omega.$$
This already defines $\mathbb{P}$ uniquely. It holds
$$\mathbb{P}(A) = \frac{\#A}{6}, \quad A \subseteq \{1, \ldots, 6\}.$$
Subsets $A \subseteq \Omega$ describe *events*. For example, $A = \{2, 4, 6\}$ is the event of rolling an even number.

2. We now toss two dice. So we select $\Omega = \{1, \ldots, 6\} \times \{1, \ldots, 6\} = \{(i, j) \mid i, j \in \{1, \ldots, 6\}\}$, $\mathfrak{F} = 2^\Omega$ and we define $\mathbb{P}$ by
$$\mathbb{P}(\{\omega\}) = \frac{1}{36}, \quad \omega \in \Omega.$$
In this case we can observe more interesting events, such as the sum of both numbers rolled being equal to an even number. We will return to this example later.

Both examples are *discrete* probability spaces (Ω is at most countable). You have seen many more examples of discrete probability spaces in the combinatorics and probability lecture.

Example 1.3 (Wheel of fortune)**.** We consider a wheel of fortune with the numbers $[0, 2\pi)$ marked on its border. We want to describe the following random experiment: We push the wheel of fortune, and it stops at a number $\omega \in [0, 2\pi)$. Hence the sample space is $\Omega = [0, 2\pi)$. We assume that all outcomes are equally likely, so the probability of any interval $[a, b] \subseteq [0, 2\pi)$ should be proportional to its length. We have seen that the Lebesgue-Measure satisfies exactly this property. We therefore choose the Borel-σ-field[1] $\mathfrak{F} = \mathcal{B}([0, 2\pi))$ on $[0, 2\pi)$ and $\mathbb{P} = \frac{1}{2\pi}\mathcal{L}^1|_{[0,2\pi)}$ as the (rescaled) Lebesgue-measure on $[0, 2\pi)$.

In probability theory the following expressions are used:

- Ω is called *sample space*, *Stichprobenraum* oder *Ergebnisraum*,
- the elements $\omega \in \Omega$ denote outcomes,
- the sets $A \in \mathfrak{F}$ are called *events*,
- $\mathbb{P}(A)$ is the probability of A.
- An event A occurs $\mathbb{P}$ almost surely, if $\mathbb{P}(A) = 1$.

Certain operations on sets can now be interpreted, for example:

- $A^c = \Omega \setminus A$: "‘A does not occur"’,
- $\cup A_i$: "‘at least one of A_i happens"’,
- $\cap A_i$: "‘all of A_i happen"’.

1.2 Random variables and expected values

Let $(\Omega, \mathfrak{F}, \mathbb{P})$ be a probability space.

Definition 1.4. A $\mathfrak{F}$-$\mathcal{B}(\mathbb{R}^d)$ measurable mapping $X = (X_1, \ldots, X_d)\colon \Omega \to \mathbb{R}^d$ is called *random vector*. If $d = 1$, we also call X *random variable*.

Since the Borel-σ-field is generated by all cuboids, it is sufficient to require

$$\Big\{\omega \mid X(\omega) \in \prod_{i=1}^{d}(a_i, b_i]\Big\} \in \mathfrak{F}, \quad \text{for all } a_1 \leq b_1, \ldots, a_d \leq b_d.$$

This is also equivalent to

$$\Big\{\omega \mid X(\omega) \in \prod_{i=1}^{d}(-\infty, t_i]\Big\} \in \mathfrak{F}, \quad \text{for all } t_1, \ldots, t_d \in \mathbb{R}.$$

Exercise 1.5. Show that $X = (X_1, \ldots, X_d)$ is a random vector if and only if $X_1, \ldots, X_d$ are random variables.

Remark. If a random variable X is a non-negative stepfunction, then it is also called *simple* random variable.

[1] The Borel-σ-field $\mathcal{B}(A)$ on a set $A \in \mathcal{B}(\mathbb{R})$ is defined as $\mathcal{B}(A) := \{A \cap B \mid B \in \mathcal{B}(\mathbb{R})\}$. The measure $\mathcal{L}^1|_A$ is the Lebesgue measure restricted to $\mathcal{B}(A)$.

Definition 1.6. The image measure $\mathbf{P}_X$ of a random vector $X = (X_1, \dots, X_d)$ is a probability measure on $\mathcal{B}(\mathbb{R}^d)$ and it is called the *distribution of* X. It is given by

$$\mathbf{P}_X : \mathcal{B}(\mathbb{R}^d) \to [0,1], \quad A \mapsto \mathbf{P}_X(A) := \mathbb{P}(X^{-1}(A)) = \mathbb{P}(X \in A) = \mathbb{P}\big(\{\omega \in \Omega \mid X(\omega) \in A\}\big).$$

The distributions $\mathbf{P}_{X_1}, \dots, \mathbf{P}_{X_d}$ of the components $X_1, \dots, X_d$ of X are called *marginal distributions*. If X has distribution $\mathbf{P}_X$, we write $X \sim \mathbf{P}_X$.

If we know the distribution of a random vector, we can compute the probability of all events of the type $X^{-1}(A) = \{X \in A\}$ for $A \in \mathcal{B}(\mathbb{R}^d)$.

Example 1.7 (Continuation of example 1.2)**.** We consider again the toss of two dice, that is, $\Omega = \{1, \dots, 6\} \times \{1, \dots, 6\}$, $\mathfrak{F} = 2^\Omega$ and $\mathbb{P}(\{\omega\}) = \frac{1}{36}$, $\omega \in \Omega$. We set

$$X : \Omega \to \mathbb{R}, \quad (i,j) \mapsto i - j.$$

The random variable X gives the difference of the number rolled with the first die and the number rolled with the second one. The possible values of this difference are $-5, -4, \dots, 0, \dots, 4, 5$. For $k \in \{-5, \dots, 5\}$ we can compute $p_k := \mathbb{P}(X = k)$. For example, $p_0 = \mathbb{P}(X = 0) = 6/36 = 1/6$ and $p_{-5} = \mathbb{P}(X = -5) = 1/36$. The distribution of X is given by

$$\mathbf{P}_X(A) = \mathbb{P}(X \in A) = \sum_{k \in A \cap \{-5,\dots,5\}} p_k = \sum_{k=-5}^{5} p_k \delta_k(A),$$

where δ_k is the Dirac measure; see the lecture notes on measure theory (examples 1.17 and 1.13).

Definition 1.8. A random vector $X : \Omega \to \mathbb{R}^d$ is called

- *absolutely continuous*, if its distribution $\mathbf{P}_X$ admits a density with respect to the Lebesgue measure $\mathcal{L}^d$.
- *singular*, if there exists a Lebesgue null set N with $\mathbb{P}(X \in N) = 1$.
- *discrete*, if there exists a countable set M with $\mathbb{P}(X \in M) = 1$.

In particular, discrete random vectors are singular.

Remark 1.9. The considerations of Example 1.7 hold under more general conditions. The distribution of a discrete random vector with values in $(x_k)_{k \in \mathbb{N}}$ and probabilities $p_k = \mathbb{P}(X = x_k)$ is given by

$$\mathbf{P}_X = \sum_{k \in \mathbb{N}} p_k \delta_{x_k}.$$

Example 1.10 (Uniform distribution)**.** Let $A \in \mathcal{B}(\mathbb{R}^d)$ be a Borel set with Lebesgue measure $0 < \mathcal{L}^d(A) < \infty$. A random vector X is called *uniformly distributed on* A, if its distribution admits the density

$$f : \mathbb{R}^d \to [0,1), x \mapsto f(x) = \frac{1}{\mathcal{L}^d(A)} \mathbb{1}_A(x).$$

Exercise 1.11. Are there any random variables that are at the same time discrete and absolutely continuous? Find an example or prove that this is not possible.

Definition 1.12. The *expected value* (or *expectation*) $\mathbf{E}(X)$ of a random variable $X\colon \Omega \to \mathbb{R}$ is defined as the integral

$$\mathbf{E}(X) := \int X d\mathbb{P},$$

if it exists. If $\mathbf{E}(|X|) < \infty$, then X is called integrable. The *k-th moment* of X is the expected value $\mathbf{E}(X^k)$, if it exists. If $\mathbf{E}(|X|^2) < \infty$, then X is called *square integrable.* The *expected value* of a random vector $X = (X_1, \dots, X_d)\colon \Omega \to \mathbb{R}^d$ is defined as $\mathbf{E}(X) = (\mathbf{E}(X_1), \dots, \mathbf{E}(X_d))$, if the expectations $\mathbf{E}(X_i)$, $i = 1, \dots, d$ all exist.

Theorem 1.13 (Lyapunoff's inequality)**.** *Let* $0 < s < t$ *and let* X *be a random variable. Then,*

$$(\mathbf{E}(|X|^s))^{1/s} \le (\mathbf{E}(|X|^t))^{1/t}.$$

In particular, X^s *is integrable for all* $0 < s < t$ *if* X^t *is integrable.*

Proof. We set $r = t/s > 1$ and $r' = t/(t-s) > 1$. Then $\frac{1}{r} + \frac{1}{r'} = 1$ and Hölder's inequality (lecture notes on measure theory) implies that for $Y = |X|^s$,

$$\mathbf{E}(|Y \cdot 1|) \le (\mathbf{E}(|Y|^r))^{1/r} (\mathbf{E}(|1|^{r'}))^{1/r'} = (\mathbf{E}(|X|^t))^{s/t},$$

which proves the claim. □

Remark 1.14. Theorem 1.13 shows that square integrable random variables are always integrable. Moreover, if the components of a random vector $X = (X_1, \dots, X_d)$ are square integrable, then $X_i X_j$ is integrable for all $i, j \in \{1, \dots, d\}$.

Exercise 1.15. Prove the second part of Remark 1.14.

According to the theorem on integration with respect to image measures (lecture notes on measure theory), it holds for any measurable function $h : \mathbb{R}^d \to \mathbb{R}$ that

$$\mathbf{E}(h(X)) = \int_\Omega h \circ X \, \mathrm{d}\mathbb{P} = \int_{\mathbb{R}^d} h \, \mathrm{d}\mathbf{P}_X,$$

if the integrals exist. Hence, the expected value of random vector X only depends on its distribution $\mathbf{P}_X$. In order to compute expectations, we therefore only need to know how to integrate with respect to measures on $\mathbb{R}^d$.

Example 1.16 (Continuation of example 1.7)**.** Let the random vector X be defined as in Example 1.7. Its distribution is $\mathbf{P}_X = \sum_{k=-5}^{5} p_k \delta_{x_k}$. According to example 2.16 in the lecture notes on measure theory, we can compute the expected value of X as follows:

$$\mathbf{E}(X) = \int X \, \mathrm{d}\mathbb{P} = \int x \, \mathrm{d}\mathbf{P}_X = \sum_{k=-5}^{5} k p_k.$$

Definition 1.17. For an integrable random variable X, its *variance* is defined as

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

Let $X = (X_1, \dots, X_d)$ be a random vector vector with square integrable components $X_1, \dots, X_d$. The matrix $\Gamma = \Gamma^X = (\Gamma^X_{ij})_{i,j=1,\dots,d}$ with entries

$$\Gamma^X_{ij} = \mathsf{Cov}(X_i, X_j) = \mathbf{E}\big[(X_i - \mathbf{E}(X_i))(X_j - \mathbf{E}(X_j))\big] = \mathbf{E}(X_i X_j) - \mathbf{E}(X_i)\mathbf{E}(X_j)$$

is called the *covariance matrix of* X.

Example 1.18 (Standard normal distribution)**.** Let

$$\varphi\colon \mathbb{R} \to \mathbb{R}, \quad x \mapsto \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

It holds $\int \varphi(x)\,\mathrm{d}x = 1$ (compare combinatorics and probability lecture), hence

$$\mathbf{P}_X\colon \mathcal{B}(\mathbb{R}) \to [0,\infty], \quad A \mapsto \int_A \varphi(x)\,\mathrm{d}x$$

defines the distribution of a random variable X by proposition 2.33 in the lecture notes on measure theory.[2] By applying the theorem on the integration with respect to image measures, we obtain

$$\mathbf{E}(X) = \int x d\mathbf{P}_X = \int x\varphi(x)\,\mathrm{d}x = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x e^{-x^2/2}\,\mathrm{d}x = 0$$

and

$$\mathsf{Var}(X) = \int (x - \mathbf{E}(X))^2 d\mathbf{P}_X = \int x^2 \varphi(x)\,\mathrm{d}x = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 e^{-x^2/2}\,\mathrm{d}x = \cdots = 1.$$

The normal distribution was discovered in the 18th century by de Moivre and Laplace as limiting distribution of the Binomial distribution. It is also called Gaussian distribution, after Karl Friedrich Gauss, who investigated this distribution.

> The normal distribution is of central importance in probability theory and will appear throughout this course. It is worthwhile to learn the corresponding formulae by heart.

Remark 1.19. As with the standard normal distribution, there is the following more general result on the computation of expected values. If a random vector X is absolutely continuous, i.e. its distribution $\mathbf{P}_X$ admits a density f with respect to the Lebesgue measure, then Theorem 2.35 in the lecture notes on measure theory implies that for any measurable function $h : \mathbb{R}^d \to \mathbb{R}$, if $\mathbf{E}(h(X))$, then

$$\mathbf{E}(h(X)) = \int_{\mathbb{R}^d} h(x)\,\mathrm{d}\mathbf{P}_X(x) = \int_{\mathbb{R}^d} h(x) f(x)\,\mathrm{d}x.$$

Exercise 1.20. Let X be an absolutely continuous random variable with density $f = (1/4)\mathbb{1}_{[-1,3]}$. Compute $\mathbf{E}(X^2)$.

Definition 1.21 (Cumulative distribution function)**.** Let $X = (X_1, \ldots, X_d)$ be a random vector. The *cumulative distribution function* (or *distribution function*) F_X of X is given by

$$F_X\colon \mathbb{R}^d \to [0,1], \quad (t_1, \ldots, t_d) \mapsto F_X(t_1, \ldots, t_d) = \mathbb{P}(X_1 \le t_1, \ldots, X_d \le t_d).$$

The distribution functions $F_{X_1}, \ldots, F_{X_d}$ are called *marginal distribution functions.*

[2]To prove the existence of the random variable X, we can choose $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathbf{P}_X)$ as probability space and set $X = \mathrm{id}$.

The distribution of a random vector X is a probability measure $\mathbf{P}_X \colon \mathcal{B}(\mathbb{R}^d) \to [0,1]$, while its distribution function is a mapping $F_X \colon \mathbb{R}^d \to [0,1]$

Remark 1.22 (Discrete distributions). A discrete random variable X takes values in a countable set $(x_k)_{k\in\mathbb{N}}$, that is

$$F_X(t) = \sum_{x_k \le t} \mathbb{P}(X = x_k).$$

Usually, one rather describes the summands $p_k = \mathbb{P}(X = x_k)$ instead of F_X itself; see Example 1.7.

Example 1.23 (Poisson distribution)**.** A random variable X with

$$\mathbb{P}(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}, \quad k = 0, 1, \ldots$$

is called *Poisson distributed.* We write $X \sim \mathrm{POIS}(\lambda)$. This is a distribution for rare events (accidents, radioactive decay, and so on).

It holds

$$\mathbf{E}(X) = \sum_{i=0}^{\infty}(e^{-\lambda}\frac{\lambda^i}{i!})i = e^{-\lambda}\sum_{i=1}^{\infty}\frac{\lambda^i}{(i-1)!} = e^{-\lambda}\lambda e^{\lambda} = \lambda.$$

For the second moment, we obtain

$$\begin{aligned}\mathbf{E}(X^2) &= e^{-\lambda}\sum_{i=1}^{\infty}\frac{\lambda^i}{i!}i^2 = e^{-\lambda}\sum_{i=1}^{\infty}\frac{\lambda^i}{i!}[i(i-1)+i] \\ &= e^{-\lambda}\left[\lambda^2\sum_{j=0}^{\infty}\frac{\lambda^j}{j!} + \lambda\sum_{j=0}^{\infty}\frac{\lambda^j}{j!}\right] = e^{-\lambda}(\lambda^2+\lambda)e^{\lambda} = \lambda^2 + \lambda.\end{aligned}$$

This implies

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

Remark 1.24 (Absolutely continuous distributions). Let $\mathbf{P}_X$ be the distribution of a random vector X with density f with respect to the Lebesgue measure $\mathcal{L}^d$; see also Example 1.18. Then the distribution function F_X satisfies

$$F_X(t_1, \ldots, t_d) = \int_{-\infty}^{t_d} \cdots \int_{-\infty}^{t_1} f(y_1, \ldots, y_d)\,\mathrm{d}y_1 \cdots \mathrm{d}y_d.$$

Therefore, a measurable non-negative function $f \colon \mathbb{R}^d \to [0, \infty]$ is the density of a distribution on $\mathbb{R}^d$, if

$$\int_{\mathbb{R}^d} f(x)\,\mathrm{d}x = 1.$$

If f is continuous at x, then

$$f(x) = \frac{\partial^d F_X}{\partial x_1 \cdots \partial x_d}(x).$$

In that way, densities can be computed from distribution functions.

Example 1.25 (Continuation of example 1.18, normal distribution)**.** The distribution function of the standard normal distribution is

$$\Phi(t) = \int_{-\infty}^{t} \varphi(x)\,\mathrm{d}x = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-x^2/2}\,\mathrm{d}x.$$

Φ is a special function and it is tabulated. It holds $\Phi(-t) = 1 - \Phi(t)$.

We define the normal distribution with parameters $\mu \in \mathbb{R}$ and $\sigma > 0$ as the distribution with density

$$f(x) = \frac{1}{\sigma}\varphi\Big(\frac{x-\mu}{\sigma}\Big) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}.$$

The corresponding distribution function F is given by

$$F(t) = \Phi\Big(\frac{x-\mu}{\sigma}\Big).$$

Let the random variable X admit this distribution. One can verify that

$$\mathbf{E}(X) = \mu, \quad \mathsf{Var}(X) = \sigma^2,$$

so the expected value of X is μ, and the variance equals σ^2. We denote $X \sim \mathcal{N}(\mu, \sigma^2)$.

Exercise 1.26. Let the random variable X be absolutely continuous with density $f(x) = ce^{-|x-a|/b}$, where $a \in \mathbb{R}$ and $b > 0$. Determine c and the distribution function of X.

Proposition 1.27. *Let X be a random vector with distribution function F_X. Then,*

1. *F_X is rectangular monotone;*
2. $\lim_{x_i \to -\infty} F_X(x_1, \dots, x_d) = 0$ *for all* $i \in \{1, \dots, d\}$, $\lim_{\min_i x_i \to \infty} F_X(x) = 1$*;*
3. *F_X is continuous from the right in each of its components, that is, for all $x = (x_1, \dots, x_d) \in \mathbb{R}^d$ and $i \in \{1, \dots, d\}$ it holds* $\lim_{s \downarrow x_i} F_X(x_1, \dots, x_{i-1}, s, x_{i+1}, \dots, x_d) = F_X(x)$.

Proof. Exercise. □

Remark 1.28. For a random variable X and its distribution function F_X, the properties in Proposition 1.27 simplify to:

1. F_X is monotone increasing;
2. $\lim_{t \to -\infty} F_X(t) = 0$, $\lim_{t \to \infty} F_X(t) = 1$;
3. F_X is continuous from the right.

In general the problem is the reverse: Given F_X, we would like to find a suitable probability space realizing F_X. Luckily, the reverse statement of Proposition 1.27 is true as well.

