Probability Theory

Matthias Löwe

Academic Year 2001/2002

Contents

1 Introduction **1**

2 Basics, Random Variables **1**

3 Expectation, Moments, and Jensen's inequality **3**

4 Convergence of random variables **7**

5 Independence **11**

6 Products and Sums of Independent Random Variables **16**

7 Infinite product probability spaces **18**

8 Zero-One Laws **23**

9 Laws of Large Numbers **26**

10 The Central Limit Theorem **35**

11 Conditional Expectation **43**

12 Martingales **50**

13 Brownian motion **57**
13.1 Construction of Brownian Motion 61

14 Appendix **66**

1 Introduction

Chance, luck, and fortune have been in the centre of human mind ever since people have started to think. The latest when people started to play cards or roll dices for money, there has also been a desire for a mathematical description and understanding of chance. Experience tells, that even in a situation, that is governed purely by chance there seems to be some regularly. For example, in a long sequence of fair coin tosses we will typically see "about half of the time" heads and "about half of the time" tails. This was formulated already by Jakob Bernoulli (published 1713) and is called a law of large numbers. Not much later the French mathematician de Moivre analyzed how much the typical number of heads in a series of n fair coin tosses fluctuates around $\frac{1}{2}n$. He thereby discovered the first form of what nowadays has become known as the Central Limit Theorem.

However, a mathematically "clean" description of these results could not be given by either of the two authors. The problem with such a description is that, of course, the average number of heads in a sequence of fair coin tosses will only **typically** converge to $\frac{1}{2}$.

One could, e.g., imagine that we are rather unlucky and toss only heads. So the principal question was: "what is the probability of an event?" The standard idea in the early days was, to define it as the limit of the average time of occurrences of the event in a long row of typical experiments. But here we are back again at the original problem of what a typical sequence is. It is natural that, even if we can define the concept of typical sequences, they are very difficult to work with. This is, why Hilbert in his famous talk on the International Congress of Mathematicians 1900 in Paris mentioned the axiomatic foundation of probability theory as the sixth of his 23 open problems in mathematics.

This problem was solved by A. N. Kolmogorov in 1933 by ingeniously making use of the newly developed field of measure theory: A probability is understood as a measure on the space of all outcomes of the random experiment. This measure is chosen in such a way that it has total mass one.

From this start-up the whole framework of probability theory was developed: Starting from the laws of large numbers over the Central Limit Theorem to the very new field of mathematical finance (and many, many others).

In this course we will meet the most important highlights from probability theory and then turn into the direction of stochastic processes, that are basic for mathematical finance.

2 Basics, Random Variables

As mentioned in the introduction the concept of Kolmogorov understands a probability on space of outcomes as measure with total mass one on this space. So in the framework of probability theory we will always consider a triple $(\Omega, \mathcal{F}, \mathbb{P})$ and call it a probability space:

Definition 2.1 *A probability space is a triple* $(\Omega, \mathcal{F}, \mathbb{P})$*, where*

- Ω *is a set,* $\omega \in \Omega$ *is called a state of the world.*

- *$\mathcal{F}$ is a σ-algebra over Ω, $A \in \mathcal{F}$ is called an event.*
- *$\mathbb{P}$ is a measure on $\mathcal{F}$ with $\mathbb{P}(\Omega) = 1$, $\mathbb{P}(A)$ is called the probability of event A.*

A probability space can be considered as an experiment we perform. Of course, in an experiment in physics one is not always interested to measure everything one could in principle measure. Such a measurement in probability theory is called a random variable.

Definition 2.2 *A random variable X is a mapping*

$$X : \Omega \to \mathbb{R}^d$$

that is measurable. Here we endow $\mathbb{R}^d$ with its Borel σ-algebra $\mathcal{B}^d$.

The important fact about random variables is that the underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$ does not really matter. For example, consider the two experiments

$$\Omega_1 = \{0, 1\}, \mathcal{F}_1 = \mathcal{P}\Omega_1, \text{ and } \mathbb{P}_1\{0\} = \frac{1}{2}$$

and

$$\Omega_2 = \{1, 2, \ldots, 6\}, \mathcal{F}_2 = \mathcal{P}\Omega_2, \text{ and } \mathbb{P}_2\{i\} = \frac{1}{6}, \quad i \in \Omega_2.$$

Consider the random variables

$$\begin{aligned} X_1 : \quad & \Omega_1 \to \mathbb{R} \\ & \omega \mapsto \omega \end{aligned}$$

and

$$\begin{aligned} X_2 : \quad & \Omega_2 \to \mathbb{R} \\ & \omega \mapsto \begin{cases} 0 & i \text{ is even} \\ 1 & i \text{ is odd.} \end{cases} \end{aligned}$$

Then

$$\mathbb{P}_1(X_1 = 0) = \mathbb{P}_2(X_2 = 0) = \frac{1}{2}$$

and therefore X_1 and X_2 have same behavior even though they are defined on completely different spaces. What we learn from this example is that what really matters for a random variable is the " distribution" $\mathbb{P} \circ X^{-1}$:

Definition 2.3 *The distribution $\mathbb{P}_X$ of a random variable $X : \Omega \to \mathbb{R}^d$, is the following probability measure on $\left(\mathbb{R}^d, \mathcal{B}^d\right)$:*

$$\mathbb{P}_X(A) := \mathbb{P}(X \in A) = \mathbb{P}\left(X^{-1}(A)\right), \quad A \in \mathcal{B}^d.$$

So the distribution of a random variable is its image measure in $\mathbb{R}^d$.

Example 2.4 *Important distributions of random variables X (we have already met in introductory courses in probability and statistics) are*

- *The Binomial distribution with parameters n and p, i.e. a random variable X is Binomially distributed with parameters n and p ($B(n,p)$-distributed), if*

$$\mathbb{P}(X=k)=\binom{n}{k}p^k(1-p)^{n-k} \qquad 0\leq k\leq n.$$

The binomial distribution is the distribution of the number of 1's in n independent coin tosses with success probability p.