Theorem 1.29. *Let $F\colon \mathbb{R}^d \to [0,1]$ be a function satisfying 1.-3. from Proposition 1.27. Then there exists a unique probability measure $\mathbb{P}\colon \mathcal{B}(\mathbb{R}^d) \to [0,1]$ with $\mathbb{P}((-\infty, x_1] \times \dots \times (-\infty, x_d]) = F(x)$ for all $x = (x_1, \dots, x_d) \in \mathbb{R}^d$.*

Proof. Since F is rectangular monotone, Theorem 1.24 in the lecture notes on measure theory implies that there exists a unique content $\tilde{\mathbb{P}}$ on $\mathfrak{R}^d$ such that for any cuboid $(\mathbf{a}^1, \mathbf{a}^2] \neq \emptyset$, it holds

$$\tilde{\mathbb{P}}((\mathbf{a}_1, \mathbf{a}_2]) = \Delta_{\mathbf{a}^1}^{\mathbf{a}^2} F.$$

In particular, $\tilde{\mathbb{P}}(A) < \infty$ for all $A \in \mathfrak{R}^d$. Together with Theorem 1.28 in the lecture notes on measure theory, property 3. implies that $\tilde{\mathbb{P}}$ is σ additive. Due to Carathéodory 's extension Theorem, there exists a measure $\mathbb{P}$ on $\sigma(\mathfrak{R}^d) = \mathcal{B}(\mathbb{R}^d)$ that extends $\tilde{\mathbb{P}}$. We obtain that for each $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$,

$$\begin{aligned}
\mathbb{P}((-\infty, x_1] \times \cdots \times (-\infty, x_d]) &= \mathbb{P}\Big(\bigcup_{n\in\mathbb{N}} (-n, x_1] \times \cdots \times (-n, x_d]\Big) \\
&= \lim_{n\to\infty} \mathbb{P}\big((-n, x_1] \times \cdots \times (-n, x_d]\big) \\
&= F(x) + \lim_{n\to\infty} \sum_{\substack{i_1,\ldots,i_d\in\{1,2\} \\ (i_1,\ldots,i_d)\neq(2,\ldots,2)}} (-1)^{i_1+\cdots+i_d} F(a_1^{i_1}, \ldots, a_d^{i_d}) = F(x),
\end{aligned}$$

using assumption 2. with $\mathbf{a}^1 = (-n, \ldots, -n)$ and $\mathbf{a}^2 = x$. Assumption 2. also implies $\mathbb{P}(\mathbb{R}^d) = \lim_{n\to\infty} F(n, \ldots, n) = 1$, hence each $\mathbb{P}$ is a probability measure. Since $\mathfrak{R}^d$ is a ring, Theorem 1.35 in the lecture notes on measure theory implies that $\mathbb{P}$ is unique. □

Example 1.30. If a random variable X is uniformly distributed on $[0, 1]$, then its distribution admits the density

$$f(x) = \mathbb{1}_{[0,1]}(x), \quad x \in \mathbb{R},$$

and we obtain the distribution function

$$F_X(t) = \mathbb{1}_{[0,1]}(t)\, t + \mathbb{1}_{(1,\infty)}(t), \quad t \in \mathbb{R}.$$

More generally, a distribution with density $f(x) = 1/(b-a)\mathbb{1}_{[a,b]}(x)$, $x \in \mathbb{R}$, is called *uniform distribution* on $[a, b]$, and we denote $X \sim \mathrm{UNIF}[a, b]$. Let $X \sim \mathrm{UNIF}[a, b]$. Then,

$$\mathbf{E}(X) = \int_a^b \frac{1}{b-a} \cdot x \,\mathrm{d}x = \frac{1}{b-a} \frac{x^2}{2}\Big]_a^b = \frac{a+b}{2}.$$

and

$$\mathbf{E}(X^2) = \frac{1}{b-a} \int_a^b x^2 \,\mathrm{d}x = \frac{b^3 - a^3}{3(b-a)} = \frac{b^2 + ab + a^2}{3}.$$

Thus

$$\mathsf{Var}(X) = \mathbf{E}(X^2) - \left(\frac{a+b}{2}\right)^2 = \frac{b^2 - 2ab + a^2}{12} = \frac{(b-a)^2}{12}.$$

Example 1.31 (Cauchy distribution)**.** The distribution function of the Cauchy distribution is

$$F(x) = \frac{1}{\pi}(\arctan x + \frac{\pi}{2}),$$

hence its density equals

$$f(x) = F'(x) = \frac{1}{\pi(1+x^2)}.$$

Let X be Cauchy distributed. We first compute

$$\mathbf{E}(X_+) = \int_0^\infty x f(x)\,\mathrm{d}x = \frac{1}{\pi}\int_0^\infty \frac{x}{1+x^2}\,\mathrm{d}x = \frac{1}{2\pi}\log(1+x^2)\Big]_0^\infty = \infty.$$

By symmetry it holds $\mathbf{E}(X_-) = \infty$, so the expected value of the Cauchy distribution does not exist.

Example 1.32 (Continuous singular distribution)**.** There exist singular distributions with continuous distribution functions!

We recursively define a sequence $(F_n)_{n\in\mathbb{N}}$ of distribution functions. For all n let $F_n(x) = 0$ for $x \le 0$ and $F_n(x) = 1$ for $x \ge 1$. For $x \in [0,1]$, we set $F_0(x) = x$ and

$$F_{n+1}(x) = \begin{cases} \frac{1}{2}F_n(3x), & \text{falls } x \in [0,1/3], \\ \frac{1}{2}, & \text{falls } x \in [1/3,2/3], \\ \frac{1}{2} + \frac{1}{2}F_n(3x-2), & \text{falls } x \in [2/3,1]. \end{cases}$$

The functions F_n are continuous distribution functions. Moreover, it holds

$$\max_{x\in[0,1]} |F_{n+1}(x) - F_n(x)| \le \frac{1}{2}\max_{x\in[0,1]} |F_n(x) - F_{n-1}(x)|,$$

so the triangle inequality implies that $(F_n)_{n\in\mathbb{N}}$ is a Cauchy sequence in the supremum norm on $[0,1]$. Hence there is a continuous limiting function F, which is also a distribution function. We now show that the corresponding distribution is singular.

We claim that $F_n|_{[0,1]}$ is constant on a length of $(1/3)\sum_{k=0}^{n-1}(2/3)^k$. Moreover, all constant pieces of F_n are also constant pieces of F_{n+1}. This implies that F is constant on a length of

$$\frac{1}{3}\sum_{k=0}^{\infty}\left(\frac{2}{3}\right)^k = \frac{1}{3}\frac{1}{1-\frac{2}{3}} = 1$$

so a random variable with distribution function F only attains values in a set with Lebesgue measure zero.

We now show the claim that $F_n|_{[0,1]}$ is constant on a length of $(1/3)\sum_{k=0}^{n-1}(2/3)^k$ by induction. For $n = 1$, this is obvious. By definition, the length of the constant pieces of $F_{n+1}|_{[0,1]}$ is

$$\frac{1}{3} + \frac{2}{3}\{\text{Length of the constant pieces of } F_n|_{[0,1]}\} = \frac{1}{3} + \frac{2}{3}\frac{1}{3}\sum_{k=0}^{n-1}\left(\frac{2}{3}\right)^k = \frac{1}{3}\sum_{k=0}^{n}\left(\frac{2}{3}\right)^k.$$

1.3 Independence

Let $(\Omega, \mathfrak{F}, \mathbb{P})$ be a probability space and I an arbitrary index set.

Definition 1.33. A family $(A_i)_{i\in I} \subseteq \mathfrak{F}$ of events is called *independent*, if for all $n \in \mathbb{N}$ and all finite subsets $\{i_1, \ldots, i_n\} \subseteq I$,

$$\mathbb{P}(A_{i_1} \cap \cdots \cap A_{i_n}) = \mathbb{P}(A_{i_1}) \cdots \mathbb{P}(A_{i_n}).$$

Example 1.34 (Continuation of example 1.2)**.** We now throw two dice, i.e. $\Omega = \{1, \dots, 6\} \times \{1, \dots, 6\} = \{(i, j) \mid i, j \in \{1, \dots, 6\}\}$, $\mathfrak{F} = 2^\Omega$ and $\mathbb{P}(\{\omega\}) = \frac{1}{36}$, $\omega \in \Omega$. Consider the events

$$A_1 := \text{"`Odd number in the first toss"'},$$
$$A_2 := \text{"`Odd number in the second toss"'},$$
$$A_3 := \text{"`Sum of the two numbers is odd"'}.$$

The set $\{A_1, A_2, A_3\}$ is not independent, but the sets $\{A_1, A_2\}$, $\{A_1, A_3\}$ and $\{A_2, A_3\}$ are all independent.

Pairwise independence does not imply independence!

Example 1.35 (Continuation of example 1.3)**.** We consider $\Omega = [0, 2\pi)$ and $\mathbb{P} = \frac{1}{2\pi}\mathcal{L}^1|_{[0,2\pi)}$. Then $\{[0, \pi], [\pi, 2\pi)\}$ is not independent, but $\{[0, 2\pi), [0, \pi]\}$ is.

We will now extend Definition 1.33.

Definition 1.36. Let $(\mathcal{A}_i)_{i\in I}$ be a family of sets of events $\mathcal{A}_i \subseteq \mathfrak{F}$. The family is called *independent*, fi for any finite subset $\{i_1, \dots, i_n\} \subseteq I$ and any choice of events

$$A_{i_k} \in \mathcal{A}_{i_k}, \quad k = 1, \dots, n$$

the family of events $(A_{i_1}, \dots, A_{i_n})$ is independent.

Exercise 1.37. Let $\mathcal{A}_1, \mathcal{A}_2 \subseteq \mathfrak{F}$ with $\mathcal{A}_1 \subseteq \mathcal{A}_2$. Show that

$$\{\mathcal{A}_1, \mathcal{A}_2\} \text{ is independent} \quad \Longleftrightarrow \quad \mathbb{P}(A) = 0 \text{ oder } \mathbb{P}(A) = 1 \text{ for all } A \in \mathcal{A}_1.$$

Definition 1.38. A family $\{X_i \mid i \in I\}$ of random vectors is called *independent*, if for any finite subset $\{i_1, \dots, i_n\} \subseteq I$ and all Borel sets $A_1, \dots, A_n \in \mathcal{B}(\mathbb{R}^d)$,

$$\mathbb{P}(X_{i_1} \in A_1, \dots, X_{i_n} \in A_n) = \mathbb{P}(X_{i_1} \in A_1) \cdots \mathbb{P}(X_{i_n} \in A_n).$$

Remark 1.39. Verify that $\{X_i \mid i \in I\}$ is independent if and only if $\{\sigma(X_i) \mid i \in I\}$ is independent, where

$$\sigma(X) := \{X^{-1}(A) \mid A \in \mathcal{B}(\mathbb{R}^d)\}$$

is the *σ-field generated by X*.

Let X, Y be two independent random variables on a probability space $(\Omega, \mathfrak{F}, \mathbb{P})$. The corresponding distributions $\mathbf{P}_X$, $\mathbf{P}_Y$ are probability measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Independence means that for all $A, B \in \mathcal{B}(\mathbb{R})$, it holds

$$\mathbf{P}_X(A)\mathbf{P}_Y(B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B) = \mathbb{P}(X \in A, Y \in B) = \mathbf{P}_{(X,Y)}(A \times B), \qquad (1.1)$$

where $\mathbf{P}_{(X,Y)}$ denotes the distribution of the random vector $(X, Y)\colon \Omega \to \mathbb{R}^2$. The unique product measure $\mathbf{P}_X \otimes \mathbf{P}_Y$ on $\mathcal{B}(\mathbb{R}^2)$ hence equals the distribution of the random vector $(X, Y)\colon \Omega \to \mathbb{R}^2$ with independent components X and Y. Product measures therefore guarantee the existence of a probability space with two independent random variables with arbitrary distributions. Since the generation of product measures is associative, this argument can be extended to show the existence of probability spaces with *finitely* many independent random variables (with arbitrary distributions).

Theorem 1.40. *Random variables $X_1, \dots, X_d$ are independent if and only if for all $t_1, \dots, t_d \in \mathbb{R}$,*

$$F_X(t_1, \dots, t_d) = F_{X_1}(t_1) \cdots F_{X_d}(t_d),$$

where $X = (X_1, \dots, X_d)$.

Proof. Exercise. Hint: Use the theorem on the uniqueness of product measures to show that for $d = 2$, $\mathbf{P}_{(X_1, X_2)} = \mathbf{P}_{X_1} \otimes \mathbf{P}_{X_2}$. The proof for general d uses the associativity of the generation of product measures. □

Corollary 1.41. *IF $X_1, \dots, X_d$ are independent absolutely continuous random variables with densities f_{X_i}, $i = 1, \dots, d$, then $X = (X_1, \dots, X_d)$ is absolutely continuous with density*

$$f(x_1, \dots, x_d) = f_{X_1}(x_1) \cdots f_{X_k}(x_d).$$

Example 1.42 (Multivariate normal distribution)**.** If $X_1, \dots, X_d$ are independent and standard normally distributed, then $X = (X_1, \dots, X_d)$ admits the density

$$f_X(x_1, \dots, x_d) = \varphi(x_1)\varphi(x_2) \cdots \varphi(x_d) = \frac{1}{(2\pi)^{d/2}} e^{-\frac{1}{2}(x_1^2 + \cdots + x_d^2)} = \frac{1}{(2\pi)^{d/2}} e^{-\frac{1}{2}\|x\|^2},$$

where $x = (x_1, x_2, \dots, x_d) \in \mathbb{R}^d$. The random vector $X = (X_1, X_2, \dots, X_d)$ admits a *d-dimensional standard normal distribution.* We now consider a linear transformation $Y := AX + b$ with an invertible $(d \times d)$-matrix A and $b \in \mathbb{R}^d$. Let $g : \mathbb{R}^d \to \mathbb{R}^d$ denote the mapping $g(y) = A^{-1}(y - b)$. For an open subset $B \subseteq \mathbb{R}^d$, the transformation theorem (see Analysis lecture) implies that

$$\begin{aligned}
\mathbb{P}(Y \in B) &= \mathbb{P}(AX + b \in B) = \mathbb{P}(X \in g(B)) \\
&= \int_{g(B)} f_X(x) \, \mathrm{d}x = \int_B f_X(g(y)) |J(y)| \, \mathrm{d}y,
\end{aligned}$$

where $J(y) = \det A^{-1}$ is the determinant of the Jacobian matrix of g. We set $\Gamma = AA^\top$. Then Γ is a symmetric, positive definite matrix, $\det \Gamma = (\det A)^2$, and

$$\|A^{-1}y\|^2 = \langle \Gamma^{-1} y, y \rangle,$$

so the distribution of Y admits the density

$$f_Y(y) = \frac{1}{(2\pi)^{d/2}\sqrt{\det \Gamma}} e^{-\frac{1}{2}\langle \Gamma^{-1}(y-b), y-b \rangle}. \tag{1.2}$$

Note that here we use the fact that the open sets generate $\mathcal{B}(\mathbb{R}^d)$, so we can apply Theorem 1.35 from the lecture notes on measure theory. A distribution with this type of density is called *d-dimensional normal distribution.*

Proposition 1.43. *Let $g, h : \mathbb{R}^d \to \mathbb{R}$ be measurable functions and X, Y independent random vectors. Then $Z := h(X)$ and $W := g(Y)$ are independent.*

Proof. For any $A, B \in \mathcal{B}(\mathbb{R})$,

$$\begin{aligned}
\mathbb{P}(Z \in A, W \in B) &= \mathbb{P}(X \in h^{-1}(A), Y \in g^{-1}(B)) \\
&= \mathbb{P}(X \in h^{-1}(A))\mathbb{P}(Y \in g^{-1}(B)) \\
&= \mathbb{P}(Z \in A)\mathbb{P}(W \in B).
\end{aligned}$$

□

Theorem 1.44. *Let X and Y be integrable random variables. If X and Y are independent, then*

$$\mathbf{E}(XY) = \mathbf{E}(X)\mathbf{E}(Y).$$

Proof. **Step 1:** Let X and Y be simple random variables, i.e. non-negative stepfunctions $X = \sum_{k=1}^n c_k \mathbb{1}_{A_k}$, $Y = \sum_{l=1}^m d_l \mathbb{1}_{B_l}$ with pairwise different c_k, d_l. Then,

$$\mathbf{E}(XY) = \mathbf{E}\Big(\sum_{k=1}^n \sum_{l=1}^m c_k d_l \underbrace{\mathbb{1}_{A_k}\mathbb{1}_{B_l}}_{=\mathbb{1}_{A_k \cap B_l}}\Big) = \sum_{k=1}^n \sum_{l=1}^m c_k d_l \mathbb{P}(A_k \cap B_l) = \sum_{k=1}^n \sum_{l=1}^m c_k d_l \mathbb{P}(A_k)\mathbb{P}(B_l),$$

because $A_k \in \sigma(X)$ and $B_l \in \sigma(Y)$ and X and Y are independent; see Remark 1.39. But now

$$\sum_{k=1}^n \sum_{l=1}^m c_k d_l \mathbb{P}(A_k)\mathbb{P}(B_l) = \sum_{k=1}^n c_k \mathbb{P}(A_k) \sum_{l=1}^m d_l \mathbb{P}(B_l) = \mathbf{E}(X)\mathbf{E}(Y).$$

Step 2: Let now $X, Y \geq 0$ and $(X_n)_{n\in\mathbb{N}}$, $(Y_n)_{n\in\mathbb{N}}$ be sequences of simple random variables with $X_n \uparrow X$, $Y_n \uparrow Y$, as constructed in the proof of Theorem 2.9 in the lecture notes on measure theory, that is, $X_n = f_n(X)$, $Y_n = f_n(Y)$ with

$$f_n \colon \mathbb{R} \to \mathbb{R}, x \mapsto \sum_{k=0}^{n2^n - 1} k2^{-n} \mathbb{1}_{[k2^{-n},(k+1)2^{-n})}(x) + n\mathbb{1}_{[n,\infty)}(x).$$

The functions f_n are measurable, so by Proposition 1.43, $X_n = f_n(X)$ and $Y_n = f_n(Y)$ are also independent for all n. By the definition of the integral, we obtain

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

Step 3: In general,

$$XY = (X_+ - X_-)(Y_+ - Y_-) = (X_+Y_+ + X_-Y_-) - (X_+Y_- + X_-Y_+),$$

so

$$\begin{aligned}\mathbf{E}(XY) &= \mathbf{E}(X_+)\mathbf{E}(Y_+) + \mathbf{E}(X_-)\mathbf{E}(Y_-) - \mathbf{E}(X_+)\mathbf{E}(Y_-) - \mathbf{E}(X_-)\mathbf{E}(Y_+) \\ &= [\mathbf{E}(X_+) - \mathbf{E}(X_-)][\mathbf{E}(Y_+) - \mathbf{E}(Y_-)] = \mathbf{E}(X)\mathbf{E}(Y).\end{aligned}$$

□

Example 1.45. Let X be distributed according to a d-dimensional standard normal distribution, and let $Y := AX + b$ with an invertible $(d \times d)$-matrix A and $b \in \mathbb{R}^d$. Then the covariance matrix of X is $\Gamma^X = I_d$, where I_d denotes the d-dimensional unit matrix. It holds

$$\mathbf{E}(Y) = A\mathbf{E}(X) + b = b, \quad \Gamma^Y = A\Gamma^X A^\top = AA^\top,$$

so the matrix $\Gamma = AA^\top$ in (1.2) is the covariance matrix Γ^Y.

1.4 Sums of independent random variables

Let X and Y be independent random variables. We are interested in the distribution of their sum, that is, of the random variable

$$Z := X + Y.$$

Since X, Y are independent, their joint distribution is given by $\mathbf{P}_{(X,Y)} = \mathbf{P}_X \otimes \mathbf{P}_Y$. By applying Fubini's Theorem, we can compute the distribution of Z. Let $A \in \mathcal{B}(\mathbb{R})$.. Then,

$$\begin{aligned}
\mathbf{P}_Z(A) &= \mathbb{P}(Z \in A) = \mathbf{E}(\mathbb{1}_A(X+Y)) \\
&= \int \mathbb{1}_A(x+y)\,\mathrm{d}(\mathbf{P}_X \otimes \mathbf{P}_Y)(x,y) = \int\int \mathbb{1}_A(x+y)\,\mathrm{d}\mathbf{P}_X(x)\,\mathrm{d}\mathbf{P}_Y(y) \\
&= \int\int \mathbb{1}_{-y+A}(x)\,\mathrm{d}\mathbf{P}_X(x)\,\mathrm{d}\mathbf{P}_Y(y) = \int \mathbf{P}_X(-y+A)\,\mathrm{d}\mathbf{P}_Y(y),
\end{aligned}$$

where $-y + A := \{-y + a \mid a \in A\}$. The measure $\mathbf{P}_Z$ is called the *convolution* $\mathbf{P}_X * \mathbf{P}_Y$ of $\mathbf{P}_X$ and $\mathbf{P}_Y$,

$$(\mathbf{P}_X * \mathbf{P}_Y)(A) := \int \mathbf{P}_X(-y+A)\,\mathrm{d}\mathbf{P}_Y(y).$$

Exercise 1.46. Let X and Y be independent and discrete with values in $\mathbb{Z}$. We set $Z := X+Y$. Let $\mathbb{P}(X = k) =: p_k$ and $\mathbb{P}(Y = k) =: q_k$, $k \in \mathbb{Z}$. Show that

$$\mathbb{P}(Z = k) = \sum_{j\in\mathbb{Z}} q_j p_{k-j}, \quad k \in \mathbb{Z}.$$

Example 1.47. Let $X \sim \mathrm{POIS}(\lambda)$, $Y \sim \mathrm{POIS}(\mu)$, that is, $p_i = e^{-\lambda}\frac{\lambda^i}{i!}$ and $q_j = e^{-\mu}\frac{\mu^j}{j!}$, $i, j \in \mathbb{N}_0$. Then,

$$\mathbb{P}(Z = k) = e^{-(\lambda+\mu)} \sum_{i=0}^{k} \frac{\lambda^i \mu^{k-i}}{i!(k-i)!} = \frac{e^{-(\lambda+\mu)}}{k!} \sum_{i=0}^{k} \binom{k}{i} \lambda^i \mu^{k-i} = e^{-(\lambda+\mu)} \frac{(\lambda+\mu)^k}{k!},$$

so the sum Z of two independent Poisson random variables is again Poisson distributed with parameter $\lambda + \mu$.