- *The Normal distribution with parameters μ and σ^2, i.e. a random variable X is normally distributed with parameters μ and σ^2 ($\mathcal{N}(\mu,\sigma^2)$-distributed), if*

$$\mathbb{P}(X\leq a)=\frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{a}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}dx.$$

Here $a\in\mathbb{R}$.

- *The Dirac distribution with atom in $a\in\mathbb{R}$, i.e. a random variable X is Dirac distributed with parameter a, if*

$$\mathbb{P}(X=b)=\delta_a(b)=\begin{cases}1 & \text{if } b=a\\ 0 & \text{otherwise.}\end{cases}$$

Here $b\in\mathbb{R}$.

- *The Poisson distribution with parameter $\lambda\in\mathbb{R}$, i.e. a random variable X is Poisson distributed with parameter $\lambda\in\mathbb{R}$ ($\mathcal{P}(\lambda)$-distributed), if*

$$\mathbb{P}(X=k)=\frac{\lambda^k e^{-\lambda}}{k!} \qquad k\in\mathbb{N}_0=\mathbb{N}\cup\{0\}.$$

- *The Multivariate Normal distribution with parameters $\mu\in\mathbb{R}^d$ and Σ, i.e. a random variable X is normally distributed in dimension d with parameters μ and Σ ($\mathcal{N}(\mu,\Sigma)$-distributed), if $\mu\in\mathbb{R}^d$, Σ is a symmetric, positive definite $d\times d$ matrix and for $A=(-\infty,a_1]\times\ldots\times(-\infty,a_d]$*

$$\mathbb{P}(X\in A)=\\ \frac{1}{\sqrt{(2\pi)^d\det\Sigma}}\int_{-\infty}^{a_1}\ldots\int_{-\infty}^{a_d}\exp\left(-\frac{1}{2}\langle\Sigma^{-1}(x-\mu),(x-\mu)\rangle\right)dx_1\ldots dx_d.$$

3 Expectation, Moments, and Jensen's inequality

We are now going to consider important characteristics of random variables

Definition 3.1 *The expectation of a random variable is defined as*

$$\mathbb{E}(X):=E_{\mathbb{P}}(X):=\int X d\mathbb{P}=\int X(\omega)d\mathbb{P}(\omega)$$

if this integral is well defined.

Notice that for $A \in \mathcal{F}$ one has $\mathbb{E}(1_A) = \int 1_A d\mathbb{P} = \mathbb{P}(A)$. Quite often one may want to integrate $f(X)$ for a function $f : \mathbb{R}^d \to \mathbb{R}^m$. How does this work?

Proposition 3.2 *Let $X : \Omega \to \mathbb{R}^d$ be a random variable and $f : \mathbb{R}^d \to \mathbb{R}$ be a measurable function. Then*

$$\int f \circ X d\mathbb{P} = \mathbb{E}[f \circ X] = \int f d\mathbb{P}_X. \tag{3.1}$$

Proof. If $f = 1_A, A \in \mathcal{B}^d$ we have

$$\mathbb{E}[f \circ X] = \mathbb{E}(1_A \circ X) = \mathbb{P}(X \in A) = \mathbb{P}_X(A) = \int 1_A d\mathbb{P}_X.$$

Hence (3.1) holds true for functions $f = \sum_{i=1}^n \alpha_i 1_{A_i}, \alpha_i \in \mathbb{R}, A_i \in \mathcal{B}^d$. The standard approximation techniques yield (3.1) for general integrable f. ■

In particular Proposition 3.2 yields that

$$\mathbb{E}[X] = \int x d\mathbb{P}_X(x).$$

Exercise 3.3 *If $X : \Omega \to \mathbb{N}_0$, then*

$$\mathbb{E}X = \sum_{n=0}^{\infty} n\mathbb{P}(X = n) = \sum_{n=1}^{\infty} \mathbb{P}(X \geq n)$$

Exercise 3.4 *If $X : \Omega \to \mathbb{R}$, then*

$$\sum_{n=1}^{\infty} \mathbb{P}(|X| \geq n) \leq \mathbb{E}(|X|) \leq \sum_{n=0}^{\infty} \mathbb{P}(|X| \geq n).$$

Thus X has an expectation, if and only if $\sum_{n=0}^{\infty} \mathbb{P}(|X| \geq n)$ converges.

Further characteristics of random variables are the p-th moments.

Definition 3.5

1. *For $p \geq 1$, the p-th moment of a random variable is defined as $\mathbb{E}(X^p)$.*
2. *The centered p-th moment of a random variable is defined as $\mathbb{E}[(X - \mathbb{E}X)^p]$.*
3. *The* **variance** *of a random variable X is its centered second moment, hence*

$$\mathbb{V}(X) := \mathbb{E}\left[(X - \mathbb{E}X)^2\right].$$

Its standard deviation σ is defined as

$$\sigma := \sqrt{\mathbb{V}(X)}.$$

Proposition 3.6 *$\mathbb{V}(X) < \infty$ if and only if $X \in \mathcal{L}^2(\mathbb{P})$. In this case*

$$\mathbb{V}(X) = \mathbb{E}\left(X^2\right) - \left(\mathbb{E}\left(X\right)\right)^2 \tag{3.2}$$

as well as

$$\mathbb{V}(X) \leq \mathbb{E}\left(X^2\right) \tag{3.3}$$

and

$$(\mathbb{E}X)^2 \leq \mathbb{E}\left(X^2\right) \tag{3.4}$$

Proof. If $\mathbb{V}(X) < \infty$, then $X - \mathbb{E}X \in \mathcal{L}^2(\mathbb{P})$. But $\mathcal{L}^2(\mathbb{P})$ is a vector space and the constant

$$\mathbb{E}X \in \mathcal{L}^2(\mathbb{P}) \Rightarrow X = (X - \mathbb{E}X) + \mathbb{E}X \in \mathcal{L}^2(\mathbb{P}).$$

On the other hand if $X \in \mathcal{L}^2(\mathbb{P}) \Rightarrow X \in \mathcal{L}^1(\mathbb{P})$. Hence $\mathbb{E}X$ exists and is a constant. Therefore also

$$X - \mathbb{E}X \in \mathcal{L}^2(\mathbb{P}).$$

By linearity of expectation then:

$$\mathbb{V}(X) = E(X - \mathbb{E}X)^2 = \mathbb{E}X^2 - 2(\mathbb{E}X)^2 + (\mathbb{E}X)^2 = \mathbb{E}X^2 - (\mathbb{E}X)^2.$$

This immediately implies (3.3) and (3.4). ■

Exercise 3.7 *Show that X and $X - \mathbb{E}X$ have the same variance.*

It will turn out in the next step that (3.4) is a special case of a much more general principle. To this end recall the concept of a **convex** function. Recall that a function $\varphi : \mathbb{R} \to \mathbb{R}$ is convex if for all $\alpha \in (0,1)$ we have

$$\varphi(\alpha x + (1-\alpha)y) \leq \alpha\varphi(x) + (1-\alpha)\varphi(y)$$

for all $x, y \in \mathbb{R}$. In a first course in analysis one learns that convexity is implied by $\varphi'' \geq 0$. On the other hand convex functions do not need to be differentiable. The following exercise shows that they are close to being differentiable.

Exercise 3.8 *Let I be an interval and $\varphi : I \to \mathbb{R}$ be a convex function. Show that the right derivative $\varphi'_+(x)$ exist for all $x \in \mathring{I}$ (the interior of x) and the left derivative $\varphi'_-(x)$ exists for all $x \in \mathring{I}$. Hence φ is continuous on $\mathring{I}$. Show moreover that φ'_+ is monotonely increasing on $\mathring{I}$ and it holds:*

$$\varphi(y) \geq \varphi(x) + \varphi'_+(x)(y - x)$$

for $x \in \mathring{I}, y \in I$.

Applying Exercise 3.8 yields the generalization of (3.4) mentioned above.

Theorem 3.9 (Jensen's inequality) *Let $X : \Omega \to \mathbb{R}$ be a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$ and assume X is $\mathbb{P}$-integrable and takes values in an open interval $I \subset \mathbb{R}$. Then $\mathbb{E}X \in I$ and for every convex*

$$\varphi : I \to \mathbb{R}$$

$\varphi \circ X$ is a random variable. If this random variable $\varphi \circ X$ is $\mathbb{P}$-integrable it holds:

$$\varphi\left(\mathbb{E}\left(X\right)\right) \leq \mathbb{E}(\varphi \circ X).$$

Proof. Assume $I = (\alpha, \beta)$. Thus $X(\omega) < \beta$ for all $\omega \in \Omega$. But then $\mathbb{E}(X) \leq \beta$, but then also $\mathbb{E}(X) < \beta$. Indeed $\mathbb{E}(X) = \beta$, implies that the strictly positive random variable $\beta - X(\omega)$ equals 0 on a set of $\mathbb{P}$ measure one, i.e. $\mathbb{P}$-almost surely. This is a contradiction. Analogously: $\mathbb{E}X > \alpha$. According to exercise 3.8 φ is continuous on $\mathring{I} = I$ hence Borel-measurable. Now we know

$$\varphi(y) \geq \varphi\left(x\right) + \varphi'_{+}\left(x\right)\left(y - x\right) \tag{3.5}$$

for all $x, y \in I$ with equality for $x = y$. Hence

$$\varphi\left(y\right) = \sup_{x \in I}\left[\varphi\left(x\right) + \varphi'_{+}\left(x\right)\left(y - x\right)\right] \tag{3.6}$$

for all $y \in I$. Putting $y = X(\omega)$ in (3.5) yields

$$\varphi \circ X \geq \varphi(x) + \varphi'_{+}(x)\left(X - x\right)$$

and by integration

$$\mathbb{E}(\varphi \circ X) \geq \varphi(x) + \varphi'_{+}(x)\left(\mathbb{E}\left(X\right) - x\right)$$

all $x \in I$. Together with (3.6) this gives

$$\mathbb{E}(\varphi \circ X) \geq \sup_{x \in I}\left[\varphi(x) + \varphi'_{+}(x)(\mathbb{E}X - x)\right] = \varphi(\mathbb{E}(X)).$$

This is the assertion of Jensen's inequality. ■

Corollary 3.10 *Let $X \in \mathcal{L}^p(\mathbb{P})$ for some $p \geq 1$. Then*

$$|\mathbb{E}(X)|^p \leq \mathbb{E}(|X|^p).$$

Exercise 3.11 *Let I be an open interval and $\varphi : I \to \mathbb{R}$ be convex. For $x_1, \ldots, x_n \in I$ and $\lambda_1, \ldots \lambda_n \in \mathbb{R}^+$ with $\sum_{i=1}^n \lambda_i = 1$, show that*

$$\varphi\left(\sum_{i=1}^{n} \lambda_i x_i\right) \leq \sum_{i=1}^{n} \lambda_i \varphi\left(x_i\right).$$

4 Convergence of random variables

Already in the course in measure theory we met three different types on convergence:

Definition 4.1 *Let (X_n) be a sequence of random variables.*

1. *X_n is stochastically convergent (or convergent in probability) to a random variable X, if for each $\varepsilon > 0$*
$$\lim_{n\to\infty} \mathbb{P}\left(|X_n - X| \geq \varepsilon\right) = 0$$

2. *X_n converges almost surely to a random variable X, if for each $\varepsilon > 0$*
$$\mathbb{P}\left(\limsup_{n\to\infty} |X_n - X| \geq \varepsilon\right) = 0$$

3. *X_n converges to a random variable X in $\mathcal{L}^p$ or in p-norm, if*
$$\lim_{n\to\infty} E\left(|X_n - X|^p\right) = 0.$$