Exercise 1.48. Let X and Y be independent and absolutely continuous with densities f and g. We set $Z := X + Y$. Show that for $A \in \mathcal{B}(\mathbb{R})$,

$$\mathbf{P}_Z(A) = \int \mathbb{1}_A(z) \int f(z-y)g(y)\,\mathrm{d}y\,\mathrm{d}z,$$

and conclude that Z is absolutely continuous with density

$$\int f(z-y)g(y)\,\mathrm{d}y, \quad z \in \mathbb{R}.$$

Example 1.49 (Normal distributions)**.** Let $X \sim \mathcal{N}(0, \sigma^2)$ and $Y \sim \mathcal{N}(0, \tau^2)$ be independent, that is

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}}, \quad g(y) = \frac{1}{\sqrt{2\pi}\tau} e^{-\frac{y^2}{2\tau^2}}.$$

In the expression $f(t)g(z-t)$, an exponential function with the exponent k

$$-\frac{1}{2}\left(\frac{t^2}{\sigma^2}+\frac{(z-t)^2}{\tau^2}\right)=-\frac{1}{2}\left[\frac{\sigma^2+\tau^2}{\sigma^2\tau^2}\left(t-\frac{\sigma^2 z}{\sigma^2+\tau^2}\right)^2+\frac{z^2}{\sigma^2+\tau^2}\right]$$

appears. Integrate over t by using

$$\int_{-\infty}^{\infty}\exp\left(-\frac{x^2}{2\alpha^2}\right)\mathrm{d}x=\sqrt{2\pi}\alpha$$

where $\alpha^2=\sigma^2\tau^2/(\sigma^2+\tau^2)$. Then

$$\begin{aligned} h(z) &= f*g(z) \\ &= \frac{1}{2\pi\sigma\tau}\exp\left(-\frac{z^2}{2(\sigma^2+\tau^2)}\right)\int_{-\infty}^{\infty}\exp\left(-\frac{1}{2\alpha^2}\left(t-\frac{\sigma^2 z}{\sigma^2+\tau^2}\right)^2\right)\mathrm{d}t \\ &= \frac{1}{\sqrt{2\pi(\sigma^2+\tau^2)}}\exp\left(-\frac{z^2}{2(\sigma^2+\tau^2)}\right). \end{aligned}$$

The random variable $Z=X+Y$ is therefore again normally distributed, more precisely $\mathcal{N}(0,\sigma^2+\tau^2)$.

Exercise 1.50. Show that any linear combination of the components of a normally distributed random vectors is normally distributed on $\mathbb{R}$.

Exercise 1.51. Let X and Y be independent standard normal random variables. Compute the density of the distribution of $Z=X^2+Y^2$.
Hint: You may use

$$\int_0^z \frac{1}{\sqrt{(z-u)u}}\mathrm{d}u=\pi, \quad z>0.$$

1.5 Infinite products of probability spaces

The associativity in the generation of product measures implies the existence of probability spaces on which we can define *finitely many* independent random variables. However, we would often like to define sequences (that is, *infinitely many*) of random variables. One can show that for any sequence of random variables $(X_n)_{n\in\mathbb{N}}$, $X_n:\Omega_n\to\mathbb{R}$ with distributions $\mathbf{P}_{X_n}$, $n\in\mathbb{N}$, there exists a probability space $(\Omega,\mathfrak{F},\mathbb{P})$ and a sequence of random variables $(\tilde{X}_n)_{n\in\mathbb{N}}$ with $\tilde{X}_n:\Omega\to\mathbb{R}$ and distributions $\mathbf{P}_{\tilde{X}_n}=\mathbf{P}_{X_n}$. To this end, one considers the product space $\Omega=\prod_{n=1}^{\infty}\Omega_n$ with a suitable product σ-field and a suitable product measure. We will consider this construction only in a special case.

We construct a probability space which models the random experiment of tossing a coin infinitely many times. In the n-th independent experiment, the coin shows head with probability $q_n\in[0,1]$ (encoded as 1) and tail with probability $1-q_n$ (encoded as 0). That is, we consider

$$\Omega:=\{(\omega_n)_{n\in\mathbb{N}}\mid\omega_n\in\{0,1\}\}=\{0,1\}^{\mathbb{N}},$$

the set of all 0-1-sequences.[3] We call a set $Z\subseteq\Omega$ of 0-1-sequences a *cylinder set*, it there exist $m\in\mathbb{N}$ and $C\subseteq\{0,1\}^m$ such that

$$Z=\{(\omega_n)_{n\in\mathbb{N}}\mid(\omega_1,\dots,\omega_m)\in C\}.$$

[3]This is an uncountable set; see lectures on Analysis.

We call (m, C) *representation* of Z. Attention! The representation (m, C) of a cylinder set is not unique.

Lemma 1.52. *The set $\mathcal{Z}$ consisting of all cylinder sets is a ring.*

Proof. It is obvious that $\emptyset \in \mathcal{Z}$. Let $Y, Z \in \mathcal{Z}$ with representations

$$Y = \{(\omega_n)_{n\in\mathbb{N}} \mid (\omega_1, \dots, \omega_m) \in B\}, \quad Z = \{(\omega_n)_{n\in\mathbb{N}} \mid (\omega_1, \dots, \omega_k) \in C\},$$

where $B \subseteq \{0,1\}^m$, $C \subseteq \{0,1\}^k$. We can assume that $m \leq k$. Then with $B' = B \times \{0,1\}^{k-m}$, it holds

$$Y = \{(\omega_n)_{n\in\mathbb{N}} \mid (\omega_1, \dots, \omega_k) \in B'\},$$

so

$$\begin{aligned} Y \setminus Z &= \big\{(\omega_n)_{n\in\mathbb{N}} \mid (\omega_1, \dots, \omega_k) \in B' \setminus C\big\}, \\ Z \setminus Y &= \big\{(\omega_n)_{n\in\mathbb{N}} \mid (\omega_1, \dots, \omega_k) \in C \setminus B'\big\}, \\ Y \cup Z &= \big\{(\omega_n)_{n\in\mathbb{N}} \mid (\omega_1, \dots, \omega_k) \in B' \cup C\big\}. \end{aligned}$$

□

For all $m \in \mathbb{N}$, let now $p_m \colon \{0,1\}^m \to [0,1]$ be defined as

$$p_m(y) = q_1^{y_1}(1-q_1)^{1-y_1} \dots q_m^{y_m}(1-q_m)^{1-y_m}.$$

Remark 1.53. The product measure on $\{0,1\}^m$ is given by

$$2^{\{0,1\}^m} \to [0,1], \quad C \mapsto \sum_{y\in C} p_m(y).$$

Lemma 1.54. *The function*

$$\tilde{\mathbb{P}} \colon \mathcal{Z} \to [0,1], \quad Z \mapsto \tilde{\mathbb{P}}(Z) = \sum_{y\in C_Z} p_{m_Z}(y)$$

defines a σ additive content on $\mathcal{Z}$, where $C_Z \subset \{0,1\}^{m_Z}$ is a representation of the cylinder set Z. It holds $\tilde{\mathbb{P}}(\Omega) = 1$.

Proof. First verify that the definition of $\tilde{\mathbb{P}}(Z)$ does not depend on the representation of Z (Exercise). Let now $Y, Z \in \mathcal{Z}$ with representations $B \subseteq \{0,1\}^m$, $C \subseteq \{0,1\}^k$. Let $m \leq k$. Then $Y \cap Z = \emptyset$ implies that $B \times \{0,1\}^{k-m} \cap C = \emptyset$, so $\tilde{\mathbb{P}}(Y \cup Z) = \tilde{\mathbb{P}}(Y) + \tilde{\mathbb{P}}(Z)$. Clearly, $\tilde{\mathbb{P}}(\Omega) = 1$, because Ω admits the representation $(1, \{0,1\})$. If $(Z_n)_{n\in\mathbb{N}}$ is a sequence in $\mathcal{Z}$ with $Z_n \cap Z_k = \emptyset$ and $Z := \bigcup_{n\in\mathbb{N}} Z_n \in \mathcal{Z}$, then only finitely many Z_n can be non-empty, because Z is a cylinder set, so the σ-additivity is a direct consequence of the additivity. □

Due to the measure extension theorem by Carathéodory and Theorem 1.35 in the lecture notes on meassure theory, there exists a unique probability measure $\mathbb{P}$ on $\sigma(\mathcal{Z})$ that extends $\tilde{\mathbb{P}}$.

Example 1.55 (Infinite toss of coin)**.** Assume $q = q_m$, i.e. the probability of head (1) is equal in each toss. Then for $y \in \{0,1\}^m$,

$$p_m(y) = p^{\#\{i \mid y_i=1\}}(1-p)^{\#\{i \mid y_i=0\}} = p^{\sum_{j=1}^m y_j}(1-p)^{m-\sum_{j=1}^m y_j}.$$

Example 1.56 (Geometric distribution)**.** We consider the probability space from Example 1.55, that is, $\Omega = \{0,1\}^{\mathbb{N}}$. X shall denote the waiting time until the first success (head, encoded as 1) in a sequence of coin tosses, i.e.

$$X(\omega) = \min\{k \in \mathbb{N} \mid \omega_1, \ldots, \omega_k = 0, \omega_{k+1} = 1\}.$$

Because $\{X = k\}$ is a cylinder set, we directly obtain

$$\mathbb{P}(X = k) = (1-p)^k p.$$

We denote this distribution as $X \sim \text{GEOM}(p)$.

Taking the derivative of $\sum_{i=0}^{\infty} x^i = (1-x)^{-1}$ (Analysis: computing with power series), we obtain $\sum_{i=1}^{\infty} i x^{i-1} = (1-x)^{-2}$ and $\sum_{i=1}^{\infty} i(i-1)x^{i-2} = 2(1-x)^{-3}$, so

$$\begin{aligned}
\mathbf{E}(X) &= p\sum_{j=1}^{\infty}(1-p)^j j = p(1-p)\sum_{j=1}^{\infty}(1-p)^{j-1} j = \frac{p(1-p)}{p^2} = \frac{1-p}{p} \\
\mathbf{E}(X^2) &= p\sum_{j=1}^{\infty}(1-p)^j j^2 = p\sum_{j=1}^{\infty}(1-p)^j[j(j-1)+j] \\
&= p\big(2(1-p)^2 p^{-3} + (1-p)p^{-2}\big) = 2\frac{(1-p)^2}{p^2} + \frac{1-p}{p}.
\end{aligned}$$

Therefore,

$$\mathsf{Var}(X) = \mathbf{E}(X^2) - \left(\frac{1-p}{p}\right)^2 = \frac{(1-p)^2}{p^2} + \frac{1-p}{p} = \frac{1-p}{p^2}.$$

Chapter 2

Convergence of random variables

2.1 Important inequalities

Theorem 2.1 (Markov's inequality). *Let X be a random variable and $\epsilon > 0$. Then,*

$$\mathbb{P}(|X| \geq \varepsilon) \leq \frac{\mathbf{E}(|X|)}{\varepsilon}.$$

Proof. It holds

$$\mathbf{E}(|X|) \geq \mathbf{E}(|X| \mathbb{1}_{\{|X| \geq \varepsilon\}}) \geq \varepsilon \mathbb{P}(|X| \geq \varepsilon).$$

□

Corollary 2.2 (Chebyshev's inequality). *If the variance of X exists, then for all $\varepsilon > 0$,*

$$\mathbb{P}(|X - \mathbf{E}(X)| \geq \varepsilon) \leq \frac{\mathsf{Var}(X)}{\varepsilon^2}.$$

Proof. With $Y = (X - \mathbf{E}(X))^2$, it holds

$$\mathbb{P}(|X - \mathbf{E}(X)| \geq \varepsilon) = \mathbb{P}(Y \geq \varepsilon^2) \leq \frac{\mathbf{E}(Y)}{\varepsilon^2}.$$

□

Corollary 2.3 (Generalized Markov inequality). *Let g be a measurable, increasing function with $g > 0$, and let X be a random variable. Then for all $\varepsilon > 0$,*

$$\mathbb{P}(X \geq \varepsilon) \leq \frac{\mathbf{E}(g(X))}{g(\varepsilon)}.$$

Theorem 2.4 (Jensen's inequality). *Let g be a convex function and X a random variable. Then,*

$$\mathbf{E}(g(X)) \geq g(\mathbf{E}(X)),$$

if these expected values exist and are finite.

Proof. Because g is convex, for any $x_0 \in \mathbb{R}$ there exists $K \in \mathbb{R}$ such that

$$g(x) \geq g(x_0) + K(x - x_0), \quad \text{for all } x \in \mathbb{R}.$$

We choose $x_0 = \mathbf{E}(X)$ and $x = X$ and apply the expectation to both sides. □

Theorem 2.5 (Kolmogorov's inequality)**.** *For independent random variables $X_1, \dots, X_k$ with $\mathbf{E}(X_i) = 0$, $i = 1, \dots, k$, and $\varepsilon > 0$, it holds:*

$$\mathbb{P}\left(\max_{1\le j\le k}\left|\sum_{i=1}^{j} X_i\right| \ge \varepsilon\right) \le \frac{1}{\varepsilon^2}\sum_{i=1}^{k} \mathsf{Var}(X_i).$$

Proof. Set $S_j = \sum_{i=1}^{j} X_i$, $j = 1, \dots, k$, and

$$A_j = \{|S_1| < \varepsilon, \dots, |S_{j-1}| < \varepsilon, |S_j| \ge \varepsilon\}.$$

The events $A_1, \dots, A_k$ are disjoint. For all $j \in \{1, \dots, k\}$, the random variables $\mathbb{1}_{A_j} S_j$ and $S_k - S_j$ are independent, since the former one only depends on $X_1, \dots, X_j$ and the latter one only on $X_{j+1}, \dots, X_k$. This implies

$$\begin{aligned}\sum_{j=1}^{k} \mathsf{Var}(X_j) &= \mathbf{E}(S_k^2) \ge \sum_{j=1}^{k} \mathbf{E}(\mathbb{1}_{A_j} S_k^2) = \sum_{j=1}^{k} \mathbf{E}(\mathbb{1}_{A_j}(S_j + (S_k - S_j))^2)\\ &\ge \sum_{j=1}^{k} \big(\mathbf{E}(\mathbb{1}_{A_j} S_j^2) + 2\mathbf{E}(\mathbb{1}_{A_j} S_j)\mathbf{E}(S_k - S_j)\big) = \sum_{j=1}^{k} \mathbf{E}(\mathbb{1}_{A_j} S_j^2)\\ &\ge \sum_{j=1}^{k} \varepsilon^2 \mathbb{P}(A_j) = \varepsilon^2 \mathbb{P}\left(\bigcup_{j=1}^{k} A_j\right) = \varepsilon^2 \mathbb{P}\left(\max_{1\le j\le k} |S_j| \ge \varepsilon\right).\end{aligned}$$

□

Exercise 2.6. Let $X_1, \dots, X_n$ be independent standard normal random variables. Prove that

$$\mathbb{P}\left(\max_{1\le k\le n}\left|\sum_{i=1}^{k} X_i\right| \ge n\right) \le \frac{1}{n}.$$

2.2 Notions of convergence

Definition 2.7. Let X and $X_1, X_2, \dots$ be random variables on a probability space $(\Omega, \mathfrak{F}, \mathbb{P})$. The sequence $(X_n)_{n\in\mathbb{N}}$ *converges almost surely to* X if

$$\mathbb{P}(\{\omega :\ \limsup_{n\to\infty} X_n(\omega) = \liminf_{n\to\infty} X_n(\omega) = X(\omega)\}) = 1,$$

and we write $X_n \xrightarrow{\text{a.s.}} X$. The sequence $(X_n)_{n\in\mathbb{N}}$ *converges in probability to* X if for all $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \varepsilon) = 0,$$

and we write $X_n \xrightarrow{\text{P}} X$.

Convergence almost surely means that $X_n(\omega) \to X(\omega)$, $n \to \infty$, holds for all ω. Convergence in probability means that the probability of the event $\{|X_n - X| > \varepsilon\}$ converges to 0 for $n \to \infty$.

Exercise 2.8. Let $\Omega = [0,1]$, $\mathfrak{F} = \mathcal{B}([0,1])$ and $\mathbb{P} = \mathcal{L}^1|_{[0,1]}$. We define $X_n = n\mathbb{1}_{A_n}$ with $A_n = [0, 1/n]$ for all $n \in \mathbb{N}$. Examine whether the sequence $(X_n)_{n\in\mathbb{N}}$ converges almost surely and/or in probability, using the definitions above.

Theorem 2.9. *Let $X, X_1, X_2, \ldots$ be random variables. Then the following statements are equivalent.*

1. $X_n \xrightarrow{\text{a.s.}} X$.
2. *For all* $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\big(\bigcup_{k\geq n}\{|X_k - X| > \varepsilon\}\big) = \mathbb{P}\big(\bigcap_{n\geq 1}\bigcup_{k\geq n}\{|X_k - X| > \varepsilon\}\big) = 0.$$

Proof. $X_n(\omega)$ does not converge to $X(\omega)$ if and only if there exists j such that $|X_n(\omega) - X(\omega)| > 1/j$ for infinitely many n, that is,

$$\{\omega \mid X_n(\omega) \not\to X(\omega)\} = \bigcup_{j\geq 1}\Big\{|X_n - X| > \frac{1}{j} \text{ for infinitely many } n\Big\}.$$

Hence, $X_n \xrightarrow{\text{a.s.}} X$ if any only if

$$\mathbb{P}\Big(\Big\{|X_n - X| > \frac{1}{j} \text{ for infinitely many } n\Big\}\Big) = 0$$

for all $j \in \mathbb{N}$. With

$$\Big\{|X_n - X| > \frac{1}{j} \text{ for infinitely many } n\Big\} = \bigcap_{n\geq 1}\bigcup_{k\geq n}\Big\{|X_k - X| > \frac{1}{j}\Big\},$$

we obtain the equivalence of 1. and 2. (first with $\varepsilon = 1/j$ and then for all $\varepsilon > 0$). □

Corollary 2.10. *Convergence almost surely implies convergence in probability.*

Proof. This is a consequence of Theorem 2.9 (2.). □

Attention: The reverse statement of Corollary 2.10 is not true!

Example 2.11. Let $\Omega = [0, 1]$ be endowed with the Lebesgue-measure. Defining

$$X_n(\omega) = \begin{cases} 1, & \omega \in [j2^{-k}, (j+1)2^{-k}], \\ 0, & \text{sonst}, \end{cases}$$

where $n = 2^k + j$, with $k = 0, 1, \ldots$ and $j = 0, \ldots, 2^k - 1$ and $X(\omega) = 0$, it holds

$$\mathbb{P}(|X_n - X| > 0) = 2^{-k} \to 0 \quad (n \to \infty),$$

so $(X_n)_{n\in\mathbb{N}}$ converges in probability to X. However, the sequence $X_n(\omega)$ contains infinitely many 0 for any ω and therefore it does not converge.

Definition 2.12. Let X and $X_1, X_2, \ldots$ be random variables and $p > 0$. Then X_n *converges in* L^p *to* X, if

$$\lim_{n\to\infty} \mathbf{E}|X_n - X|^p = 0$$

and we write $X_n \xrightarrow{L^p} X$.

Theorem 2.13. *Convergence in L^p implies convergence in probability.*

Proof. Markov's inequality (Theorem 2.1) implies

$$\mathbb{P}(|X_n - X| > \varepsilon) = \mathbb{P}(|X_n - X|^p > \varepsilon^p) \leq \varepsilon^{-p}\mathbf{E}|X_n - X|^p.$$

□

Convergence in L^p is neither sufficient nor necessary for convergence almost surely.

Theorem 2.14 (Convergence of series)**.** *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent square integrable random variables. If the sequences $\sum_{n=1}^{\infty}\mathbf{E}(X_n)$ and $\sum_{n=1}^{\infty}\mathsf{Var}(X_n)$ both converge absolutely, then the sequence $\sum_{n=1}^{\infty} X_n$ converges almost surely.*

Proof. We can assume without loss of generality that $\mathbf{E}(X_n) = 0$ for all n. Let $S_n = \sum_{i=1}^{n} X_i$ and $m \in \mathbb{N}$. Kolmogorov's inequality (Theorem 2.5) implies that for all $\varepsilon > 0$,

$$\begin{aligned}
\mathbb{P}\left(\sup_{n>m}|S_n - S_m| > \varepsilon\right) &= \mathbb{P}\left(\bigcup_{k\geq 1}\left\{\max_{m<n\leq m+k}|S_n - S_m| > \varepsilon\right\}\right) \\
&= \lim_{k\to\infty}\mathbb{P}\left(\max_{m<n\leq m+k}|S_n - S_m| > \varepsilon\right) \\
&\leq \limsup_{k\to\infty}\frac{1}{\varepsilon^2}\sum_{n=m+1}^{m+k}\mathsf{Var}(X_n) \\
&= \frac{1}{\varepsilon^2}\sum_{n>m}\mathsf{Var}(X_n).
\end{aligned}$$

By assumption,

$$\lim_{m\to\infty}\mathbb{P}\left(\sup_{n>m}|S_n - S_m| > \varepsilon\right) = 0.$$

The following Lemma 2.16 shows that then S_n converges almost surely. □

Exercise 2.15. Let $(Z_n)_{n\in\mathbb{N}}$ be a sequence of independent random variables with $Z_n \in \{-1, 1\}$. Let $\mathbb{P}(Z_n = 1) = p \in [0,1]$ and $\mathbb{P}(Z_n = -1) = 1 - p$. Is there a $p \in [0,1]$ such that

$$\sum_{n=1}^{\infty}\frac{Z_n}{n}$$

converges almost surely?

Lemma 2.16. *A sequence $(X_n)_{n\in\mathbb{N}}$ of random variables converges almost surely if any only if for all $\varepsilon > 0$,*

$$\lim_{n\to\infty}\mathbb{P}\left(\sup_{k,l\geq n}|X_k - X_l| \geq \varepsilon\right) = 0. \tag{2.1}$$

This condition is equivalent to

$$\lim_{n\to\infty}\mathbb{P}\left(\sup_{k\geq 0}|X_{n+k} - X_n| \geq \varepsilon\right) = 0.$$

Proof. A sequence of real numbers converges if and only if it is a Cauchy sequence. Now we can use similar arguments as in Theorem 2.9 □

2.3 Zero-one-laws

Let $(\Omega, \mathfrak{F}, \mathbb{P})$ be a probability space.

2.3.1 Borel-Cantelli Lemma

Theorem 2.17 (Borel-Cantelli Lemma). *Let $(A_k)_{k\in\mathbb{N}}$ be a sequence of events and*

$$A := \{\omega \mid \omega \in A_k \text{ for infinitely many } k\} = \bigcap_{n\geq 1} \bigcup_{m\geq n} A_m =: \limsup_{k\to\infty} A_k$$

1. *If $\sum_{k=1}^{\infty} \mathbb{P}(A_k) < \infty$, then $\mathbb{P}(A) = 0$.*
2. *If the sequence $(A_k)_{k\in\mathbb{N}}$ is independent and $\sum_{k=1}^{\infty} \mathbb{P}(A_k) = \infty$, then $\mathbb{P}(A) = 1$.*

Proof. 1. The definition of A implies that

$$A \subseteq \bigcup_{k=n}^{\infty} A_k, \quad \text{for } n = 1, 2, \ldots,$$

so

$$\mathbb{P}(A) \leq \sum_{k=n}^{\infty} \mathbb{P}(A_k), \quad \text{for } n = 1, 2, \ldots.$$

Because the sequence $\sum_{k=1}^{\infty} \mathbb{P}(A_k)$ converges, its remainder $r_m = \sum_{k=m}^{\infty} \mathbb{P}(A_k)$ converges to zero, which proves the claim.

2. We first prove the following claim (see Analysis). For any sequence $(\alpha_n)_{n\in\mathbb{N}}$ of numbers $\alpha_n \in [0, 1]$,

$$\sum_{n=1}^{\infty} \alpha_n = \infty \quad \Rightarrow \quad \prod_{n=1}^{\infty} (1 - \alpha_n) = 0.$$

Now note that the sequence of complements $(A_k^c)_{k\in\mathbb{N}}$ is also independent. Therefore,

$$1 - \mathbb{P}(A) = \mathbb{P}(A^c) = \mathbb{P}\Big(\bigcup_{n\geq 1} \bigcap_{k\geq n} A_k^c\Big) = \lim_{n\to\infty} \mathbb{P}\Big(\bigcap_{k\geq n} A_k^c\Big) = \lim_{n\to\infty} \lim_{N\to\infty} \mathbb{P}\Big(\bigcap_{k=n}^{N} A_k^c\Big)$$

$$= \lim_{n\to\infty} \lim_{N\to\infty} \prod_{k=n}^{N} \mathbb{P}\big(A_k^c\big) = \lim_{n\to\infty} \lim_{N\to\infty} \prod_{k=n}^{N} (1 - \mathbb{P}\big(A_k\big))$$

With the claim above, it follows that

$$\lim_{N\to\infty} \prod_{k=n}^{N} (1 - \mathbb{P}\big(A_k\big)) = 0$$

for all $n = 1, 2, \ldots$, which gives statement (2.).

□

Exercise 2.18. Let $p \in (0, 1]$. Show that there exist a probability space and a sequence of events $(A_n)_{n\in\mathbb{N}}$ with $\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty$ and $\mathbb{P}(\limsup_{n\to\infty} A_n) = p$.

Corollary 2.19. *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of random variables and X a random variable. If for all $\varepsilon > 0$,*

$$\sum_{n=1}^{\infty} \mathbb{P}\big(|X_n - X| > \varepsilon\big) < \infty,$$

then $X_n \xrightarrow{\text{a.s.}} X$ for $n \to \infty$.

Proof. Exercise. □

Corollary 2.20. *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of random variables which converges to X in probability, i.e. $X_n \xrightarrow{\text{P}} X$. Then there exists a subsequence $(X_{n_k})_{k\in\mathbb{N}}$ with $X_{n_k} \xrightarrow{\text{a.s.}} X$ for $k \to \infty$.*

Proof. We define the index sequence n_k inductively. Let $n_1 := 0$, and define n_k to be the smallest integer $n > n_{k-1}$ such that

$$\mathbb{P}(|X_m - X| > 2^{-k}) \leq 2^{-k} \quad \text{f"ur alle } m \geq n.$$

(Such a n exists because for all ε $(= 2^{-k})$, the sequence $\mathbb{P}(|X_m - X| > \varepsilon)$ converges to zero.) Then,

$$\sum_{k=1}^{\infty} \mathbb{P}(|X_{n_k} - X| > 2^{-k}) \leq \sum_{k=1}^{\infty} 2^{-k} < \infty.$$

Corollary 2.19 yields the desired result. □

Exercise 2.21. Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of random variables and X a random variable. Assume that any subsequence $(X_n)_{n\in\mathbb{N}}$ contains a subsequence which converge almost surely to X. Show that then $X_n \xrightarrow{\text{P}} X$.

2.3.2 Kolmogorov's zero-one law

Definition 2.22. Let $(\mathcal{A}_n)_{n\in\mathbb{N}}$ be a sequence of σ-fields $\mathcal{A}_n \subseteq \mathfrak{F}$. Let

$$\mathfrak{A}_n := \sigma\Big(\bigcup_{m\geq n} \mathcal{A}_m\Big)$$

the σ-field generated by $\mathcal{A}_n, \mathcal{A}_{n+1}, \ldots$. Then

$$\mathfrak{A}_\infty := \bigcap_{n\geq 1} \mathfrak{A}_n$$

is called the σ-field of *terminal events* of the sequence $(\mathcal{A}_n)_{n\in\mathbb{N}}$.

Theorem 2.23 (Kolmogorov's zero-one law)**.** *Let $(\mathcal{A}_n)_{n\in\mathbb{N}}$ be an independent sequence of σ-fields $\mathcal{A}_n \subseteq \mathfrak{F}$. Then for any terminal event $A \in \mathfrak{A}_\infty$, either $\mathbb{P}(A) = 0$ or $\mathbb{P}(A) = 1$.*

Before proving this Theorem, we consider an important application.

Definition 2.24. Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of random variables. A *terminal event* of the sequence $(X_n)_{n\in\mathbb{N}}$ is a terminal event of the sequence $(\sigma(X_n))_{n\in\mathbb{N}}$, i.e. an element of the σ-field

$$\mathfrak{A}_\infty = \bigcap_{n=1}^{\infty} \mathfrak{A}_n,$$

where

$$\mathfrak{A}_n = \sigma\Big(\bigcup_{m\geq n} \sigma(X_m)\Big).$$

Example 2.25. Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of random variables. Then,

1. $\{\limsup_{n\to\infty} X_n < \infty\}$ is a terminal event.
2. Let $S_n = X_1+\cdots+X_n$. Then $\{\limsup_{n\to\infty} S_n/n = \liminf_{n\to\infty} S_n/n = 0\}$ and $\{(S_n)_{n\in\mathbb{N}} \text{ converges}\}$ are terminal events.
   Hence, Kolmogorov's zero-one law shows that for a sequence $(X_n)_{n\in\mathbb{N}}$ of independent random variables, the series
   $$\sum_{n=1}^{\infty} X_n$$
   converges with probability 0 or 1. Theorem 2.14 gives one possible condition under which this happens with probability 1.
3. In general, $\{X_n = 0 \ \forall n \geq 1\}$ and $\{\limsup_{n\to\infty} S_n \leq c\} = \{\limsup_{n\to\infty} \sum_{n=1}^{N} X_n \leq c\}$ are not terminal events.

To prove the zero-one law, we need the following theorem.

Theorem 2.26. *Let $\{\mathcal{A}_i \mid i \in I\}$ be an independent family of $\cap$-stable sets of events $\mathcal{A}_i \subseteq \mathfrak{F}$. Then the family $\{\sigma(\mathcal{A}_i) \mid i \in I\}$ is independent.*

Proof. A family of events is independent if and only if all its finite sub-families are independent, so we can assume that I is finite. For an arbitrary $i_0 \in I$, define

$$\mathcal{B}_{i_0} = \{B \in \mathfrak{F} \mid \{\mathcal{A}_i \mid i \in I\backslash\{i_0\}\} \cup \{\{B\}\} \text{ is independent}\}.$$

One can show that $\mathcal{B}_{i_0}$ is a Dynkin system.

Clearly, $\mathcal{A}_{i_0} \subseteq \mathcal{B}_{i_0}$. By assumption, $\mathcal{A}_{i_0}$ is $\cap$-stable, so Theorem 1.15 in the lecture notes on measure theory implies that

$$\sigma(\mathcal{A}_{i_0}) = \mathfrak{D}(\mathcal{A}_{i_0}) \subset \mathcal{B}_{i_0}.$$

By construction of $\mathcal{B}_{i_0}$, the family $\{\mathcal{A}_i \mid i \in I\}$ remains independent of we replace $\mathcal{A}_{i_0}$ by a subset of $\mathcal{B}_{i_0}$, for example by $\sigma(\mathcal{A}_{i_0})$. Therefore, $\{\mathcal{A}_i \mid i \in I\backslash\{i_0\}\} \cup \{\sigma(\mathcal{A}_{i_0})\}$ is independent. Since I is finite, we can repeat this procedure infinitely often and obtain that $\{\sigma(\mathcal{A}_i) \mid i \in I\}$ is independent. □

Proof of Theorem 2.23. Let $A \in \mathfrak{A}_\infty$. We define $\mathfrak{M}$ to be the family of all events which are independent of A, that is,

$$\mathfrak{M} := \big\{M \in \mathfrak{F} \mid \mathbb{P}(A \cap M) = \mathbb{P}(A)\mathbb{P}(M)\big\}.$$

We will show that $A \in \mathfrak{M}$, which implies $\mathbb{P}(A) = \mathbb{P}(A \cap A) = \mathbb{P}(A)\mathbb{P}(A)$ and therefore $\mathbb{P}(A) \in \{0, 1\}$.

Step 1: Show that $\mathfrak{M}$ is a Dynkin system.

Step 2: We show that for any $n \in \mathbb{N}$: $\mathfrak{A}_{n+1}$ is independent of $\mathfrak{B}_n := \sigma(\mathcal{A}_1 \cup \cdots \cup \mathcal{A}_n)$. For arbitrary $k \in \mathbb{N}$, $l \in \{1, \ldots, n\}$ let $J = \{j_1, \ldots, j_k\} \subseteq \{n+1, n+2, \ldots\}$, $I = \{i_1, \ldots, i_l\} \subseteq \{1, \ldots, n\}$ be finite non-empty subsets. For

$$B_s \in \mathcal{A}_{j_s},\ s = 1, \ldots, k, \quad C_t \in \mathcal{A}_{i_t},\ t = 1, \ldots, l$$

the events

$$B_1 \cap \cdots \cap B_k, \quad \text{and} \quad C_1 \cap \cdots \cap C_l$$

are independent. So the following two families $\mathcal{F}, \mathcal{G}$ of sets are independent:

$$\mathcal{F} := \{B_1 \cap \cdots \cap B_k \mid k \geq 1, \{j_1, \ldots, j_k\} \subseteq \{n+1, n+2, \ldots\}, B_s \in \mathcal{A}_{j_s}, s = 1, \ldots, k\}$$
$$\mathcal{G} := \{C_1 \cap \cdots \cap C_l \mid 1 \leq l \leq n, \{i_1, \ldots, i_l\} \subseteq \{1, \ldots, n\}, C_t \in \mathcal{A}_{i_t}, t = 1, \ldots, l\}.$$

It holds $\mathfrak{A}_{n+1} = \sigma(\mathcal{F})$ and $\mathcal{B}_n = \sigma(\mathcal{G})$. The families $\mathcal{F}, \mathcal{G}$ are $\cap$-stable, so we can apply Theorem 2.26.

Step 3: We know that $A \in \mathfrak{A}_{n+1}$, so Step 2 shows $\mathfrak{B}_n \subseteq \mathfrak{M}$ for all $n \in \mathbb{N}$. Thus

$$\mathfrak{B}_0 := \bigcup_{n=1}^{\infty} \mathfrak{B}_n \subseteq \mathfrak{M}.$$

The family of sets $\mathfrak{B}_0$ is $\cap$-stable, so Theorem 1.15 in the lecture notes on measure theory implies that

$$\sigma(\mathfrak{B}_0) \subseteq \mathfrak{M}.$$

For all $n \geq 1$, $\mathcal{A}_n \subseteq \mathfrak{B}_0$ due to the definition of $\mathfrak{B}_n$, so also $\mathfrak{A}_n \subseteq \sigma(\mathfrak{B}_0)$, and in particular $\mathfrak{A}_\infty \subseteq \sigma(\mathfrak{B}_0)$. Therefore,

$$\mathfrak{A}_\infty \subseteq \sigma(\mathfrak{B}_0) \subseteq \mathfrak{M},$$

and $A \in \mathfrak{M}$, which is what we had to show.

□

2.4 Law of large numbers

Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of random variables. The laws of large numbers are statemebts about the convergence of

$$\frac{1}{n} S_n := \frac{1}{n}(X_1 + \cdots + X_n)$$

for $n \to \infty$. The *weak law* describes the convergence of S_n/n in probability, the *strong law* convergence almost surely.

Theorem 2.27 (Weak law of large numbers)**.** *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent and identically distributed random variables. Then,*

$$\lim_{n\to\infty} \mathbf{E}\big(|S_n/n - \mathbf{E}(X_1)|\big) = 0$$

and

$$S_n/n \xrightarrow{\text{P}} \mathbf{E}(X_1).$$

Proof. See course on combinatorics and probability. □

Exercise 2.28. Prove Theorem 2.27 for square integrable random variables. (Hint: Chebychev's inequality.)

Theorem 2.29 (Cantelli's Theorem)**.** *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent random variables. Let $S_n = X_1 + \cdots + X_n$. If $\mathbf{E}(X_n - \mathbf{E}(X_n))^4 \leq C$for some constant C and all $n \geq 1$, then $(S_n - \mathbf{E}(S_n))/n \xrightarrow{\text{a.s.}} 0$.*

Proof. Without loss of generality, assume that $\mathbf{E}X_n = 0$ for n. (Otherwise consider $\tilde{X}_n = X_n - \mathbf{E}(X_n)$.) Due to the Borel-Cantelli Lemma (Corollary 2.19), we have to show that

$$\sum_{n=1}^{\infty} \mathbb{P}\Big(\Big|\frac{S_n}{n}\Big| \geq \varepsilon\Big) < \infty.$$

Markov's inequality (Theorem 2.1) implies that this series is bounded by

$$\varepsilon^{-4} \sum_{n=1}^{\infty} \mathbf{E}\left(\left|\frac{S_n}{n}\right|^4\right). \tag{2.2}$$

Because $\mathbf{E}(X_n) = 0$, we have

$$\begin{aligned}\mathbf{E}(S_n^4) &= \sum_{i=1}^{n} \mathbf{E}(X_i^4) + 6 \sum_{1\leq i<j\leq n} \mathbf{E}(X_i^2)\mathbf{E}(X_j^2) \\ &\leq nC + 6 \sum_{1\leq i<j\leq n} \sqrt{\mathbf{E}(X_i^4)\mathbf{E}(X_j^4)} \\ &\leq nC + 6\frac{n(n-1)}{2}C.\end{aligned}$$

So the series (2.2) converges. □

Theorem 2.30 (Kolmogorov's Theorem)**.** *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent integrable random variables. Let $S_n = X_1 + \cdots + X_n$. If*

$$\sum_{n=1}^{\infty} \frac{\mathsf{Var}(X_n)}{n^2} < \infty,$$

then $(S_n - \mathbf{E}(S_n))/n \xrightarrow{\text{a.s.}} 0$.

Proof. We can assume that $\mathbf{E}(X_n) = 0$ for all n. Set

$$Y_n := \max_{k\in\{1,\ldots,2^n\}} |S_k|.$$

For $2^{n-1} \leq m \leq 2^n$, it follows that

$$|m^{-1}S_m| \leq 2^{-n+1} \max_{k\in\{1,\ldots,2^n\}} |S_k| = 2^{-n+1}Y_n.$$

So it is sufficient to show that

$$\mathbb{P}(\lim_{n\to\infty} 2^{-n}Y_n = 0) = 1.$$

Due to the Borel-Cantelli Lemma (Corollary 2.19), it is sufficient to show

$$\sum_{n=1}^{\infty} \mathbb{P}(2^{-n} Y_n \geq \varepsilon) < \infty$$

for all$\varepsilon > 0$. Kolmogorov's inequality (Theorem 2.5) implies that

$$\mathbb{P}(2^{-n} Y_n \geq \varepsilon) = \mathbb{P}(Y_n \geq 2^n \varepsilon) \leq \frac{1}{\varepsilon^2 2^{2n}} \sum_{k=1}^{2^n} \mathsf{Var}(X_k).$$

This shows

$$\begin{aligned} \sum_{n=1}^{\infty} \mathbb{P}(2^{-n} Y_n \geq \varepsilon) &\leq \varepsilon^{-2} \sum_{n=1}^{\infty} 2^{-2n} \sum_{k=1}^{2^n} \mathsf{Var}(X_k) \\ &\leq \varepsilon^{-2} \sum_{k=1}^{\infty} \mathsf{Var}(X_k) \sum_{n:\, 2^n \geq k} 4^{-n} \\ &\leq \varepsilon^{-2} \sum_{k=1}^{\infty} \mathsf{Var}(X_k) \frac{4k^{-2}}{1 - 1/4} < \infty, \end{aligned}$$

where we used

$$\sum_{n:\, 2^n \geq k} 4^{-n} \leq \sum_{n=[\log_2(k)]}^{\infty} 4^{-n} = 4^{-[\log_2(k)]} \sum_{n=0}^{\infty} 4^{-n} \leq 4^{-(\log_2(k)-1)} \frac{1}{1 - 1/4} = k^{-2} \frac{4}{1 - 1/4}$$

in the last inequality. □

Theorem 2.31 (Strong law of large numbers). *Let $(X_n)_{n \in \mathbb{N}}$ be a sequence of independent and identically distributed random variables with $\mathbf{E}(|X_1|) < \infty$. Let $S_n = X_1 + \cdots + X_n$. Then S_n/n converges almost surely to $\mathbf{E}(X_1)$, i.e. $S_n/n \xrightarrow{\text{a.s.}} \mathbf{E}(X_1)$.*

The prove uses the following Lemma, which is an exercise.

Lemma 2.32. *For any random variable $X \geq 0$,*

$$\sum_{n=1}^{\infty} \mathbb{P}(X \geq n) \leq \mathbf{E}(X) \leq 1 + \sum_{n=1}^{\infty} \mathbb{P}(X \geq n).$$

Proof of Theorem 2.31. We can again assume that $\mathbf{E}(X_n) = \mathbf{E}(X_1) = 0$. Lemma 2.32 implies that

$$\mathbf{E}(|X_1|) < \infty \iff \sum_{n=1}^{\infty} \mathbb{P}(|X_1| \geq n) < \infty \iff \sum_{n=1}^{\infty} \mathbb{P}(|X_n| \geq n) < \infty,$$

where the second equivalence is true because the X_n are identically distributed. Together with the Borel-Cantelli Lemma (Theorem (Satz 2.17), we deduce that

$$\mathbb{P}\Big(\bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} \{|X_k| \geq k\} \Big) = \mathbb{P}(|X_n| \geq n \text{ for infinitely many } n) = 0,$$

so for any $\omega \in \Omega$ there exists $n_0 = n_0(\omega) \in \mathbb{N}$ such that

$$|X_n(\omega)| < n \quad \text{for all } n \geq n_0(\omega).$$

Let

$$Y_n := X_n \mathbb{1}\{|X_n| \leq n\}.$$

The above inequality implies that

$$\frac{S_n}{n} = \frac{X_1 + \cdots + X_n}{n} \xrightarrow{\text{a.s.}} 0 \quad \Longleftrightarrow \quad \frac{Y_1 + \cdots + Y_n}{n} \xrightarrow{\text{a.s.}} 0.$$

It holds

$$\mathbf{E}(Y_n) = \mathbf{E}(X_n \mathbb{1}\{|X_n| \leq n\}) = \mathbf{E}(X_1 \mathbb{1}\{|X_1| \leq n\}) \to \mathbf{E}(X_1) = 0,$$

This implies

$$\frac{1}{n}\sum_{k=1}^{n} \mathbf{E}(Y_k) \to 0, \tag{2.3}$$

(Exercise). We set $Z_n = Y_n - \mathbf{E}(Y_n)$. Then,

$$\frac{Y_1 + \cdots + Y_n}{n} \xrightarrow{\text{a.s.}} 0 \quad \Longleftrightarrow \quad \frac{Z_1 + \cdots + Z_n}{n} \xrightarrow{\text{a.s.}} 0$$

because (2.3).