Already in measure theory we proved:

Theorem 4.2 *1. If X_n converges to X almost surely or in $\mathcal{L}^p$, then it also converges stochastically. None of the converses is true.*

2. Almost sure convergence does not imply convergence in $\mathcal{L}^p$ and vice versa.

Definition 4.3 *Let $(\Omega, \mathcal{F})$ be a measurable, topological space endowed with its Borel σ-algebra $\mathcal{F}$. This means $\mathcal{F}$ is generated by the topology on Ω. Moreover for each $n \in \mathbb{N}$ let μ_n and μ be probability measures on $(\Omega, \mathcal{F})$. We say that μ_n converges weakly to μ, if for each bounded, continuous, and real valued function $f : \Omega \to \mathbb{R}$ (we will write $\mathcal{C}^b(\Omega)$ for the space of all such functions) it holds*

$$\mu_n(f) := \int f d\mu_n \to \int f d\mu =: \mu(f) \text{ as } n \to \infty \tag{4.1}$$

Theorem 4.4 *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of real valued random variables on a space $(\Omega, \mathcal{F}, \mathbb{P})$. Assume (X_n) converges to a random variable X stochastically. Then $\mathbb{P}_{X_n}$ (the sequence of distributions) converges weakly to $\mathbb{P}_X$,i.e.*

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

or equivalently

$$\lim_{n\to\infty} \mathbb{E}\left(f \circ X_n\right) = \mathbb{E}\left(f \circ X\right)$$

for all $f \in \mathcal{C}^b(\mathbb{R})$.

If X is constant $\mathbb{P}$-a.s. the converse also holds true.

Proof. First assume $f \in \mathcal{C}^b(\mathbb{R})$ is uniformly continuous. Then for $\varepsilon > 0$ there is a $\delta > 0$ such that for any $x', x'' \in \mathbb{R}$

$$|x' - x''| < \delta \Rightarrow |f(x') - f(x'')| < \varepsilon.$$

Define $A_n := \{|X_n - X| \geq \delta\}, n \in \mathbb{N}$. Then

$$\begin{aligned}
\left| \int f d\mathbb{P}_{X_n} - \int f d\mathbb{P}_X \right| &= |\mathbb{E}(f \circ X_n) - \mathbb{E}(f \circ X)| \\
&\leq \mathbb{E}(|f \circ X_n - f \circ X|) \\
&= \mathbb{E}(|f \circ X_n - f \circ X| \circ 1_{A_n}) + \mathbb{E}\left(|f \circ X_n - f \circ X| \circ 1_{A_n^c}\right) \\
&\leq 2\|f\| \mathbb{P}(A_n) + \varepsilon \mathbb{P}(A_n^c) \\
&\leq 2\|f\| \mathbb{P}(A_n) + \varepsilon.
\end{aligned}$$

Here we used the notation

$$\|f\| := \sup\{|f(x)| : x \in \mathbb{R}\}$$

and that

$$|f \circ X_n - f \circ X| \leq |f \circ X_n| + |f \circ X| \leq 2\|f\|.$$

But since $X_n \to X$ stochastically, $\mathbb{P}(A_n) \to 0$ as $n \to \infty$, so for n large enough $P(A_n) \leq \frac{\varepsilon}{2\|f\|}$. Thus for such n

$$\left| \int f d\mathbb{P}_{X_n} - \int f d\mathbb{P}_X \right| \leq 2\varepsilon.$$

Now let $f \in \mathcal{C}^b(\mathbb{R})$ be arbitrary and denote $I_n := [-n, n]$. Since $I_n \uparrow \mathbb{R}$ as $n \to \infty$ we have $\mathbb{P}_X(I_n) \uparrow 1$ as $n \to \infty$. Thus for all $\varepsilon > 0$ there is $n_0(\varepsilon) =: n_0$ such that

$$1 - \mathbb{P}_X(I_{n_0}) = \mathbb{P}_X(\mathbb{R} \setminus I_{n_0}) < \varepsilon.$$

We choose the continuous function u_{n_0} in the following way:

$$u_{n_0}(x) = \begin{cases} 1 & x \in I_{n_0} \\ 0 & |x| \geq n_0 + 1 \\ -x + n_0 + 1 & n_0 < x < n_0 + 1 \\ x + n_0 + 1 & -n_0 - 1 < x < -n_0 \end{cases}$$

Eventually put $\bar{f} := u_{n_0} \cdot f$. Since $\bar{f} \equiv 0$ outside the compact set $[-n_0 - 1, n_0 + 1]$, the function $\bar{f}$ is uniformly continuous (and so is u_{n_0}) and hence

$$\lim_{n \to \infty} \int \bar{f} d\mathbb{P}_{X_n} = \int \bar{f} d\mathbb{P}_X$$

as well as

$$\lim_{n \to \infty} \int u_{n_0} d\mathbb{P}_{X_n} = \int u_{n_0} d\mathbb{P}_X;$$

thus also

$$\lim_{n \to \infty} \int (1 - u_{n_0}) d\mathbb{P}_{X_n} = \int (1 - u_{n_0}) d\mathbb{P}_X.$$

By the triangle inequality

$$\left|\int f d\mathbb{P}_{X_n} - \int f d\mathbb{P}_X\right| \leq \int \left|f - \bar{f}\right| d\mathbb{P}_{X_n} + \left|\int \bar{f} d\mathbb{P}_{X_n} - \int \bar{f} d\mathbb{P}_X\right| + \int \left|f - \bar{f}\right| d\mathbb{P}_X. \tag{4.2}$$