The sequence $(Z_n)_{n\in\mathbb{N}}$ consists of independent square integrable random variables. For the variances, we obtain that

$$\begin{aligned}
\sum_{n=1}^{\infty} \frac{\mathsf{Var}(Z_n)}{n^2} &\leq \sum_{n=1}^{\infty} \frac{\mathbf{E}(Y_n^2)}{n^2} \\
&= \sum_{n=1}^{\infty} \frac{\mathbf{E}(X_n^2 \mathbb{1}\{|X_n| \leq n\})}{n^2} \\
&= \sum_{n=1}^{\infty} \frac{\mathbf{E}(X_1^2 \mathbb{1}\{|X_1| \leq n\})}{n^2} \\
&= \sum_{n=1}^{\infty} \frac{1}{n^2} \sum_{k=1}^{n} \mathbf{E}(X_1^2 \mathbb{1}\{k-1 < |X_1| \leq k\}) \\
&= \sum_{k=1}^{\infty} \mathbf{E}(X_1^2 \mathbb{1}\{k-1 < |X_1| \leq k\}) \sum_{n=k}^{\infty} \frac{1}{n^2} \\
&\leq 2\sum_{k=1}^{\infty} \frac{1}{k}\mathbf{E}(X_1^2 \mathbb{1}\{k-1 < |X_1| \leq k\}) \\
&\leq 2\sum_{k=1}^{\infty} \mathbf{E}(|X_1| \mathbb{1}\{k-1 < |X_1| \leq k\}) = 2\mathbf{E}(|X_1|) < \infty,
\end{aligned}$$

where we have used

$$\sum_{n=k}^{\infty} \frac{1}{n^2} \leq \frac{2}{k}, \quad k \geq 1.$$

This inequality follows from $\sum_{n=1}^{\infty} n^{-2} = 1 + \sum_{n=2}^{\infty} n^{-2}$ and

$$\sum_{n=k}^{\infty} \frac{1}{n^2} \leq \int_{k-1}^{\infty} \frac{1}{x^2} dx = \frac{1}{k-1} \leq \frac{2}{k}, \quad k \geq 2.$$

The claim follows by Kolmogorov's Theorem (Theorem 2.30). □

Chapter 3

Central limit theorem

3.1 Motivation

Classical probability theory considers sequences $(X_n)_{n\in\mathbb{N}}$ of independent and identically distributed (*i.i.d.*) random variables, and investigates the behaviour of sums

$$S_n = X_1 + X_2 + \cdots + X_n$$

for $n \to \infty$. The strong law of large numbers implies that S_n/n converges almost surely to $\mathbf{E}(X_1)$.

However, it is important to determine how large the deviation between S_n/n (average value) and $\mathbf{E}(X_1)$ (theoretical expected value) is. The following theorem is one of the most important results in probability theory.

Theorem (Central limit heorem). *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent and identically distributed random variables. If $\mathsf{Var}(X_1) = \sigma^2 < \infty$, then*

$$\mathbb{P}\left(\frac{\sqrt{n}}{\sigma}\Big(\frac{S_n}{n} - \mu\Big) \le t\right) \xrightarrow{n\to\infty} \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{t} e^{-x^2/2}dx = \Phi(t), \quad \textit{f"ur alle } t \in \mathbb{R}.$$

In other words

a) The random variable S_n/n has a degenerate asymptotic distribution, that is, $F(t) = 0$ if $t < \mu$ and $F(t) = 1$ otherwise.

b) The random variable
$$\frac{\sqrt{n}}{\sigma}\left(\frac{S_n}{n} - \mu\right) = \frac{S_n - n\mu}{\sqrt{n}\sigma}$$
has an asymptotic $\mathcal{N}(0,1)$ distribution.

3.2 Convergence in distribution

Definition 3.1. A sequence $(X_n)_{n\in\mathbb{N}}$ of random variables *converges in distribution to* a random variable X, if

$$F_{X_n}(t) \to F_X(t), \quad n \to \infty,$$

for all t with $F_X(t-) := \lim_{h\downarrow 0} F_X(t-h) = F_X(t)$; that is, for all t, where F_X is continuous. We write $X_n \xrightarrow{d} X$.

Remark. Convergence of $F_{X_n}(t) \xrightarrow{n\to\infty} F_X(t)$ for ALL t is a too strong condition. Here is an example: We define $X_n := n^{-1}$ for $n \geq 1$ and $X = 0$ as deterministic random variables. Intuitively, we would expect $X_n \xrightarrow{d} X$. Indeed, $F_{X_n}(t) = \mathbb{1}_{[n^{-1},\infty)}(t)$ converges to $F_X(t) = \mathbb{1}_{[0,\infty)}(t)$ for all $t \neq 0$. But there is no convergence at $t = 0$, ($F_{X_n}(0) = 0 \not\to F_X(0) = 1$), where F_X has a jump (discontinuity).

Exercise 3.2. Let X be a random variable and $X_n := X + 1/n$, $n \in \mathbb{N}$. Show that $(X_n)_{n\in\mathbb{N}}$ converges in distribution to X.

We need (especially for multi-dimensional distributions) an equivalent characterization of convergence in distribution.

We use the following notation:

C_b^0 continuous, bounded functions on $\mathbb{R}$;

C_c^2 twice continuously differentiable functions on $\mathbb{R}$ with compact support.

Lemma 3.3. *If $(X_n)_{n\in\mathbb{N}}$ converges in distribution to X, then*

$$\lim_{n\to\infty} \mathbf{E}(f(X_n)) = \mathbf{E}(f(X))$$

for all $f \in C_b^0$.

Proof. Let $f \in C_b^0$. Then there exists $a > 0$ with $|f(x)| \leq a$ for all x. Let $\varepsilon > 0$. We choose $t \in \mathbb{R}$ such that F_X is continuous at t and $-t$ and $F_X(-t) < \frac{\varepsilon}{16a}$ and $1 - F_X(t) < \frac{\varepsilon}{16a}$. Then there exists $n_0(\varepsilon)$ such that $F_{X_n}(-t) < \frac{\varepsilon}{16a}$ and $1 - F_{X_n}(t) < \frac{\varepsilon}{16a}$ for all $n \geq n_0(\varepsilon)$, by assumption. The function f is uniformly continuous on the compact interval $[-t, t]$, so the interval $[-t, t]$ can be partitioned by points $t_0 = -t < t_1 < t_2 < \cdots < t_k = t$ in such a way that F_X is continuous at all t_i and

$$\max_{t_i \leq x \leq t_{i+1}} f(x) - \min_{t_i \leq x \leq t_{i+1}} f(x) \leq \frac{\varepsilon}{8},$$

for $i = 0, 1, \ldots, k-1$. Then,

$$\Big|\mathbf{E}(f(X)) - \sum_{i=0}^{k-1} f(t_i)\mathbb{P}(t_i < X \leq t_{i+1})\Big| \leq \mathbf{E}(|f(X)|\mathbb{1}_{\{|X|\geq t\}}) + \frac{\varepsilon}{8}\mathbb{P}(|X| \leq t) \leq a\frac{\varepsilon}{8a} + \frac{\varepsilon}{8} = \frac{\varepsilon}{4}. \tag{3.1}$$

An analogous inequality holds for X_n instead of X for $n \geq n_0(\varepsilon)$. Now choose $n_1(\varepsilon) \geq n_0(\varepsilon)$ large enough such that for all $n \geq n_1(\varepsilon)$,

$$|F_X(t_i) - F_{X_n}(t_i)| < \frac{\varepsilon}{4ka}, \quad \text{für } i = 0, 1, \ldots, k,$$

so

$$|\mathbb{P}(t_i < X \leq t_{i+1}) - \mathbb{P}(t_i < X_n \leq t_{i+1})| \leq |F_X(t_{i+1}) - F_{X_n}(t_{i+1})| + |F_X(t_i) - F_{X_n}(t_i)| \leq \frac{\varepsilon}{2ka}. \tag{3.2}$$

This gives

$$\left|\sum_{i=0}^{k-1} f(t_i)\mathbb{P}(t_i < X \leq t_{i+1}) - \sum_{i=0}^{k-1} f(t_i)\mathbb{P}(t_i < X_n \leq t_{i+1})\right| \leq ka\frac{\varepsilon}{2ka} = \frac{\varepsilon}{2}. \tag{3.3}$$

The combination of (3.1), (3.2) and (3.3) now implies

$$
\begin{aligned}
|\mathbf{E}(f(X)) - \mathbf{E}(f(X_n))| \leq & \left| \mathbf{E}(f(X)) - \sum_{i=0}^{k-1} f(t_i)\mathbb{P}(t_i < X \leq t_{i+1}) \right| + \\
& + \left| \mathbf{E}(f(X_n)) - \sum_{i=0}^{k-1} f(t_i)\mathbb{P}(t_i < X_n \leq t_{i+1}) \right| + \\
& + \left| \sum_{i=0}^{k-1} f(t_i)[\mathbb{P}(t_i < X \leq t_{i+1}) - \mathbb{P}(t_i < X_n \leq t_{i+1})] \right| \leq \varepsilon.
\end{aligned}
$$

□

Lemma 3.4. *If* $\lim_{n\to\infty} \mathbf{E}(f(X_n)) = \mathbf{E}(f(X))$ *for all* $f \in C_c^2$*, then* $(X_n)_{n\in\mathbb{N}}$ *converges in distribution to* X*.*

Proof. For the proof, we construct two special functions in C_c^2.

Let $s < t$, $\varepsilon > 0$ and $g : \mathbb{R} \to \mathbb{R}, x \mapsto g(x)$ with $g(x) = 1$ for $x \in [s - \varepsilon/3, t + \varepsilon/3]$, $g(x) = 0$ for $x \in (-\infty, s - 2\varepsilon/3)$ or $x \in (t + 2\varepsilon/3, \infty)$. We expand g linearly to a continuous function on $\mathbb{R}$, as illustrated in the following figure.

Define

$$h(x) = \frac{3}{\varepsilon} \int_{x-\varepsilon/6}^{x+\varepsilon/6} g(y)dy, \qquad f(x) = \frac{3}{\varepsilon} \int_{x-\varepsilon/6}^{x+\varepsilon/6} h(y)dy.$$

The function h is represented by the dashed line in the figure. This construction yields $f \in C_2^c$ as a smoothed version of g with $f(x) \in [0, 1]$, $f(x) = 1$ on $[s, t]$, $f(x) = 0$ for $x \leq s - \varepsilon$ and $x \geq t + \varepsilon$. This implies

$$\mathbb{P}(s < X_n \leq t) \leq \mathbf{E}(f(X_n)) \xrightarrow{n\to\infty} \mathbf{E}(f(X)) \leq \mathbb{P}(s - \varepsilon < X \leq t + \varepsilon).$$

If s and t are points where F_X is continuous, then one can let ε converge to 0, and obtains

$$\limsup_{n\to\infty} \mathbb{P}(s < X_n \leq t) \leq \mathbb{P}(s < X \leq t).$$

In a similar way, but with a function $\tilde{f}$ that goes to zero on the outside of $[s, t]$ and is equal to one on $[s + \varepsilon, t - \varepsilon]$, one can show that

$$\liminf_{n\to\infty} \mathbb{P}(s < X_n \leq t) \geq \mathbb{P}(s < X \leq t).$$

Together, if $\mathbb{P}(X = s) = \mathbb{P}(X = t) = 0$:

$$F_{X_n}(t) - F_{X_n}(s) \xrightarrow{n\to\infty} F_X(t) - F_X(s). \tag{3.4}$$

Let now $\varepsilon > 0$ and $\bar{s}, \bar{t}$ such that $\bar{s}$, $\bar{t}$ are points of continuity of F_X and

$$F_X(\bar{t}) - F_X(\bar{s}) > 1 - \frac{\varepsilon}{4}.$$

Then due to (3.4), there exists n_0 such that for $n \geq n_0$,

$$1 - F_{X_n}(\bar{s}) \geq F_{X_n}(\bar{t}) - F_{X_n}(\bar{s}) > 1 - \frac{\varepsilon}{4},$$

so $F_{X_n}(\bar{s}) < \varepsilon/4$ for all $n \geq n_0$. For a particular t which is a point of continuity of F_X, choose $n_1 \geq n_0$ such that for all $n \geq n_1$,

$$|F_{X_n}(t) - F_{X_n}(\bar{s}) - (F_X(t) - F_X(\bar{s}))| < \frac{\varepsilon}{2}.$$

Then for $n \geq n_1$,

$$|F_{X_n}(t) - F_X(t)| \leq |F_{X_n}(t) - F_{X_n}(\bar{s}) - (F_X(t) - F_X(\bar{s}))| + |F_{X_n}(\bar{s}) - F_X(\bar{s})| < \frac{\varepsilon}{2} + 2\frac{\varepsilon}{4} = \varepsilon.$$

□

Exercise 3.5. Assume that the sequence of random variables $(X_n)_{n\in\mathbb{N}}$ converges in distribution to X. Show that then $\lim_{r\to\infty} \limsup_{n\to\infty} \mathbb{P}(|X_n| > r) = 0$.
Hint: Consider the function $f(x) = (1 - (r - |x|)_+)_+$.

The statements of Lemma 3.3 and Lemma 3.4 can be further generalized to the *Portemanteau Theorem.*

Theorem 3.6 (Portemanteau Theorem)**.** *Let X be a random variable and $(X_n)_{n\in\mathbb{N}}$ a sequence of random variables. Then the following statements are equivalent.*

1. $X_n \xrightarrow{d} X$, $n \to \infty$.
2. *For all $f \in \mathcal{C}_b^0$,*
$$\lim_{n\to\infty} \mathbf{E}(f(X_n)) = \mathbf{E}(f(X)).$$
3. *For all $f \in C_c^2$,*
$$\lim_{n\to\infty} \mathbf{E}(f(X_n)) = \mathbf{E}(f(X)).$$
4. *For all open sets $G \subseteq \mathbb{R}$,*
$$\liminf_{n\to\infty} \mathbb{P}(X_n \in G) \geq \mathbb{P}(X \in G).$$
5. *For all closed sets $A \subseteq \mathbb{R}$,*
$$\limsup_{n\to\infty} \mathbb{P}(X_n \in A) \leq \mathbb{P}(X \in A).$$
6. *For all Borel sets $C \subseteq \mathbb{R}$ with $\mathbf{P}(X \in \partial C) = 0$,*
$$\lim_{n\to\infty} \mathbb{P}(X_n \in C) = \mathbb{P}(X \in C).$$

Proof. $\boxed{1. \implies 2.}$ Lemma 3.3.
$\boxed{2. \implies 3.}$ Trivial.
$\boxed{3. \implies 1.}$ Lemma 3.4.
$\boxed{4. \iff 5.}$ Exercise.
$\boxed{\text{4. und 5.} \implies 6.}$ For $C \in \mathcal{B}(\mathbb{R})$,

$$\mathbb{P}(X_n \in C^\circ) \leq \mathbb{P}(X_n \in C) \leq \mathbb{P}(X_n \in \overline{C}),$$

so 4. and 5. imply that

$$\mathbb{P}(X \in C^\circ) \le \liminf_{n\to\infty} \mathbb{P}(X_n \in C^\circ) \le \liminf_{n\to\infty} \mathbb{P}(X_n \in C)$$
$$\le \limsup_{n\to\infty} \mathbb{P}(X_n \in C) \le \limsup_{n\to\infty} \mathbb{P}(X_n \in \overline{C}) \le \mathbb{P}(X \in \overline{C}).$$

Since $\partial C = \overline{C}\backslash C^\circ$ and $\mathbb{P}(X \in \partial C) = 0$, it holds $\mathbb{P}(X \in C^\circ) = \mathbb{P}(X \in C) = \mathbb{P}(X \in \overline{C})$, which gives the desired result.

6. $\implies$ 1. Let F_X be continuous at t. Then, $\mathbb{P}(X \in \{t\}) = 0$, and 6. implies that

$$\lim_{n\to\infty} F_{X_n}(t) = \lim_{n\to\infty} \mathbb{P}(X_n \in (-\infty, t]) = \mathbb{P}(X \in (-\infty, t]) = F_X(t).$$

2. $\implies$ 5. Let $A \subseteq \mathbb{R}$ be closed. For all $x \in \mathbb{R}$ define

$$d(x, A) := \inf\{|x - y| \mid y \in A\}.$$

For $m \in \mathbb{N}$, we set

$$G_m := \Big\{x \in \mathbb{R} \;\Big|\; d(x, A) < \frac{1}{m}\Big\}.$$

Then $G_m \downarrow A$, $m \to \infty$. Let now $\varepsilon > 0$. Then there exists m such that

$$\mathbb{P}(X \in G_m) = \mathbf{P}_X(G_m) \le \mathbf{P}_X(A) + \varepsilon = \mathbb{P}(X \in A) + \varepsilon$$

We set $f : \mathbb{R} \to \mathbb{R}, x \mapsto \psi(m\, d(x, A))$ with $\psi(t) = \mathbb{1}_{(-\infty,0]}(t) + (1 - t)\mathbb{1}_{(0,1]}(t)$. For $x \in A$ it holds $d(x, A) = 0$, so $f(x) = 1$. If $x \in G_m^c$, i.e. $x \in A^c$ with $d(x, A) \ge 1/m$, then $f(x) = 0$. So,

$$\mathbb{P}(X_n \in A) = \mathbf{E}(\mathbb{1}_{\{X_n \in A\}} f(X_n)) \le \mathbf{E}(f(X_n)),$$

and

$$\mathbf{E}(f(X)) = \mathbf{E}(\mathbb{1}_{\{X \in G_m\}} f(X)) \le \mathbb{P}(X \in G_m) \le \mathbb{P}(X \in A) + \varepsilon.$$

The function f is continuous and bounded. Due to 2. we obtain $\lim_{n\to\infty} \mathbf{E}(f(X_n)) = \mathbf{E}(f(X))$, which proves the claim. □

Corollary 3.7. *Two random variables X, Y are equal in distribution if and only if for all $f \in \mathcal{C}_c^2$,*

$$\mathbf{E}(f(X)) = \mathbf{E}(f(Y)).$$

Exercise 3.8. Let $g : \mathbb{R} \to \mathbb{R}$ be a continuous function and $(X_n)_{n\in\mathcal{N}}$ sequence of random variables converging in distribution to X. Show that then $(g(X_n))_{n\in\mathbb{N}}$ converges in distribution to $g(X)$.

Theorem 3.9. *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of random variables converging in probability to a random variable X. Then $(X_n)_{n\in\mathbb{N}}$ converges in distribution to X, that is,*

$$X_n \xrightarrow{\mathrm{P}} X \quad \implies \quad X_n \xrightarrow{d} X.$$

Proof. Let $f \in \mathcal{C}_c^2$. Since f is continuous with compact support, f is uniformly continuous. Let $\varepsilon > 0$. Then there exists $\delta > 0$, such that for all $x, y \in \mathbb{R}$ with $|x - y| \le \delta$, it holds $|f(x) - f(y)| < \varepsilon$. We define $A_n = \{|X_n - X| > \delta\}$. Then,

$$|\mathbf{E}(f(X_n)) - \mathbf{E}(f(X))| \le \mathbf{E}(\mathbb{1}_{A_n}|f(X_n) - f(X)|) + \mathbf{E}(\mathbb{1}_{A_n^c}|f(X_n) - f(X)|)$$
$$\le 2M\mathbb{P}(A_n) + \varepsilon\mathbb{P}(A_n^c) \le 2M\mathbb{P}(A_n) + \varepsilon,$$

where $M = \sup_{x\in\mathbb{R}} |f(x)|$. Convergence in probability of $(X_n)_{n\in\mathbb{N}}$ implies that $\mathbb{P}(A_n) \to 0$ as $n \to \infty$. With $\varepsilon \downarrow 0$, the claim follows by Lemma 3.4. □

Remark. In the exercises you will show the following partial converse statement of Theorem 3.9: If $(X_n)_{n\in\mathbb{N}}$ converges in distribution to a constant $c \in \mathbb{R}$, then $(X_n)_{n\in\mathbb{N}}$ also converges in probability to c. In general, convergence in distribution is weaker than convergence in probability, as the following example shows.

Example 3.10. Let $\Omega = [0, 1)$ be equipped with the Borel σ-field and the Lebesgue measure. For $n \in \mathbb{N}$ let

$$A_n = \Big[0, \frac{1}{2^n}\Big) \cup \Big[\frac{2}{2^n}, \frac{3}{2^n}\Big) \cup \cdots \cup \Big[\frac{2^n-2}{2^n}, \frac{2^n-1}{2^n}\Big).$$

and $X_n = \mathbb{1}_{A_n}$. Then $\mathbb{P}(X_n = 1) = \mathbb{P}(X_n = 0) = 1/2$ for all $n \in \mathbb{N}$ and $(X_n)_{n\in\mathbb{N}}$ converges in distribution to X_1. However, for $m \neq n$ and all $\varepsilon > 0$,

$$\mathbb{P}(|X_m - X_n| > \varepsilon) = \mathbb{P}(X_m = 0, X_n = 1) + \mathbb{P}(X_m = 1, X_n = 0) = \frac{1}{2}\cdot\frac{1}{2} + \frac{1}{2}\cdot\frac{1}{2} = \frac{1}{2},$$

that is, $(X_n)_{n\in\mathbb{N}}$ does not converge in probability.

3.3 Characteristic functions

Characteristic functions of random variables (or random vectors) "‘characterise"’ their distribution. They are closely related to the Fourier transform, which is studied in Analysis.

3.3.1 Fourier transform

To define the Fourier transform, we need to define integrals of complex-valued functions. These should be understoof componentwise, that is, for a measure μ on Ω and a measurable function $f\colon \Omega \to \mathbb{C}$, we define

$$\int f \,\mathrm{d}\mu := \int \mathrm{Re} f \,\mathrm{d}\mu + i \int \mathrm{Im} f \,\mathrm{d}\mu,$$

where we require that $\mathrm{Re}f$ and $\mathrm{Im}f$ are integrable.

Definition 3.11. For a Lebesgue integrable function $f\colon \mathbb{R} \to \mathbb{R}$, define

$$\widehat{f}\colon \mathbb{R} \to \mathbb{C}, \quad t \mapsto \widehat{f}(t) = \int f(x) e^{-itx} \,\mathrm{d}x = \int_{-\infty}^{\infty} f(x)\cos(tx)\,\mathrm{d}x - i \int_{-\infty}^{\infty} f(x)\sin(tx)\,\mathrm{d}x.$$

The function $\widehat{f}$ is called *Fourier transform* of f.