For large $n \geq n_1(\varepsilon)$, $\left|\int \bar{f} d\mathbb{P}_{X_n} - \int \bar{f} d\mathbb{P}_X\right| \leq \varepsilon$. Furthermore from $0 \leq 1 - u_{n_0} \leq 1_{\mathbb{R}\setminus I_{n_0}}$ we obtain

$$\int (1 - u_{n_0})\, d\mathbb{P}_X \leq \mathbb{P}_X\left(\mathbb{R} \setminus I_{n_0}\right) < \varepsilon,$$

so that for all $n \geq n_2(\varepsilon)$ also

$$\int (1 - u_{n_0})\, d\mathbb{P}_{X_n} < \varepsilon.$$

This yields

$$\int \left|f - \bar{f}\right| d\mathbb{P}_X = \int |f|\,(1 - u_{n_0})\, d\mathbb{P}_X \leq \|f\|\,\varepsilon$$

on the one hand and

$$\int \left|f - \bar{f}\right| d\mathbb{P}_{X_n} \leq \|f\|\,\varepsilon$$

all $n \geq n_2(\varepsilon)$ on the other. Hence we obtain from (4.2) for large n:

$$\left|\int f d\mathbb{P}_{X_n} - \int f d\mathbb{P}_X\right| \leq 2\,\|f\|\,\varepsilon + \varepsilon.$$

This proves weak convergence of the distributions.

For the converse let $\eta \in \mathbb{R}$ and assume $X \equiv \eta$ (X is identically equal to $\eta \in \mathbb{R}$ $\mathbb{P}$-almost surely). This means $\mathbb{P}_X = \delta_\eta$ where δ_η is the Dirac measure concentrated in η. For the open interval $I = (\eta - \varepsilon, \eta + \varepsilon), \varepsilon > 0$ we may find $f \in \mathcal{C}^b(\mathbb{R})$ with $f \leq 1_I$ and $f(\eta) = 1$.

Then

$$\int f d\mathbb{P}_{X_n} \leq \mathbb{P}_{X_n}(I) = P(X_n \in I) \leq 1.$$

Since we assumed weak convergence of $\mathbb{P}_{X_n}$ to $\mathbb{P}_X$ we know that

$$\int f d\mathbb{P}_{X_n} \to \int f d\mathbb{P}_X = f(\eta) = 1$$

as $n \to \infty$. Since $\int f d\mathbb{P}_X \leq \mathbb{P}(X_n \in I) \leq 1$, this implies

$$P(X_n \in I) \to 1$$

as $n \to \infty$. But

$$\{X_n \in I\} = \{|X_n - \eta| \leq \varepsilon\}$$

and thus

$$\begin{aligned} \mathbb{P}(|X_n - X| \geq \varepsilon) &= \mathbb{P}(|X_n - \eta| \geq \varepsilon) \\ &= 1 - \mathbb{P}(|X_n - \eta| < \varepsilon) \to 0 \end{aligned}$$

for all $\varepsilon > 0$. This means X_n converges stochastically to X. ■

Definition 4.5 *Let X_n, X be random variables on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. If $\mathbb{P}_{X_n}$ converges weakly to $\mathbb{P}_X$ we also say that X_n converges to X in distribution.*

Remark 4.6 *If the random variables in Theorem 4.4 are $\mathbb{R}^d$-valued and the $f : \mathbb{R}^d \to \mathbb{R}$ belong to $\mathcal{C}^b(\mathbb{R}^d)$ the statement of the theorem stays valid.*

Example 4.7 *For each sequence $\sigma_n > 0$ with $\lim \sigma_n = 0$ we have*

$$\lim_{n\to\infty} \mathcal{N}(0, \sigma_n^2) = \delta_0.$$

Here δ_0 denotes the Dirac measure concentrated in $0 \in \mathbb{R}$. Indeed, substituting $x = \sigma y$ we obtain

$$\frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} f(x) e^{-\frac{x^2}{2\sigma^2}} dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} f(\sigma y) dy$$

Now for all $y \in \mathbb{R}$

$$\left| \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} f(\sigma y) \right| \leq \|f\| e^{-\frac{y^2}{2}}$$

which is integrable. Thus by dominated convergence

$$\begin{aligned} \lim_{n\to\infty} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma_n^2}} e^{-\frac{x^2}{2\sigma_n^2}} f(x) dx &= \lim_{n\to\infty} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} f(\sigma_n y) dy \\ &= \int_{-\infty}^{\infty} \lim_{n\to\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} f(\sigma_n y) dy = f(0) = \int f d\delta_0. \end{aligned}$$

Exercise 4.8 *Show that the converse direction in Theorem 4.4 is not true, if we drop the assumption that $X \equiv \eta$ $\mathbb{P}$-a.s.*

Exercise 4.9 *Assume the following holds for sequence (X_n) of random variables on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$:*

$$\mathbb{P}(|X_n| > \varepsilon) < \varepsilon$$

For all n large enough (larger than $n(\varepsilon)$) for each given $\varepsilon > 0$. Is this equivalent with stochastic convergence of X_n to 0 ?

Exercise 4.10 *For a sequence of Poisson distributions (π_{α_n}) with parameters $\alpha_n > 0$ show that*

$$\lim_{n\to\infty} \pi_{\alpha_n} = \delta_0,$$

if $\lim_{n\to\infty} \alpha_n = 0$. Is there a probability measure μ on $\mathcal{B}^1$, with

$$\lim_{n\to\infty} \pi_{\alpha_n} = \mu \quad \textit{(weakly)}$$

if $\lim \alpha_n = \infty$?

5 Independence

The concept of independence of events is one of the most essential in probability theory. It is met already in the introductory courses. Its background is the following:

Assume we are given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. For two events $A, B \in \mathcal{F}$ with $\mathbb{P}(B) > 0$ one may ask, how the probability of the event A changes, if we know already that B has happened, we only need to consider the probability of $A \cap B$. To obtain a probability again we normalize by $\mathbb{P}(B)$ and get the conditional probability of A given B:

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

now we would call A and B independent, if the knowledge that B has happened does not change the probability that A will happen or not, i.e. if $\mathbb{P}(A \mid B)$ and $\mathbb{P}(A)$ are the same

$$\mathbb{P}(A \mid B) = P(A).$$

This in other words means A and B are independent, if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B).$$

More generally we define

Definition 5.1 *A family $(A_i)_{i \in I}$ of events on $\mathcal{F}$ is called independent if for each choice of different indices $i_1, \ldots, i_n \in I$*

$$\mathbb{P}(A_{i_1} \cap .. \cap A_{i_n}) = \mathbb{P}(A_{i_1}) \ldots \mathbb{P}(A_{i_n}) \tag{5.1}$$

Exercise 5.2 *Give an example of a sequence of events that are pairwise independent, i.e. each pair of events from this sequence is independent, but not independent (i.e. all events together not independent).*

We generalize Definition 5.1 to set systems in the following way:

Definition 5.3 *For each $i \in I$ let $\mathcal{E}_i \subset \mathcal{F}$ be a collection of events. $(\mathcal{E}_i)_{i \in I}$ are called independent if (5.1) holds for each $i_1, \ldots, i_n \in I$, each $n \in \mathbb{N}$ and each $A_{i_\nu} \in \mathcal{E}_{i_\nu}, \nu = 1, \ldots, n$.*

Exercise 5.4 *Show the following*

1. *A family $(\mathcal{E}_i)_{i \in I}$ is independent, if and only if every finite sub-family is independent.*
2. *Independence of $(\mathcal{E}_i)_{i \in I}$ is maintained, if we reduce the families $(\mathcal{E}_i)$. More precisely, let $(\mathcal{E}_i)$ be independent and $\mathcal{E}'_i \subseteq \mathcal{E}_i$ then also the families $(\mathcal{E}'_i)$ are independent.*
3. *If for all $n \in \mathbb{N}$, $(\mathcal{E}_i^n)_{i \in I}$ are independent and for all $n \in \mathbb{N}$ and $i \in I$, $\mathcal{E}_i^n \subseteq \mathcal{E}_i^{n+1}$, then $(\bigcup_n \mathcal{E}_i^n)_{i \in I}$ are independent.*

Exercise 5.5 *If $(\mathcal{E}_i)_{i\in I}$ are independent, then so are the Dynkin-systems $(\mathcal{D}(\mathcal{E}_i))_{i\in I}$. Here $\mathcal{D}(\mathcal{A})$ is the Dynkin system generated by $\mathcal{A}$, it coincides with the intersection of all Dynkin systems containing $\mathcal{A}$. (See the Appendix for definitions.)*

Corollary 5.6 *Let $(\mathcal{E}_i)_{i\in I}$ be an independent family of $\cap$-stable sets $\mathcal{E}_i \subseteq \mathcal{F}$. Then also the families $(\sigma(\mathcal{E}_i))_{i\in I}$ of the σ-algebras generated by the $\mathcal{E}_i$ are independent.*

Theorem 5.7 *Let $(\mathcal{E}_i)_{i\in I}$ be an independent family of $\cap$-stable sets $\mathcal{E}_i \subseteq \mathcal{F}$ and*

$$I = \dot{\cup}_{j\in J} I_j$$

with $I_i \cap I_j = \emptyset, i \neq j$. Let $\mathcal{A}_j := \sigma\left(\cup_{i\in I_j}\mathcal{E}_i\right)$. Then also $(\mathcal{A}_j)_{j\in J}$ is independent.

Proof. For $j \in J$ let $\tilde{\mathcal{E}}_j$ be the system of all sets of the form

$$E_{i_1} \cap \ldots \cap E_{i_n}$$

where $\emptyset \neq \{i_1, \ldots, i_n\} \subseteq I_j$ and $E_{i_\nu} \in \mathcal{E}_{i_\nu}$, $\nu = 1, \ldots, n$, are arbitrary. Then $\tilde{\mathcal{E}}_j$ is $\cap$-stable. As an immediate consequence of the independence of the $(\mathcal{E}_i)_{i\in I}$ also the $(\tilde{\mathcal{E}}_j)_{j\in J}$ are independent. Eventually $\mathcal{A}_j = \sigma(\tilde{\mathcal{E}}_j)$. Thus the assertion follows from Corollary 5.6. ■

In the next step we want to show that events that depend on all but finitely many σ-algebras of a countable family of independent σ-algebras can only have probability zero or one. To this end we need the following definition.

Definition 5.8 *Let $(\mathcal{A}_n)_n$ be a sequence of σ-algebras from $\mathcal{F}$ and*

$$\mathcal{T}_n := \sigma\left(\bigcup_{m=n}^{\infty} \mathcal{A}_m\right)$$

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

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

is called the σ-algebra of the tail events.

Exercise 5.9 *Why is $\mathcal{T}_\infty$ a σ-algebra?*

This is now the result announced above:

Theorem 5.10 (Kolmogorov's Zero-One-Law) *Let $(\mathcal{A}_n)$ be an independent sequence of σ-algebras $\mathcal{A}_n \subseteq \mathcal{F}$. Then for every tail event $A \in \mathcal{T}_\infty$ it holds*

$$\mathbb{P}(A) = 0 \quad \text{or} \quad \mathbb{P}(A) = 1.$$

Proof. Let $A \in \mathcal{T}_\infty$ and $\mathcal{D}$ the system of all sets $D \in \mathcal{F}$ that are independent of A. We want to show that $A \in \mathcal{D}$:

In the exercise below we show that $\mathcal{D}$ is a Dynkin system. By Theorem 5.7 the σ-algebra $\mathcal{T}_{n+1}$ is independent of the σ-algebra

$$\overline{\mathcal{A}}_n := \sigma\left(\mathcal{A}_1 \cup \ldots \cup \mathcal{A}_n\right).$$

Since $A \in \mathcal{T}_{n+1}$ we know that $\overline{\mathcal{A}}_n \subseteq \mathcal{D}$ for every $n \in \mathbb{N}$. Thus

$$\overline{\mathcal{A}} := \bigcup_{n=1}^{\infty} \overline{\mathcal{A}}_n \subseteq \mathcal{D}$$