It holds

$$|\widehat{f}(t)| \le \int |f(x)e^{-itx}|\,\mathrm{d}x = \int |f|\,\mathrm{d}x < \infty, \tag{3.5}$$

since f is integrable, so the Fourier transform of a Lebesgue integrable function is bounded. It is not necessarily integrable. If this is the case, then the Fourier transform can be inverted. The following theorem is proved in Analysis (see for example *Königsberger, Analysis II*).

Theorem 3.12. *Let $f : \mathbb{R} \to \mathbb{R}$ be a Lebesgue integrable function such that its Fourier transform $\widehat{f}$ is also Lebesgue integrable. Then for all $y \in \mathbb{R}$,*

$$f(y) = \frac{1}{2\pi} \int_{\mathbb{R}} e^{ity} \widehat{f}(t)\,\mathrm{d}t.$$

For our purposes, the following proposition will be useful

Proposition 3.13. *For any $f \in C_c^2$, the Fourier transform $\widehat{f}$ is Lebesgue integrable.*

Proof. Let $K \in \mathbb{R}$ such that $f(x) = 0$ for $|x| \geq K$. Since f is differentiable, we know

$$\begin{aligned}\widehat{f'}(t) &= \int_{-\infty}^{\infty} f'(x)e^{-itx}\,\mathrm{d}x = f(x)e^{-itx}\Big]_{-\infty}^{\infty} + it\int_{-\infty}^{\infty} f(x)e^{-itx}\,\mathrm{d}x \\ &= it\widehat{f}(t)\,,\end{aligned}$$

because $f(x) = 0$ for $|x| \geq K$. Iterating this shows that $\widehat{f''}(t) = -t^2\widehat{f}(t)$. Since f'' is integrable,

$$|\widehat{f''}(t)| \leq c$$

for a constant $c > 0$ due to (3.5). It follows that

$$|\widehat{f}(t)| \leq \min\Big\{\int |f|\,\mathrm{d}x, \frac{c}{t^2}\Big\}$$

so $\widehat{f}$ is integrable on $\mathbb{R}$. □

Corollary 3.14. *For any $f \in C_c^2$,*

$$f(y) = \frac{1}{2\pi}\int_{\mathbb{R}} e^{ity}\widehat{f}(t)\,\mathrm{d}t, \quad y \in \mathbb{R}.$$

Exercise 3.15. Compute the Fourier transform of the following functions.

1. $f = \frac{1}{b-a}\mathbb{1}_{[a,b]}$ for $a < b$.
2. $x \mapsto f(x) = e^{-\lambda|x|}$ for $\lambda > 0$.
3. $x \mapsto f(x) = \frac{\lambda^p}{\Gamma(p)}x^{p-1}e^{-\lambda x}\mathbb{1}_{(0,\infty)}(x)$ for $p > 0$, $\lambda > 0$.
   Recall: $\Gamma(p) = \int_0^\infty y^{p-1}e^{-y}\,\mathrm{d}y$.

3.3.2 Definition and properties of characteristic functions

To define characteristic functions, we need expected values (i.e. integrals) of complex-valued random variables. As in Section 3.3.1, they are defined componentwise, that is, for $X\colon \Omega \to \mathbb{C}$, we set

$$\mathbf{E}(X) := \int \mathrm{Re}X\,\mathrm{d}\mathbb{P} + i\int \mathrm{Im}X\,\mathrm{d}\mathbb{P} = \mathbf{E}(\mathrm{Re}\,X) + i\mathbf{E}(\mathrm{Im}\,X),$$

where we assume that ReX and ImX are integrable.

For independent $\mathbb{C}$–valued random variables X and Y with $\mathbf{E}(|X|) < \infty$ and $\mathbf{E}(|Y|) < \infty$, it holds

$$\mathbf{E}(XY) \;=\; \mathbf{E}(X)\mathbf{E}(Y), \tag{3.6}$$

which follows from the corresponding result for real-valued random variables.

We now define the characteristic function of real-valued random variables $X : \Omega \to \mathbb{R}$.

Definition 3.16. The *characteristic function* of a real-valued random variable $X : \Omega \to \mathbb{R}$ is defined as the function $\varphi_X : \mathbb{R} \to \mathbb{C}$ with

$$\varphi_X(t) := \mathbf{E}(e^{itX}) = \mathbf{E}(\cos(tX)) + i\mathbf{E}(\sin(tX)), \quad t \in \mathbb{R}.$$

Exercise 3.17. Show: If φ is the characteristic function of a random variable, then $\overline{\varphi}$, $|\varphi|^2$ and $\operatorname{Re}\varphi$ are also characteristic functions.
Hint: Construct random variables with the corresponding characteristic functions.

The following theorem summarizes the most important properties of characteristic functions.

Theorem 3.18. (C1) *Let X and Y be independent random variables. Then,*

$$\varphi_{X+Y}(t) = \varphi_X(t)\varphi_Y(t), \quad t \in \mathbb{R}.$$

(C2) *Let X be a random variable. Then its characteristic function satisfies*

$$|\varphi_X(t)| \leq 1 = \varphi_X(0), \quad t \in \mathbb{R}.$$

It is uniformly continuous on $\mathbb{R}$, more precisely,

$$|\varphi_X(t+h) - \varphi_X(t)| \leq \mathbf{E}(\min\{2, |hX|\}) \leq |h|\mathbf{E}(|X|)$$

for all $t, h \in \mathbb{R}$.

(C3) *If $\mathbf{E}(|X|^m) < \infty$ for a $m \in \mathbb{N}$, then φ_X is m times continuously differentiable with*

$$\varphi_X^{(m)}(t) = \frac{d^m}{dt^m}\varphi_X(t) = i^m\mathbf{E}(X^m e^{itX}),$$

so in particular

$$\mathbf{E}(X^m) = i^{-m}\varphi_X^{(m)}(0).$$

(C4) *The characteristic function is* positive definite*, that is, for all $n \in \mathbb{N}$, $t_1, \ldots, t_n \in \mathbb{R}$, $a_1, \ldots, a_n \in \mathbb{C}$,*

$$\sum_{k,l=1}^{n} a_k \bar{a}_l \varphi_X(t_k - t_l) \geq 0.$$

Remark 3.19. *Bochner's Theorem* shows that the converse of (C2) are (C4) true. More precisely, it states that any continuous positive definite function $\varphi : \mathbb{R} \to \mathbb{C}$ with $\varphi(0) = 1$ is the characteristic function of a random variable.

To show properties (C2) and (C3), we use the following lemma.

Lemma 3.20. *For $m \in \mathbb{N}_0$ and $x \in \mathbb{R}$,*

$$\Big| e^{ix} - \sum_{k=0}^{m} \frac{(ix)^k}{k!} \Big| \leq \min\Big\{\frac{2|x|^m}{m!}, \frac{|x|^{m+1}}{(m+1)!}\Big\}.$$

Proof. We set $r_m(x) = e^{ix} - \sum_{k=0}^{m}(ix)^k/(k!)$ and show the claim by induction over m.
Induction start: It holds

$$r_0(x) = e^{ix} - 1 = i\int_0^x e^{iy}dy,$$

so $|r_0(x)| \leq \min\{2, |x|\}$.
Observation: In general,

$$r_{m+1}(x) = i\int_0^x r_m(y)dy \quad \text{und} \quad r_{m+1}(x) = r_m(x) - \frac{(ix)^{m+1}}{(m+1)!}.$$

Induction assumption: For a $m \in \mathbb{N}_0$ let $|r_m(x)| \leq |x|^{m+1}/(m+1)!$ for all $x \in \mathbb{R}$.
Induction step: For $x \geq 0$,

$$|r_{m+1}(x)| \leq \int_0^x |r_m(y)|dy \leq \int_0^x \frac{y^{m+1}}{(m+1)!}dy = \frac{|x|^{m+2}}{(m+2)!}$$

and similarly for $x < 0$. On the other hand,

$$|r_{m+1}(x)| \leq |r_m(x)| + \frac{|x|^{m+1}}{(m+1)!} \leq \frac{2|x|^{m+1}}{(m+1)!}.$$

□

Proof of Theorem 3.18. (C1) follows by the definition of the characteristic function and (3.6). For (C2), we can use that

$$|\varphi_X(t)| = |\mathbf{E}(e^{itX})| \leq \mathbf{E}(|e^{itx}|) = 1 = \varphi_X(0).$$

With Lemma 3.20 for $m = 0$ we obtain

$$\begin{aligned} |\varphi_X(t+h) - \varphi_X(t)| &= |\mathbf{E}(e^{i(t+h)X} - e^{itX})| = |\mathbf{E}(e^{itX}(e^{ihX} - 1))| \\ &\leq \mathbf{E}(|e^{itX}||e^{ihX} - 1|) \leq \mathbf{E}(\min\{2, |hX|\}). \end{aligned}$$

The proof of (C3) is an exercise.
For (C4):

$$\sum_{k,l=1}^{n} a_k\bar{a}_l\varphi_X(t_k - t_l) = \mathbf{E}\Big(\sum_{k,l=1}^{n} a_k\bar{a}_l e^{i(t_k-t_l)X}\Big) = \mathbf{E}\Big(\sum_{k,l=1}^{n} a_k e^{it_kX}\overline{a_l e^{it_lX}}\Big) = \mathbf{E}\Big(\Big|\sum_{k=1}^{n} a_k e^{it_kX}\Big|^2\Big) \geq 0.$$

□

Exercise 3.21. Let the random variable X have characteristic function φ_X.

1. What is the characteristic function of $Y := aX + b$ for $a, b \in \mathbb{R}$.

2. Prove: If the distribution of X is symmetric, i.e. $F_X = F_{-X}$, then φ_X is real-valued and even.

Example 3.22 (Normal distribution)**.** We compute the characteristic function of the Gaussian distribution. Let $X \sim \mathcal{N}(\mu, \sigma^2)$. Then,

$$\begin{aligned}
\varphi_X(t) &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{(x-\mu)^2}{2\sigma^2}} e^{itx} dx = \frac{1}{\sqrt{2\pi}} e^{i\mu t} \int_{-\infty}^{\infty} e^{-y^2/2} e^{it\sigma y} dy \\
&= \frac{1}{\sqrt{2\pi}} e^{i\mu t} \int_{-\infty}^{\infty} e^{-y^2/2} \cos(t\sigma y) dy,
\end{aligned}$$

where we have used the substitution $y = (x-\mu)/\sigma$ and the fact that $e^{-y^2/2}$ is an even function. We set

$$f(t) = \int_{-\infty}^{\infty} e^{-y^2/2} \cos(t\sigma y) dy.$$

Then $f(0) = \sqrt{2\pi}$, and by the monotone convergence theorem

$$\begin{aligned}
\frac{d}{dt} f(t) &= -\sigma \int_{-\infty}^{\infty} e^{-y^2/2} y \sin(t\sigma y) dy \\
&= \sigma \Big[e^{-y^2/2} \sin(t\sigma y) \Big]_{-\infty}^{\infty} - \sigma^2 t \int_{-\infty}^{\infty} e^{-y^2/2} \cos(t\sigma y) dy \\
&= -\sigma^2 t f(t).
\end{aligned}$$

The funciton $f(t) = \sqrt{2\pi} e^{-(\sigma t)^2/2}$ is the solution of this differential equation. So we obtain

$$\varphi_X(t) = e^{i\mu t} e^{-(\sigma t)^2/2} = \exp\Big(i\mu t - \frac{\sigma^2}{2} t^2 \Big).$$

With property (C3) of Theorem 3.18 we can compute all moments of the standard normal distribution $\mathcal{N}(0, \sigma^2)$. Let $X \sim \mathcal{N}(0, \sigma^2)$, then

$$\varphi_X(t) = e^{-(\sigma t)^2/2} = \sum_{n=0}^{\infty} \frac{(-1)^n}{n!} \Big(\frac{\sigma^2 t^2}{2} \Big)^n,$$

so

$$\varphi_X^{(2n)}(0) = \frac{(2n)!(-1)^n \sigma^{2n}}{n! 2^n}$$

and therefore

$$\mathbf{E}(X^{2n}) = \frac{(2n)!}{n! 2^n} \sigma^{2n}.$$

The $(2n-1)$-th moments are all zero due to symmetry.

3.3.3 Uniqueness and weak convergence

Characteristic functions are very helpful to make statements about weak convergence. There is the following theorem.

Theorem 3.23. *Let X and X_n, $n \geq 1$ be random variables. The sequence $(X_n)_{n\in\mathbb{N}}$ converges in distribution to X if and only if*

$$\lim_{n\to\infty} \varphi_{X_n}(t) = \varphi_X(t) \quad \textit{for all } t \in \mathbb{R}.$$

Proof. $\boxed{\Longrightarrow}$ Assume that $X_n \xrightarrow{d} X$. Lemma 3.3 implies that for all $t \in \mathbb{R}$,

$$\lim_{n\to\infty} \mathbf{E}(\cos(tX_n)) = \mathbf{E}(\cos(tX)), \qquad \lim_{n\to\infty} \mathbf{E}(\sin(tX_n)) = \mathbf{E}(\sin(tX)),$$

so $\lim_{n\to\infty} \varphi_{X_n}(t) = \varphi_X(t)$ for all $t \in \mathbb{R}$.
$\boxed{\Longleftarrow}$ Let now $\lim_{n\to\infty} \varphi_{X_n}(t) = \varphi_X(t)$, $t \in \mathbb{R}$. Let $f \in C_c^2$. Corollary 3.14 and Fubini's Theorem yield

$$\int f \,\mathrm{d}\mathbf{P}_X = \frac{1}{2\pi} \int\int e^{ixy} \hat{f}(x)\,\mathrm{d}x\,\mathrm{d}\mathbf{P}_X(y) = \frac{1}{2\pi}\int\int e^{ixy}\,\mathrm{d}\mathbf{P}_X(y)\hat{f}(x)\,\mathrm{d}x = \frac{1}{2\pi}\int \varphi_X(x)\hat{f}(x)\,\mathrm{d}x,$$

so

$$\int f\,\mathrm{d}\mathbf{P}_{X_n} = \frac{1}{2\pi}\int \varphi_{X_n}\hat{f}\,\mathrm{d}x, \quad n \in \mathbb{N}, \quad \text{und} \quad \int f\,\mathrm{d}\mathbf{P}_X = \frac{1}{2\pi}\int \varphi_X \hat{f}\,\mathrm{d}x.$$

It holds $|\varphi_{X_n}\hat{f}| \leq |\hat{f}|$, and $\hat{f}$ is Lebesgue integrable due to Proposition 3.13, so we can apply the Lebesgue convergence theorem and obtain

$$\lim_{n\to\infty}\int \varphi_{X_n}\hat{f}\,\mathrm{d}x = \int \varphi_X\hat{f}\,\mathrm{d}x,$$

so

$$\lim_{n\to\infty}\int f\,\mathrm{d}\mathbf{P}_{X_n} = \int f\,\mathrm{d}\mathbf{P}_X$$

and the claim follows by Lemma 3.4. □

The essential property of characteristic functions is that the distribution $\mathbf{P}_X$ of a random variable X is defined uniquely by φ_X.

Corollary 3.24. *Random variables X and Y are equal in distribution, that is $\mathbf{P}_X = \mathbf{P}_Y$, if and only if*

$$\varphi_X(t) = \varphi_Y(t) \quad \textit{for all } t \in \mathbb{R}.$$

Proof. Follows from Theorem 3.23. □

3.3.4 Examples of characteristic functions

If X admits a discrete distribution with values $\{x_j\}_{j\in\mathbb{N}}$, then

$$\varphi_X(t) = \sum_j e^{itx_j}\mathbb{P}(X = x_j).$$

If X is absolutely continuous with density f, then

$$\varphi_X(t) = \int_{\mathbb{R}} f(x)e^{itx}dx.$$

The following table gives a summary of important characteristic functions.

Distribution	Notation	characteristic function $\varphi_X(t)$
Binomial	$\mathrm{BIN}(n,p)$	$(pe^{it}+1-p)^n$
Poisson	$\mathrm{POIS}(\lambda)$	$\exp(\lambda(e^{it}-1))$
Geometric	$\mathrm{GEOM}(p)$	$p/(1-(1-p)e^{it})$
Uniform	$\mathrm{UNIF}(a,b)$	$(e^{itb}-e^{ita})/(it(b-a))$
Normal	$\mathcal{N}(\mu,\sigma^2)$	$\exp(i\mu t-(t\sigma)^2/2)$
Gamma	$\mathrm{Gamma}(p,\lambda)$	$(1-(it)/\lambda)^{-p}$
Cauchy		$\exp(-\lvert t\rvert)$

Exercise 3.25. Let $\lambda \in (0,1)$ and let $(X_n)_{n\in\mathbb{N}}$ be a sequence of random variables with $X_n \sim \mathrm{BIN}(n,p_n)$, where $p_n = \lambda/n$. Use characteristic functions and Theorem 3.23 to show that $X_n \xrightarrow{d} X$, where $X \sim \mathrm{POIS}(\lambda)$.

We have seen in example 3.22 that the characteristic function is useful for computing moments. On the other hands, one may try to compute the characteristic function (and thereby the distribution function) by using moments. This is only possible if the Taylor series

$$\sum_{k=0}^{\infty} \frac{t^k}{k!}\varphi_X^{(k)}(0)$$

converges in a neighborhood of 0. As the following example shows, this is not always the case.

Example 3.26. The random variable $X > 0$ admits a *Lognormal distribution* if $\log X$ is normally distributed. If $\log X \sim \mathcal{N}(0,1)$, one can easily show that the density of X is given by

$$f(x) = \frac{1}{\sqrt{2\pi}} x^{-1} e^{-(\log x)^2/2}, \quad x > 0.$$

We show that the lognormal distribution is not uniquely defined by its moments. Define

$$f_a(x) = f(x)(1 + a\sin(2\pi \log x))$$

for $a \in [-1,1]$. Then $f_a(x) \geq 0$. We now show that f_a is a density function with moments equal to the ones of f. For this, it is sufficient to show that

$$\int_0^\infty x^k f(x)\sin(2\pi\log x)dx = 0, \quad k = 0,1,\ldots.$$

It is easy to verify that

$$\begin{aligned}\int_0^\infty x^k f(x)\sin(2\pi\log x)dx &\overset{\log x = t}{=} \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{1}{2}t^2+kt}\sin(2\pi t)dt \\ &\overset{t=k-y}{=} \frac{1}{\sqrt{2\pi}} e^{\frac{1}{2}k^2}\int_{-\infty}^{\infty} e^{-\frac{1}{2}y^2}\sin(2\pi y)dy.\end{aligned}$$

The last integrand is an uneven function, so the integral equals zero.

3.4 The central limit theorem

3.4.1 Identically distributed random variables

Theorem 3.27 (Central limit theorem)**.** *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent, identically distributed and square integrable random variables with $\mu = \mathbf{E}(X_1)$ and $\sigma^2 = \mathsf{Var}(X_1) < \infty$. Then,*

$$\mathbb{P}\left(\frac{\sqrt{n}}{\sigma}\Big(\frac{S_n}{n} - \mu\Big) \le t\right) \xrightarrow{n\to\infty} \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{t} e^{-x^2/2}dx = \Phi(t) \quad \text{for all } t \in \mathbb{R},$$

where $S_n = \sum_{k=1}^n X_k$, or in other words,

$$\frac{S_n - n\mu}{\sqrt{n}\sigma} \xrightarrow{d} \mathcal{N}(0,1).$$

Proof. We set $Y_n = (X_n - \mu)/\sigma$. Then

$$\varphi'_{Y_1}(0) = i\mathbf{E}(Y_1) = 0,$$
$$\varphi''_{Y_1}(0) = -\mathbf{E}(Y_1^2) = -\mathsf{Var}(Y_1) = -1.$$

Due to Lemma 3.20 with $m = 2$,

$$|e^{itY_1} - 1 - itY_1 + \frac{1}{2}t^2Y_1^2| \le \min\{|tY_1|^2, |tY_1|^3/6\},$$

so

$$t^{-2}(e^{itY_1} - 1 - itY_1 + \frac{1}{2}t^2Y_1^2) \to 0, \quad t \to 0.$$

Furthermore,

$$t^{-2}|e^{itY_1} - 1 - itY_1 + \frac{1}{2}t^2Y_1^2| \le |Y_1|^2$$

so the Lebesgue convergence theorem with dominating function $|Y_1|^2$ implies that

$$t^{-2}\mathbf{E}\Big(e^{itY_1} - 1 - itY_1 + \frac{1}{2}t^2Y_1^2\Big) \to 0, \quad t \to 0,$$

and therefore

$$\varphi_{Y_1}(t) = 1 - \frac{1}{2}t^2(1 + \alpha(t)),$$

with a function α such that $\alpha(t) \to 0$ as $t \to 0$.

Let $\tilde{S}_n = \sum_{k=1}^n Y_n$. Then $(\sqrt{n}/\sigma)(S_n/n - \mu) = \tilde{S}_n/\sqrt{n}$ and we obtain by Theorem 3.18 (C1) that

$$\begin{aligned}\varphi_{\tilde{S}_n/\sqrt{n}}(t) &= \varphi_{\tilde{S}_n}(t/\sqrt{n}) = \big(\varphi_{Y_1}(t/\sqrt{n})\big)^n \\ &= \left(1 - \frac{1}{2}\frac{t^2}{n}\big(1 + \alpha(t/\sqrt{n})\big)\right)^n \to e^{-\frac{1}{2}t^2}, \quad n \to \infty.\end{aligned}$$

The last convergence is due to the following Lemma 3.28, which is an exercise. The limit is exactly the characteristic function of the $\mathcal{N}(0,1)$ distribution, see Example 3.22. □

Lemma 3.28. *Let $(a_n)_{n\in\mathbb{N}}$ be a sequence of complex numbers with $a_n \to 0$, $n \to \infty$, $x \in \mathbb{R}$. Then,*

$$\left(1 + \frac{x + a_n}{n}\right)^n \to e^x.$$

Exercise 3.29. Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent and uniformly distributed random variables on $[0,2]$. Find an approximation for the probability $\mathbb{P}(X_1 + \cdots + X_{200} > 110)$.