Obviously $\left(\overline{\mathcal{A}}_n\right)$ is increasing. For $E, F \in \overline{\mathcal{A}}$ there thus exists n with $E, F \in \overline{\mathcal{A}}_n$ and hence $E \cap F \in \overline{\mathcal{A}}_n$ and thus $E \cap F \in \overline{\mathcal{A}}$. This means that $\overline{\mathcal{A}}$ is $\cap$-stable. Since $\overline{\mathcal{A}} \subseteq \mathcal{D}$, i.e. $(\overline{\mathcal{A}}, \{A\})$ is independent, from Exercise 5.5 we conclude that $(\mathcal{D}(\overline{\mathcal{A}}), \{A\})$ is independent, so that $\sigma\left(\overline{\mathcal{A}}\right) = \mathcal{D}\left(\overline{\mathcal{A}}\right) \subseteq \mathcal{D}$. Moreover $\mathcal{A}_n \subseteq \overline{\mathcal{A}}$, all n. Hence $\mathcal{T}_1 = \sigma(\cup \mathcal{A}_n) \subseteq \sigma\left(\overline{\mathcal{A}}\right)$. Therefore

$$A \in \mathcal{T}_\infty \subseteq \mathcal{T}_1 \subseteq \sigma\left(\overline{\mathcal{A}}\right) \subseteq \mathcal{D}.$$

Therefore A is independent of A, i.e. it holds

$$\mathbb{P}(A) = \mathbb{P}(A \cap A) = \mathbb{P}(A) \cdot \mathbb{P}(A) = (\mathbb{P}(A))^2.$$

Hence $\mathbb{P}(A) \in \{0, 1\}$ is asserted. ■

Exercise 5.11 *Show that $\mathcal{D}$ from the proof of Theorem 5.10 is a Dynkin system.*

As an immediate consequence of the Kolmogorov Zero-One-Law (Theorem 5.10) we obtain

Theorem 5.12 (Borel's Zero-One-Law) *For each independent sequence $(A_n)_n$ of events $A_n \in \mathcal{F}$ we have*

$$\mathbb{P}(A_n \text{ for infinitely many } n) = 0$$

or

$$\mathbb{P}(A_n \text{ for infinitely many } n) = 1,$$

i.e.

$$\mathbb{P}\left(\limsup A_n\right) \in \{0, 1\}.$$

Proof. Let $\mathcal{A}_n = \sigma(A_n)$, i.e. $\mathcal{A}_n = \{\emptyset, \Omega, A_n, A_n^c\}$. It follows that $(\mathcal{A}_n)_n$ is independent. For $Q_n := \bigcup_{m=n}^{\infty} A_m$ we have $Q_n \in \mathcal{T}_n$. Since $(\mathcal{T}_n)_n$ is decreasing we even have $Q_m \in \mathcal{T}_n$ for all $m \geq n, n \in \mathbb{N}$. Since $(Q_n)_n$ is decreasing we obtain

$$\limsup_{n \to \infty} A_n = \bigcap_{k=1}^{\infty} Q_k = \bigcap_{k=j}^{\infty} Q_k \in \mathcal{T}_j$$

for all $j \in \mathbb{N}$. Hence $\limsup A_n \in \mathcal{T}_\infty$. Hence the assertion follows from Kolmogorov's zero-one law. ■

Exercise 5.13 *In every $\mathcal{F}$ the pairs (A,B) (any $A,B\in\mathcal{F}$ with $\mathbb{P}(A)=0$ or $\mathbb{P}(A)=1$ or $\mathbb{P}(B)=0$ or $\mathbb{P}(B)=1$) are pairs of independent sets. If these are the only pairs of independent sets we call $\mathcal{F}$ independence-free. Show that the following space is independence-free:*

$$\Omega=\mathbb{N},\ \mathcal{F}=\mathcal{P}(\Omega),\ \text{and}\ \mathbb{P}(\{k\})=2^{-k!}$$

for each $k\geq 2$, $\mathbb{P}(\{1\})=1-\sum_{k=2}^{\infty}\mathbb{P}(k)$. (Hint: Door zo nodig over te gaan op complementen, mag je aannemen dat $1\notin A$ en $1\notin B$.)

A special case of the above abstract setting is the concept of independent random variables. This will be introduced next. Again we work over a probability space $(\Omega,\mathcal{F},\mathbb{P})$.

Definition 5.14 *A family of random variables $(X_i)_i$ is called independent, if the σ-algebras $(\sigma(X_i))_i$ generated by them are independent.*

For finite families there is another criterion (which is important, since by definition of independence we only need to check the independence of finite families).

Theorem 5.15 *Let $X_1,\ldots,X_n$ be a sequence of n random variables with values in measurable spaces $(\Omega_i,\mathcal{A}_i)$ with $\cap$-stable generators $\mathcal{E}_i$ of $\mathcal{A}_i$. $X_1,\ldots,X_n$ are independent if and only if*

$$\mathbb{P}(X_1\in E_1,\ldots,X_n\in E_n)=\prod_{i=1}^{n}\mathbb{P}(X_i\in E_i)$$

for all $E_i\in\mathcal{E}_i, i=1,\ldots,n$.

Proof. Put

$$\mathcal{G}_i:=\left\{X_i^{-1}(E_i),E_i\in\mathcal{E}_i\right\}.$$

Then $\mathcal{G}_i$ generates $\sigma(X_i)$. $\mathcal{G}_i$ is $\cap$-stable and $\Omega\in\mathcal{G}_i$. According to Corollary 5.6 we need to show the independence of $(\mathcal{G}_i)_{i=1\ldots n}$, which is equivalent with

$$\mathbb{P}(G_1\cap\cdots\cap G_n)=\mathbb{P}(G_1)\cdots\mathbb{P}(G_n)$$

for all choices of $G_i\in\mathcal{G}_i$. Sufficiency is evident, since we may choose $G_i=\Omega$ for appropriate i. ■

Exercise 5.16 *Random variables $X_1,\ldots,X_{n+1}$ are independent with values in $(\Omega_i,\mathcal{F}_i)$ if and only if $X_1,\ldots,X_n$ are independent and X_{n+1} is independent of $\sigma(X_1,\ldots,X_n)$.*

The following theorem states that a measurable deformation of an independent family of random variarables stays independent:

Theorem 5.17 *Let $(X_i)_{i\in I}$ be a family of independent random variables X_i with values in $(\Omega_i,\mathcal{A}_i)$ and let*

$$f_i:(\Omega_i,\mathcal{A}_i)\to(\Omega_i',\mathcal{A}_i')$$

be measurable. Then $(f_i(X_i))_{i\in I}$ is independent.