Corollary 3.30 (Theorem of de Moivre-Laplace). *Assume that S_n is $\mathrm{BIN}(n,p)$-distributed with $p \in (0,1)$. Then for $-\infty \le t_1 \le t_2 \le \infty$*

$$\mathbb{P}\left(t_1 \le \frac{S_n - np}{\sqrt{np(1-p)}} \le t_2\right) \to \frac{1}{\sqrt{2\pi}} \int_{t_1}^{t_2} e^{-x^2/2} dx = \Phi(t_2) - \Phi(t_1).$$

The Theorem of de Moivre-Laplace is useful to approximate the Binomial distribution with large n by the Gaussian distribution.

3.4.2 Lindeberg Theorem

The central limit theorem can be proved under weaker conditions. The random variables $(X_n)_{n\in\mathbb{N}}$ are still independent, but not necessarily identically distributed. It is only important that there is no X_{n_0} that *outweighs* the others, that is, $\mathsf{Var}(X_{n_0})$ should be small compared to $\mathsf{Var}(S_n) = \sum_{k=1}^n \mathsf{Var}(X_k)$.

We use the following notation:

$$S_n := \sum_{k=1}^n X_n,$$
$$\sigma_n^2 := \mathsf{Var}(X_n),$$
$$s_n^2 := \mathsf{Var}(S_n) = \sum_{k=1}^n \sigma_k^2.$$

Theorem 3.31 (Lindeberg Theorem). *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent square integrable random variables with $\sigma_n^2 = \mathsf{Var}(X_n) > 0$, which satisfies the* Lindeberg condition

$$\lim_{n\to\infty} L_n(c) = 0, \quad \text{für alle } c > 0,$$

where

$$L_n(c) := \frac{1}{s_n^2} \sum_{k=1}^n \mathbf{E}\big((X_k - \mathbf{E}(X_k))^2 \mathbb{1}_{\{|X_k - \mathbf{E}(X_k)| > c\, s_n\}}\big)$$

with $S_n = \sum_{k=1}^n X_k$, $s_n^2 = \mathsf{Var}(S_n) = \sum_{k=1}^n \sigma_k^2$. Then

$$\frac{S_n - \mathbf{E}(S_n)}{s_n} \xrightarrow{d} \mathcal{N}(0,1), \quad n \to \infty.$$

For the proof, we need the following auxiliary results.

Lemma 3.32. *The Lindeberg condition implies*

$$\lim_{n\to\infty} \max_{k\in\{1,\dots,n\}} \frac{\sigma_k^2}{s_n^2} = 0.$$

Proof. Let $c > 0$. Then with $\tilde{X}_k := X_k - \mathbf{E}(X_k)$

$$\sigma_k^2 = \mathbf{E}(\tilde{X}_k^2) = \mathbf{E}(\mathbb{1}_{\{|\tilde{X}_k| \le c\, s_n\}} \tilde{X}_k^2) + \mathbf{E}(\mathbb{1}_{\{|\tilde{X}_k| > c\, s_n\}} \tilde{X}_k^2) \le c^2\, s_n^2 + L_n(c)\, s_n^2,$$

so

$$\max_{1 \le k \le n} \frac{\sigma_k^2}{s_n^2} \le c^2 + L_n(c).$$

By letting first $n \to \infty$ and then $c \downarrow 0$, the desired result follows. □

Lemma 3.33. *For complex numbers $a_1, \dots, a_n, b_1, \dots, b_n \in \mathbb{C}$ with $|a_k| \le 1$, $|b_k| \le 1$, $k = 1, \dots, n$,*

$$\Big| \prod_{k=1}^n a_k - \prod_{k=1}^n b_k \Big| \le \sum_{k=1}^n |a_k - b_k|.$$

Proof. Exercise. □

We set

$$X_{nk} := \frac{1}{s_n}(X_k - \mathbf{E}(X_k)), \quad \sigma_{nk}^2 := \mathsf{Var}(X_{nk}) = \frac{\sigma_k^2}{s_n^2}$$

Let φ_{nk} be the characteristic function of X_{nk} and φ_n the characteristic function of $(S_n - \mathbf{E}(S_n))/s_n$. It holds

$$\frac{S_n - \mathbf{E}(S_n)}{s_n} = \sum_{k=1}^n X_{nk},$$

so by Theorem 3.18 (C1),

$$\varphi_n(t) = \prod_{k=1}^n \varphi_{nk}(t).$$

Proof of Theorem 3.31. We show that for all $t \in \mathbb{R}$,

$$|\varphi_n(t) - e^{-t^2/2}| \to 0, \quad n \to \infty. \tag{3.7}$$

The theorem then follows by Theorem 3.23.

Due to $\sum_{k=1}^n \sigma_{nk}^2 = 1$, one can use Lemma 3.33 to obtain

$$\begin{aligned}
|\varphi_n(t) - e^{-t^2/2}| &= \Big| \prod_{k=1}^n \varphi_{nk}(t) - \prod_{k=1}^n e^{-\sigma_{nk}^2 t^2/2} \Big| \\
&\le \sum_{k=1}^n \big| \varphi_{nk}(t) - e^{-\sigma_{nk}^2 t^2/2} \big| \\
&\le \underbrace{\sum_{k=1}^n \Big| \varphi_{nk}(t) - 1 + \frac{\sigma_{nk}^2 t^2}{2} \Big|}_{=:A} + \underbrace{\sum_{k=1}^n \Big| e^{-\sigma_{nk}^2 t^2/2} - 1 + \frac{\sigma_{nk}^2 t^2}{2} \Big|}_{=:B}
\end{aligned}$$

We first consider A. Let $c > 0$. Lemma 3.20 with $m = 2$ shows that

$$\begin{aligned}
\sum_{k=1}^{n}\Big|\varphi_{nk}(t) - 1 + \frac{\sigma_{nk}^2 t^2}{2}\Big| &\le \sum_{k=1}^{n} t^2 \mathbf{E}\big(\min\{X_{nk}^2, t|X_{nk}|^3/6\}\big)\\
&= \sum_{k=1}^{n} t^2\Big(\mathbf{E}\big(\mathbb{1}_{\{|X_{nk}|\le c\}}\min\{X_{nk}^2, t|X_{nk}|^3/6\}\big) + \mathbf{E}\big(\mathbb{1}_{\{|X_{nk}|>c\}}\min\{X_{nk}^2, t|X_{nk}|^3/6\}\big)\Big)\\
&\le \frac{t^3}{6}\sum_{k=1}^{n}\mathbf{E}(\mathbb{1}_{\{|X_{nk}|\le c\}}|X_{nk}|^3) + t^2\sum_{k=1}^{n}\mathbf{E}(\mathbb{1}_{\{|X_{nk}|>c\}}X_{nk}^2)\\
&\le \frac{t^3}{6}c\underbrace{\sum_{k=1}^{n}\mathbf{E}(X_{nk}^2)}_{=1} + t^2 L_n(c).
\end{aligned}$$

By letting first $n \to \infty$ and then $c \downarrow 0$, it follows that $A \to 0$.

For B, we have

$$\sum_{k=1}^{n}\Big|e^{-\sigma_{nk}^2 t^2/2} - 1 + \frac{\sigma_{nk}^2 t^2}{2}\Big| \le \sum_{k=1}^{n}\frac{\sigma_{nk}^4 t^4}{4}e^{\sigma_{nk}^2 t^2/2} \le \frac{t^4}{4}e^{(t^2/2)\max_{1\le k\le n}\sigma_{nk}^2}\max_{1\le k\le n}\sigma_{nk}^2 \to 0, \quad n\to\infty,$$

because of Lemma 3.32. In the first inequality in the formula above, we have used that for $x \in \mathbb{R}$,

$$\Big|e^x - 1 - x\Big| = \Big|\sum_{k=2}^{\infty}\frac{x^k}{k!}\Big| \le |x|^2\sum_{k=2}^{\infty}\frac{|x|^{k-2}}{k!} = |x|^2\sum_{k=0}^{\infty}\frac{|x|^k}{(k+2)!} \le |x|^2 e^{|x|}.$$

□

Exercise 3.34. Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent random variables with $|X_n| \le n^{1/3}$ and $\mathsf{Var}(X_n) = 1$ for all $n \in \mathbb{N}$. Is the Lindeberg condition satisfied?

The following theorem is weaker than the Lindeberg Theorem, but it uses a simpler assumption (Lyapunov condition).

Theorem 3.35 (Lyapunov Theorem)**.** *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent square integrable random variables with $\mathsf{Var}(X_n) > 0$ and $S_n = \sum_{k=1}^n X_k$. If for some $\delta > 0$ the* Lyapunov condition *is satisfied, that is,*

$$\overline{L}_n(\delta) = \frac{1}{\mathsf{Var}(S_n)^{(2+\delta)/2}}\sum_{k=1}^{n}\mathbf{E}(|X_k - \mathbf{E}X_k|^{2+\delta}) \to 0, \quad n\to\infty,$$

then

$$\frac{S_n - \mathbf{E}(S_n)}{\sqrt{\mathsf{Var}(S_n)}} \xrightarrow{d} \mathcal{N}(0,1), \quad n\to\infty.$$

Proof. Exercise. □

Exercise 3.36. Let $(X_n)_{n\in\mathbb{N}}$ be a uniformly bounded sequence of random variables, that is, there exists $c > 0$ such that $|X_n| \le c$ for all $n \in \mathbb{N}$, and we assume that $\lim_{n\to\infty} s_n = \infty$. Show that the Lyapunov condition is satisfied for all $\delta > 0$

3.4.3 Multidimensional generalizations

There are central limit theorems also for random vectors.

Definition 3.37. A sequence $(X_n)_{n\in\mathbb{N}}$ of d-dimensional random vectors *converges in distribution* to the d-dimensional random vector X if

$$\mathbf{E}(f(X_n)) \to \mathbf{E}(f(X)), \quad n \to \infty$$

for all continuous bounded functions $f\colon \mathbb{R}^d \to \mathbb{R}$.

Remark 3.38. According to Lemma 3.3 and 3.4, the above definition is equivalent to Definition 3.1 in the case $d = 1$.

One defines the *characteristic function* o the random vector $X = (X_1, \dots, X_d)^\top$ by

$$\varphi_X\colon \mathbb{R}^d \to \mathbb{C}, \quad t = (t_1, \dots, t_d)^\top \mapsto \varphi_X(t) := \mathbf{E}[e^{i\langle t, X\rangle}] = \mathbf{E}\Big[\exp\Big(i\sum_{k=1}^{d} t_k X_k\Big)\Big].$$

All properties of Theorem 3.18 can be given in a multivariate setting, and Theorems 3.24 und 3.23 continue to hold. Also the arguments of chapter 3.4 are applicable.

Chapter 4

Conditional expectations

Let $(\Omega, \mathfrak{F}, \mathbb{P})$ be a probability space.

4.1 Definition

Let $\mathcal{A} \subseteq \mathfrak{F}$ be a (sub)-σ-field. A random variable $X : \Omega \to \mathbb{R}$ is called $\mathcal{A}$-measurable (or measurable with respect to $\mathcal{A}$), if

$$X^{-1}(B) \in \mathcal{A}, \quad \text{for all } B \in \mathcal{B}(\mathbb{R}).$$

For example, the constant function $Y : \Omega \to \mathbb{R}, \omega \mapsto c$ for a $c \in \mathbb{R}$ is measurable with respect to the trivial σ-field $\{\emptyset, \Omega\} \subseteq \mathfrak{F}$, because

$$Y^{-1}(B) = \begin{cases} \Omega, & \text{if } c \in B \\ \emptyset, & \text{if } c \notin B. \end{cases}$$

Let now X be an arbitrary integrable random variable. We can interpret its expectation $\mathbf{E}(X)$ as a constant mapping $Y : \Omega \to \mathbb{R}, \omega \mapsto \mathbf{E}(X)$, which is $\{\emptyset, \Omega\}$-measurable.

For a set $A \in \mathfrak{F}$ with $\mathbb{P}(A) \in (0,1)$, the conditional probability on A is the probability measure

$$\mathbb{P}(\cdot|A) : \mathfrak{F} \to [0,1], \quad B \mapsto \mathbb{P}(B|A) := \frac{\mathbb{P}(B \cap A)}{\mathbb{P}(A)}.$$

Now, the conditional expectation of X with respect to A can be defined as

$$\mathbf{E}(X|A) := \int X(\omega)\mathbb{P}(\,\mathrm{d}\omega|A).$$

One can show that

$$\mathbf{E}(X|A) = \frac{\mathbf{E}(X\mathbb{1}_A)}{\mathbb{P}(A)}.$$

It holds

$$\mathbf{E}(X|A^c) = \frac{\mathbf{E}(X) - \mathbf{E}(X\mathbb{1}_A)}{1 - \mathbb{P}(A)} = \frac{\mathbf{E}(X) - \mathbb{P}(A)\mathbf{E}(X|A)}{1 - \mathbb{P}(A)}.$$

We set

$$Y : \Omega \to \mathbb{R}, \quad \omega \mapsto \mathbb{1}_A(\omega)\mathbf{E}(X|A) + \mathbb{1}_{A^c}(\omega)\mathbf{E}(X|A^c).$$

Let $\mathcal{A} = \sigma(\{A\}) = \{A, A^c, \emptyset, \Omega\} \subseteq \mathfrak{F}$. The random variable Y is $\mathcal{A}$-measurable. It holds $\mathbf{E}(Y) = \mathbf{E}(X)$ and Y is denoted by $\mathbf{E}(X|\mathcal{A})$. Note that here, the element in the condition is the σ-field $\mathcal{A}$, which must not be confused with the event A.

We can interpret σ-fields as "information sets". The smaller the σ-field with respect to which a random variable X is measurable, the less information it encodes.

We will now generalize the concept described above to arbitrary σ-fields $\mathcal{A} \subseteq \mathfrak{F}$. For that purpose, we need the following theorem.

Theorem 4.1. *Let X be a non-negative (integrable) random variable on $(\Omega, \mathfrak{F}, \mathbb{P})$, and let $\mathcal{A} \subseteq \mathfrak{F}$ be a σ-field. Then there exists a non-negative (integrable) $\mathcal{A}$-measurable random variable Y such that*

$$\mathbf{E}(\mathbb{1}_C Y) = \mathbf{E}(\mathbb{1}_C X), \quad \textit{for all } C \in \mathcal{A}. \tag{4.1}$$

The random variable Y is unique almost surely.

Proof. Let first X be non-negative, i.e. $X \geq 0$.

Let P_0 be the restriction of the probability measure $\mathbb{P}$ to $\mathcal{A}$. We define the measure Q as

$$Q : \mathcal{A} \to [0, \infty], \quad C \mapsto \mathbf{E}(\mathbb{1}_C X) = \int_C X \, \mathrm{d}\mathbb{P}.$$

It follows from this definition that $P_0(C) = \mathbb{P}(C) = 0$ for a $C \in \mathcal{A}$ implies that $Q(C) = 0$, i.e. Q is absolutely continuous with respect to P_0. The Radon-Nikodym Theorem then shows that Q admits a density with respect to P_0, that is, a $\mathcal{A}$-measurable function $Y \geq 0$ such that

$$\int_C Y \, \mathrm{d}P_0 = \int_C X \, \mathrm{d}\mathbb{P}, \text{for all } C \in \mathcal{A}.$$

Since Y is $\mathcal{A}$-measurable, it follows by definition of the integral with respect to a measure that $\int_C Y \, \mathrm{d}P_0 = \int_C Y \, \mathrm{d}\mathbb{P}$ for all $C \in \mathcal{A}$. Theorem 3.1 in the lecture notes on measure theory implies that Y is P_0 almost surely unique. Let Y' be another density of Q with respect to P_0. Because Y, Y' are $\mathcal{A}$-measurable, it follows that $\{Y = Y'\} \in \mathcal{A}$, so $1 = P_0(Y = Y') = \mathbb{P}(Y = Y')$ and Y is $\mathbb{P}$ almost surely unique.

Let now X be an integrable random variable. We decompose it into positive and negative part, $X = X_+ - X_-$. The first part of the theorem shows that there exist $\mathcal{A}$-measurable random variables $Y_1 \geq 0$ and $Y_2 \geq 0$ with

$$\mathbf{E}(\mathbb{1}_C Y_1) = \mathbf{E}(\mathbb{1}_C X_+), \quad \mathbf{E}(\mathbb{1}_C Y_2) = \mathbf{E}(\mathbb{1}_C X_-) \quad \text{for all } C \in \mathcal{A}.$$

Choosing $C = \Omega$ yields that Y_1, Y_2 are integrable. So $Y = Y_1 - Y_2$ is integrable and satisfies (4.1). Let Y' be another $\mathcal{A}$-measurable integrable solution of (4.1). Then

$$\mathbf{E}(\mathbb{1}_C X) = \mathbf{E}(\mathbb{1}_C Y_1) - \mathbf{E}(\mathbb{1}_C Y_2) = \mathbf{E}(\mathbb{1}_C Y'_+) - \mathbf{E}(\mathbb{1}_C Y'_-),$$

so

$$\mathbf{E}(\mathbb{1}_C (Y_1 + Y'_-)) = \mathbf{E}(\mathbb{1}_C (Y_2 + Y'_+)).$$

The $\mathcal{A}$-measurability of $Y_1 + Y'_-$, $Y_2 + Y'_+$ implies that both are densities of the measure

$$\mathcal{A} \to [0, \infty), \quad C \mapsto \mathbf{E}(\mathbb{1}_C (Y_1 + Y'_-)) = \mathbf{E}(\mathbb{1}_C (Y_2 + Y'_+)).$$

But then Theorem 3.1 in the lecture notes on measure theory implies that $Y_1 + Y'_- = Y_2 + Y'_+$ P_0 almost surely, so $\mathbb{P}$ almost surely with the same argument as above. $\square$

Definition 4.2. Let X be a non-negative or integrable random variable and $\mathcal{A} \subseteq \mathfrak{F}$ be a σ-field. Then the random variable Y from Theorem 4.1 is called the *conditional expectation of X with respect to $\mathcal{A}$*, in symbols:

$$\mathbf{E}(X|\mathcal{A}) := Y.$$

Hence the conditional expectation $\mathbf{E}(X|\mathcal{A})$ of X with respect to $\mathcal{A}$ is a $\mathcal{A}$ measurable random variable with

$$\mathbf{E}(\mathbf{E}(X|\mathcal{A})\mathbb{1}_C) = \mathbf{E}(X\mathbb{1}_C), \quad \text{for all } C \in \mathcal{A}. \tag{4.2}$$

The conditional expectation $\mathbf{E}(X|\mathcal{A})$ is only $\mathbb{P}$ almost surely unique. One says that there are different, but $\mathbb{P}$ almost surely equal *versions* of the conditional expectation.

Proposition 4.3. *If X is integrable and almost surely non-negative, then $\mathbf{E}(X|\mathcal{A}) \geq 0$ almost surely.*

Proof. For all $C \in \mathcal{A}$,

$$0 \leq \mathbf{E}(X\mathbb{1}_C) = \mathbf{E}(\mathbf{E}(X|\mathcal{A})\mathbb{1}_C).$$

Since $\mathbf{E}(X|\mathcal{A})$ is $\mathcal{A}$ measurable, it holds that $C_0 := \{\mathbf{E}(X|\mathcal{A}) < 0\} \in \mathcal{A}$, so

$$0 \leq \mathbf{E}(\mathbf{E}(X|\mathcal{A})\mathbb{1}_{C_0}) \leq 0,$$

and hence $\mathbb{P}(C_0) = 0$, because is negative $\mathbf{E}(X|\mathcal{A})$ on C_0. $\square$

Exercise 4.4. Let $(\Omega, \mathfrak{F}, \mathbb{P})$ be a probability space with $\Omega = \{0,1\}^2$, $\mathfrak{F} = 2^\Omega$ and $\mathbb{P}(\{\omega\}) = 1/4$ for all $\omega \in \Omega$. We define random variables X and Y on Ω by $X(\omega) = \omega_1 + \omega_2$ and $Y(\omega) = \omega_1$. Show that for $\mathcal{A} = \sigma(X)$,

$$\mathbb{E}(Y|\mathcal{A}) = \mathbb{1}_{\{(1,1)\}} + \frac{1}{2}\mathbb{1}_{\{(0,1),(1,0)\}}.$$

4.2 Properties

We now derive properties of conditional expectations. Let X, Y be integrable random variables and $\mathcal{A} \subseteq \mathfrak{F}$ a σ-field.

The following statements are direct consequences of Theorem 4.1.

(A) $\mathbf{E}(\mathbf{E}(X|\mathcal{A})) = \mathbf{E}(X)$

(B) X $\mathcal{A}$-measurable $\implies$ $\mathbf{E}(X|\mathcal{A}) = X$ almost surely

(C) $X = Y$ almost surely $\implies$ $\mathbf{E}(X|\mathcal{A}) = \mathbf{E}(Y|\mathcal{A})$ almost surely

(D) $\mathbf{E}(\alpha X + \beta Y|\mathcal{A}) = \alpha\mathbf{E}(X|\mathcal{A}) + \beta\mathbf{E}(Y|\mathcal{A})$ for all $\alpha, \beta \in \mathbb{R}$

As with the usual expected value, monotonicity is also preserved under conditional expectations.

(E) $X \leq Y$ almost surely $\implies$ $\mathbf{E}(X|\mathcal{A}) \leq \mathbf{E}(Y|\mathcal{A})$ almost surely

(F) $|\mathbf{E}(X|\mathcal{A})| \leq \mathbf{E}(|X|\,|\mathcal{A})$ almost surely

Proof. **(E)** Because $Z = Y - X \geq 0$ almost surely, Theorem 4.3 implies that $\mathbf{E}(Z|\mathcal{A}) \geq 0$ almost surely. Now we can apply **(D)**.
(F) Clearly, $X \leq |X|$ and $-X \leq |X|$, so due to **(E)** (and **(D)**),

$$\max\{\mathbf{E}(X|\mathcal{A}), \mathbf{E}(-X|\mathcal{A})\} = |\mathbf{E}(X|\mathcal{A})| \leq \mathbf{E}(|X|\,|\mathcal{A}) \quad \text{almost surely.}$$

□

Also the monotone convergence theorem and Lebesgue's theorem of dominated convergence hold for conditional expectations.

(G) Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of random variables $X_n \geq 0$ with $X_n \uparrow X$. Then

$$\lim_{n\to\infty} \mathbf{E}(X_n|\mathcal{A}) = \mathbf{E}(\lim_{n\to\infty} X_n|\mathcal{A}) = \mathbf{E}(X|\mathcal{A}) \quad \text{fast sicher.}$$

(H) Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of random variables with $X_n \xrightarrow{\text{a.s.}} X$, and assume that there exists Y such that $|X_n| \leq Y$ for all $n \in \mathbb{N}$. Then,

$$\lim_{n\to\infty} \mathbf{E}(X_n|\mathcal{A}) = \mathbf{E}(X|\mathcal{A}) \quad \text{almost surely}$$

and

$$\mathbf{E}(|X_n - X|\,|\mathcal{A}) \to 0, \quad \text{almost surely.}$$

Proof. **(G)** Due to **(C)** and **(E)** we can assume that $\mathbf{E}(X_n|\mathcal{A}) \leq \mathbf{E}(X_{n+1}|\mathcal{A})$ for all n on Ω. (Note: Countable unions of null sets are again null sets!) Let now $C \in \mathcal{A}$. Then the monotone convergence theorem implies that

$$\mathbf{E}(\mathbb{1}_C\mathbf{E}(X|\mathcal{A})) = \lim_{n\to\infty} \mathbf{E}(\mathbb{1}_C\mathbf{E}(X_n|\mathcal{A})) = \lim_{n\to\infty} \mathbf{E}(\mathbb{1}_C X_n) = \mathbf{E}(\mathbb{1}_C X),$$

which yields the claim, because $\lim_{n\to\infty} \mathbf{E}(X_n|\mathcal{A})$ is $\mathcal{A}$-measurable.
(H) We define $Z_n := \sup_{m\geq n} |X_m - X|$. Since $X_n \xrightarrow{\text{a.s.}} X$, we obtain $Z_n \downarrow 0$ almost surely. Because X_n and X are integrable, **(D)**, **(F)** and **(E)** imply that

$$|\mathbf{E}(X_n|\mathcal{A}) - \mathbf{E}(X|\mathcal{A})| = |\mathbf{E}(X_n - X|\mathcal{A})| \leq \mathbf{E}(|X_n - X||\mathcal{A}) \leq \mathbf{E}(Z_n|\mathcal{A}), \quad \text{almost surely.}$$

Because $0 \leq \mathbf{E}(Z_{n+1}|\mathcal{A}) \leq \mathbf{E}(Z_n|\mathcal{A})$ almost surely for all n, the limit $\zeta = \lim_{n\to\infty} \mathbf{E}(Z_n|\mathcal{A})$ exists almost surely. Due to **(A)**,

$$0 \leq \mathbf{E}(\zeta) \leq \mathbf{E}(\mathbf{E}(Z_n|\mathcal{A})) = \mathbf{E}(Z_n) \to 0, \quad n \to \infty.$$

The last convergence statement follows by the dominated convergence theorem, because $0 \leq Z_n \leq 2Y$. So $\mathbf{E}(\zeta) = 0$ and hence $\zeta = 0$ almost surely □

(I) Let Y be a $\mathcal{A}$-measurable random variable, X an integrable random variable and YX integrable. Then,

$$\mathbf{E}(YX|\mathcal{A}) = Y\mathbf{E}(X|\mathcal{A}) \quad \text{almost surely.} \tag{4.3}$$

Proof. If $Y = \mathbb{1}_B$ for $B \in \mathcal{A}$, then for arbitrary $C \in \mathcal{A}$,

$$\mathbf{E}(\mathbb{1}_C YX) = \mathbf{E}(\mathbb{1}_{C\cap B}X) = \mathbf{E}(\mathbb{1}_{C\cap B}\mathbf{E}(X|\mathcal{A})) = \mathbf{E}(\mathbb{1}_C Y\mathbf{E}(X|\mathcal{A}))$$

Because $Y\mathbf{E}(X|\mathcal{A})$ is $\mathcal{A}$-measurable, this implies (4.3) for indicator variables Y. The same argument gives (4.3) also for simple random variables Y.
Let now $Y \geq 0$ be a $\mathcal{A}$-measurable random variable such that XY is integrable. There exist simple $\mathcal{A}$-measurable random variables with $Y_n \uparrow Y$ almost surely. For all Y_n, the result proved above implies that

$$\mathbf{E}(Y_n X|\mathcal{A}) = Y_n\mathbf{E}(X|\mathcal{A}) \quad \text{almost surely.}$$

It holds $|Y_n X| \leq |YX|$. Since YX is integrable, it follows by **(H)** that $\mathbf{E}(Y_n X|\mathcal{A}) \to \mathbf{E}(YX|\mathcal{A})$ almost surely. Moreover, $Y_n\mathbf{E}(X|\mathcal{A}) \to Y\mathbf{E}(X|\mathcal{A})$, which yields the claim. General random variables Y are decomposed into positive and negative part. □

(J) For σ-fields $\mathcal{A}_1 \subseteq \mathcal{A}_2 \subseteq \mathfrak{F}$, it holds

$$\mathbf{E}(\mathbf{E}(X|\mathcal{A}_1)|\mathcal{A}_2) = \mathbf{E}(\mathbf{E}(X|\mathcal{A}_2)|\mathcal{A}_1) = \mathbf{E}(X|\mathcal{A}_1) \quad \text{almost surely.}$$

Proof. The random variable $\mathbf{E}(X|\mathcal{A}_1)$ is $\mathcal{A}_1$-measurable, so also $\mathcal{A}_2$-measurable, and therefore $\mathbf{E}(\mathbf{E}(X|\mathcal{A}_1)|\mathcal{A}_2) = \mathbf{E}(X|\mathcal{A}_1)$ almost surely. For all $C \in \mathcal{A}_1 \subseteq \mathcal{A}_2$,

$$\mathbf{E}(\mathbb{1}_C\mathbf{E}(X|\mathcal{A}_1)) = \mathbf{E}(\mathbb{1}_C X) = \mathbf{E}(\mathbb{1}_C\mathbf{E}(X|\mathcal{A}_2)),$$

so

$$\mathbf{E}(\mathbf{E}(X|\mathcal{A}_2)|\mathcal{A}_1) = \mathbf{E}(X|\mathcal{A}_1) \quad \text{fast sicher.}$$

□

(K) Assume that the σ-field $\mathcal{A}$ is independent of the integrable random variable X. Then,

$$\mathbf{E}(X|\mathcal{A}) = \mathbf{E}(X) \quad \text{almost surely.}$$

Proof. The constant random variable $\mathbf{E}(X)$ is $\mathcal{A}$-measurable. Let $C \in \mathcal{A}$. Due to independence,

$$\mathbf{E}(\mathbb{1}_C X) = \mathbf{E}(\mathbb{1}_C)\mathbf{E}(X) = \mathbf{E}(\mathbf{E}(X)\mathbb{1}_C),$$

which proves the statement. □

Exercise 4.5. Let X_1 and X_2 be independent random variables with $\mathbb{E}(X_i) = \mathbb{E}(X_i^2) = 1$, $i = 1, 2$ and define $\mathcal{A} = \sigma(X_1)$. Show that then,

$$\mathbb{E}((X_1 + X_2)^2|\mathcal{A}) = (X_1 + 1)^2.$$

4.3 Factorization of the conditional expectation

Let X, Y be random variables on $(\Omega, \mathfrak{F}, \mathbb{P})$. If X is integrable (or non-negative), we define

$$\mathbf{E}(X|Y) := \mathbf{E}(X|\sigma(Y)).$$

If Y only takes values 0 and 1, i.e. $\mathbb{P}(Y = 0) = p \in (0,1)$ and $\mathbb{P}(Y = 1) = 1 - p$, then $\sigma(Y) = \{A, A^c, \emptyset, \Omega\}$, where $A = Y^{-1}(\{0\})$. If X is integrable, we obtain

$$\mathbf{E}(X|Y) = \mathbf{E}(X|\sigma(Y)) = \mathbb{1}_A \mathbf{E}(X|A) + \mathbb{1}_{A^c} \mathbf{E}(X|A^c), \quad \text{almost surely.}$$

Knowing the value of Y, we would like to determine the conditional expectation of X given $Y = 0$ (or $Y = 1$). In the above case, it is clear that for $y \in \{0, 1\}$, we should choose

$$\mathbf{E}(X|Y = y) := f(y) := \frac{1}{1-p}\big(y\mathbf{E}(X) + (1 - y - p)\mathbf{E}(X|A)\big)$$

because then,

$$\mathbf{E}(X|Y) = f(Y) \quad \text{almost surely.}$$

We now want to generalize the above arguments for an arbitrary random variable Y. For this, we need the following theorem.

Theorem 4.6. *Let Y, Z be random variables. Then Z is $\sigma(Y)$-measurable if and only if there exists a measurable function $f : \mathbb{R} \to \mathbb{R}$ with $Z = f(Y)$.*

Proof. Exercise □

Corollary 4.7. *Let X, Y be random variables and X integrable. Then there exists a measurable function $f : \mathbb{R} \to \mathbb{R}$ such that*

$$\mathbf{E}(X|Y) = f(Y)$$

almost surely, and f is $\mathbf{P}_Y$ almost surely unique.

Proof. The existence of f is proved in an exercise. We show that f $\mathbf{P}_Y$ is almost surely unique. For all $B \in \mathcal{B}(\mathbb{R})$,

$$\int \mathbb{1}_B f \,\mathrm{d}\mathbf{P}_Y = \int (\mathbb{1}_B \circ Y)(f \circ Y)\,\mathrm{d}\mathbb{P} = \mathbf{E}(\mathbb{1}_{Y^{-1}(B)} f(Y)) = \mathbf{E}(\mathbb{1}_{Y^{-1}(B)} \mathbf{E}(X|Y)) = \mathbf{E}(\mathbb{1}_{Y^{-1}(B)} X).$$

Let now $g : \mathbb{R} \to \mathbb{R}$ be another measurable function such that $\mathbf{E}(X|Y) = g(Y)$ almost surely. The above computation shows that the $\mathbb{P}$-integrability implies the $\mathbf{P}_Y$-integrability of f and g (choose $B = \mathbb{R}$). So for all $B \in \mathcal{B}(\mathbb{R})$,

$$\int \mathbb{1}_B f \,\mathrm{d}\mathbf{P}_Y = \int \mathbb{1}_B g \,\mathrm{d}\mathbf{P}_Y.$$

With $B = \{f \geq g\}$ we obtain that $\mathbb{1}_B(f - g) = \mathbb{1}_{\{f \geq g\}}(f - g) = 0$ $\mathbf{P}_Y$ almost surely, and analogously $\mathbb{1}_{\{g \geq f\}}(g - f) = 0$ $\mathbf{P}_Y$ almost surely, which proves the claim. □

With the function f from Corollary 4.7, we can therefore define

$$\mathbf{E}(X|Y = y) := f(y).$$

4.4 Martingales

In this chapter we introduce martingales and shortly illustrate two important theorems of martingale theory. Martingales are one of the main tools in the theory of stochastic processes, and they are treated extensively in the lectures Stochastic Processes I and II. This chapter only gives a non-comprehensive preview on this interesting and central topic of probability theory.

Let $(\Omega, \mathfrak{F}, \mathbb{P})$ be a probability space.

Definition 4.8. Let $\mathbb{A} = (\mathcal{A}_n)_{n\in\mathbb{N}}$ be a family of σ-fields $\mathcal{A}_n \subseteq \mathfrak{F}$, $n \in \mathbb{N}$. The family $\mathbb{A}$ is called *filtration* if

$$\mathcal{A}_n \subseteq \mathcal{A}_{n+1}, \quad \text{for all } n \in \mathbb{N}.$$

A sequence $(X_n)_{n\in\mathbb{N}}$ of random variables is called ($\mathbb{A}$-)*adapted*, if X_n is $\mathcal{A}_n$-measurable for all $n \in \mathbb{N}$.

Example 4.9. Any sequence $(X_n)_{n\in\mathbb{N}}$ of random variables is adapted to the filtration $(\sigma(X_1, \dots, X_n))_{n\in\mathbb{N}}$.

Definition 4.10. Let $\mathbb{A} = (\mathcal{A}_n)_{n\in\mathbb{N}}$ be a filtration and $(X_n)_{n\in\mathbb{N}}$ a sequence of integrable random variables. $(X_n)_{n\in\mathbb{N}}$ is called *martingale*, if for all $n \in \mathbb{N}$,

$$\mathbb{E}(X_{n+1}|\mathcal{A}_n) = X_n \quad \text{almost surely.}$$

Example 4.11 (Sums of independent centred random variables)**.** Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent integrable random variables with $\mathbb{E}(X_n) = 0$ for all $n \in \mathbb{N}$. We set $S_n = \sum_{k=1}^n X_k$ for $n \in \mathbb{N}$. Then $(S_n)_{n\in\mathbb{N}}$ is a martingale with respect to $(\sigma(X_1, \dots, X_n))_{n\in\mathbb{N}}$, because

$$\begin{aligned}
\mathbb{E}(S_{n+1}|\sigma(X_1, \dots, X_n)) &= \mathbb{E}(X_{n+1} + S_n|\sigma(X_1, \dots, X_n)) \\
&= \mathbb{E}(X_{n+1}|\sigma(X_1, \dots, X_n)) + \mathbb{E}(S_n|\sigma(X_1, \dots, X_n)) \\
&= \mathbb{E}(X_{n+1}) + S_n = S_n
\end{aligned}$$

where we have used properties (K) and (B) of conditional expectations in the second-last step.

Example 4.12 (Products of independent random variables)**.** Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent, integrable random variables with $\mathbb{E}(X_n) = 1$ for all $n \in \mathbb{N}$. We set $M_n = \prod_{k=1}^n X_k$ for $n \in \mathbb{N}$. Then $(M_n)_{n\in\mathbb{N}}$ is a martingale with respect to $(\sigma(X_1, \dots, X_n))_{n\in\mathbb{N}}$, because

$$\begin{aligned}
\mathbb{E}(M_{n+1}|\sigma(X_1, \dots, X_n)) &= \mathbb{E}(X_{n+1}M_n|\sigma(X_1, \dots, X_n)) \\
&= M_n\mathbb{E}(X_{n+1}|\sigma(X_1, \dots, X_n)) \\
&= M_n\mathbb{E}(X_{n+1}) = M_n
\end{aligned}$$

where we first use property (I) and then (K) of conditional expectations.

Example 4.13 (Successive predictions)**.** Let Y be an integrable random variable and $\mathbb{A} = (\mathcal{A}_n)_{n\in\mathbb{N}}$ an arbitrary filtration. Then $(X_n)_{n\in\mathbb{N}}$ with

$$X_n = \mathbb{E}(Y|\mathcal{A}_n)$$

is a martingale. From the definition it is obvious that $(X_n)_{n\in\mathbb{N}}$ is $\mathbb{A}$-adapted. Property (A) shows that all X_n are integrable, and because of property (J),

$$\mathbb{E}(X_{n+1}|\mathcal{A}_n) = \mathbb{E}(\mathbb{E}(Y|\mathcal{A}_{n+1})|\mathcal{A}_n) = \mathbb{E}(Y|\mathcal{A}_n) = X_n.$$

Definition 4.14. Let $\mathbb{A} = (\mathcal{A}_n)_{n\in\mathbb{N}}$ be a filtration. A mapping $T : \Omega \to \mathbb{N} \cup \{\infty\}$ is called *stopping time*, if

$$\{T \leq n\} \in \mathcal{A}_n \quad \text{for all } n \in \mathbb{N}.$$

Theorem 4.15 (Stopping time)**.** *Let T be a stopping time and $(X_n)_{n\in\mathbb{N}}$ a martingale with respect to a filtration $\mathbb{A}$. Then the stopped process $(X_{\min\{n,T\}})_{n\in\mathbb{N}}$ is a martingale too. If $T < \infty$ almost surely and if $(X_{\min\{n,T\}})_{n\in\mathbb{N}}$ is dominated by an integrable function, then*

$$\mathbb{E}(X_T) = \mathbb{E}(X_1).$$

Example 4.16 (Ruin problem)**.** Let $(Y_n)_{n\in\mathbb{N}}$ be a sequence of independent and identically distributed random variables with $\mathbb{P}(Y_n = 1) = \mathbb{P}(Y_n = -1) = 1/2$. We define $S_n = \sum_{k=1}^n Y_k$. Then $(S_n)_{n\in\mathbb{N}}$ is a martingale due to example 4.11, with the filtration $(\sigma(X_1, \dots, X_n))_{n\in\mathbb{N}}$. The martingale $(S_n)_{n\in\mathbb{N}}$ is also called *simple symmetric random walk.* We can interpret S_n as the gain/loss of a player in a sequence of independent games against a bank, and each game has a probability of winning of 1/2 and a constant investment of 1.

For $a, b \in \mathbb{N}$ we define

$$T := \inf\{n \geq 1 \mid S_n \notin (-a, b)\}.$$

The random variable T is a stopping time. If a is interpreted as the starting capital of the player and b as the starting capital of the bank, then T describes the point of time when the game ends because either the player is ruined ($S_T = -a$) or the bank ($S_T = b$). The probability of ruin of the player is

$$p := \mathbb{P}(S_T = -a).$$

We can compute p by applying the stopping theorem (Theorem 4.15). The Borel-Cantelli Lemma implies that $\mathbb{P}(T < \infty) = 1$, and $|S_{\min\{n,T\}}| \leq \max\{a, b\}$. Therefore

$$0 = \mathbb{E}(S_1) = \mathbb{E}(S_T) = p(-a) + (1-p)b,$$

so

$$p = \frac{b}{b+a}$$

which is the bank's share of the total capital.

Theorem 4.17 (Convergence theorem)**.** *Let $(X_n)_{n\in\mathbb{N}}$ be a martingale with*

$$\sup_{n\in\mathbb{N}} \mathbb{E}(|X_n|) < \infty.$$

Then $(X_n)_{n\in\mathbb{N}}$ converges almost surely to an integrable random variableX_∞.

Example 4.18. We again consider a simple symmetric random walk $(S_n)_{n\in\mathbb{N}}$ as in example 4.16. The martingale $(S_n)_{n\in\mathbb{N}}$ does not converge, because $|S_{n+1} - S_n| = 1$ for all $n \in \mathbb{N}$. Still, the convergence theorem allows us to derive an interesting result on the random walk. We define for $c \in \mathbb{Z}$ the stopping time

$$T_c := \inf\{n > 1 | S_n = c\}$$

Theorem 4.15 shows that $(S_{\min\{n,T_c\}})_{n\in\mathbb{N}}$ is a martingale. If $c > 0$, then $S_{\min\{n,T_c\}} \leq c$, so $(S_{\min\{n,T_c\}})_+ \leq c$ and therefore

$$\begin{aligned}
\sup_{n\in\mathbb{N}} \mathbb{E}(|S_{\min\{n,T_c\}}|) &= \sup_{n\in\mathbb{N}}((\mathbb{E}(S_{\min\{n,T_c\}}))_+ + (\mathbb{E}(S_{\min\{n,T_c\}}))_-) \\
&= \sup_{n\in\mathbb{N}}(2(\mathbb{E}(S_{\min\{n,T_c\}}))_+ - \mathbb{E}(S_{\min\{1,T_c\}})) \leq 2c - \mathbb{E}(S_{\min\{1,T_c\}}) < \infty.
\end{aligned}$$

If $c < 0$, one can apply an analogous argument. Hence the convergence theorem (Theorem 4.17) implies that $\lim_{n\to\infty} S_{\min\{n,T_c\}}$ exists almost surely, which shows (because $(S_n)_{n\in\mathbb{N}}$ almost surely does not converge) that $\mathbb{P}(T_c < \infty) = 1$ for all $c \in \mathbb{Z}$, so also $\mathbb{P}(T_c < \infty \text{ for all } c \in \mathbb{Z}) = 1$. This shows that

$$\mathbb{P}(\liminf_{n\to\infty} S_n = -\infty, \limsup_{n\to\infty} S_n = +\infty) = 1.$$

We conclude that the simple symmetric random walk almost surely oscillates almost surely in an arbitrarily wide range.

The convergence theorem for martingales can also be used, for example, to derive the law of large numbers not only for sequences $(X_n)_{n\in\mathbb{N}}$ of independent random variables, but also for sequences with certain forms of dependency.