Proof. Let $i_1, \ldots, i_n \in I$. Then

$$\begin{aligned} & \mathbb{P}\left(f_{i_1}(X_{i_1}) \in A'_{i_1}, \ldots, f_{i_n}(X_{i_n}) \in A'_{i_n}\right) \\ = \; & \mathbb{P}\left(X_{i_1} \in f_{i_1}^{-1}\left(A'_{i_1}\right), \ldots, X_{i_n} \in f_{i_n}^{-1}\left(A'_{i_n}\right)\right) \\ = \; & \prod_{\nu=1}^{n} \mathbb{P}\left(X_{i_\nu} \in f_{i_\nu}^{-1}\left(A'_{i_\nu}\right)\right) \\ = \; & \prod_{\nu=1}^{n} \mathbb{P}\left(f_{i_\nu}(X_{i_\nu}) \in A'_{i_\nu}\right) \end{aligned}$$

by the independence of $(X_i)_{i\in I}$. Here the $A'_{i_\nu} \in \mathcal{A}'_{i_\nu}$ were arbitrary. ■

Already Theorem 5.15 gives rise to the idea that independence of random variables may be somehow related to product measures. This is made more precise in the following theorem. To this end let $X_1, \ldots, X_n$ be random variables such that

$$X_i : (\Omega, \mathcal{F}) \to (\Omega_i, \mathcal{A}_i).$$

Define

$$Y := X_1 \otimes \cdots \otimes X_n : \Omega \to \Omega_1 \times \cdots \times \Omega_n$$

Then the distribution of Y which we denote by $\mathbb{P}_Y$ can be computed as $\mathbb{P}_Y = \mathbb{P}_{X_1 \otimes \cdots \otimes X_n}$. Note that $\mathbb{P}_Y$ is a probability measure on $\otimes_{i=1}^n \mathcal{A}_i$.

Theorem 5.18 *The random variables $X_1, \ldots, X_n$ are independent if and only if their distribution is the product measure of the individual distributions, i.e. if*

$$\mathbb{P}_{X_1 \otimes \ldots \otimes X_n} = \mathbb{P}_{X_1} \otimes \cdots \otimes \mathbb{P}_{X_n}$$

Proof. Let $A_i \in \mathcal{A}_i$, $i = 1, \ldots, n$. Then with

$$Y = X_1 \otimes \cdots \otimes X_n :$$

$$\mathbb{P}_Y\left(\prod_{i=1}^{n} A_i\right) = \mathbb{P}\left(Y \in \prod_{i=1}^{n} A_i\right) = \mathbb{P}\left(X_1 \in A_1, \ldots, X_n \in A_n\right)$$

as well as

$$\mathbb{P}_{X_i}(A_i) = \mathbb{P}(X_i \in A_i) \qquad i = 1, \ldots n.$$

Now $\mathbb{P}_Y$ is the product measure of the $\mathbb{P}_{X_i}$ if and only if

$$\mathbb{P}_Y(A_1 \times \ldots \times A_n) = \mathbb{P}_{X_1}(A_1) \ldots \mathbb{P}_{X_n}(A_n).$$

But this is identical with

$$\mathbb{P}\left(X_1 \in A_1, \ldots, X_n \in A_n\right) = \prod_{i=1}^{n} \mathbb{P}\left(X_i \in A_i\right).$$

But according to Theorem 5.15 this is equivalent with the independence of the X_i. ■

6 Products and Sums of Independent Random Variables

In this section we will study independent random variables in greater detail.

Theorem 6.1 *Let $X_1, \ldots, X_n$ be independent, real-valued random variables. Then*

$$\mathbb{E}\left(\prod_{i=1}^{n} X_i\right) = \prod_{i=1}^{n} \mathbb{E}\left(X_i\right) \tag{6.1}$$

if $\mathbb{E}X_i$ is well defined (and finite) for all i. (6.1) shows that then also $\mathbb{E}\left(\prod_{i=1}^{n} X_i\right)$ is well defined.

Proof. We know that $Q := \otimes_{i=1}^{n} \mathbb{P}_{X_i}$ is the joint distribution of the $X_1, \ldots, X_n$. By Proposition 3.2 and Fubini's theorem

$$\begin{aligned}
\mathbb{E}\left(\left|\prod_{i=1}^{n} X_i\right|\right) &= \int |x_1 \ldots x_n| \, dQ(x_1, \ldots, x_n) \\
&= \int \ldots \int |x_1| \ldots |x_n| \, d\mathbb{P}_{X_1}(x_1) \ldots d\mathbb{P}_{X_n}(x_n) \\
&\quad \int |x_1| \, d\mathbb{P}_{X_1}(x_1) \ldots \int |x_n| \, d\mathbb{P}_{X_n}(x_n)
\end{aligned}$$

This shows that integrability of the X_i implies integrability of $\prod_{i=1}^{n} X_i$. In this case the equalities are also true without absolute values. This proves the result. ■

Exercise 6.2 *For any two random variables X, Y that are integrable Theorem 6.1 tells that independence of X, Y implies that*

$$\mathbb{E}\left(X \cdot Y\right) = \mathbb{E}\left(X\right)\mathbb{E}\left(Y\right)$$

Show that the converse is not true.

Definition 6.3 *For any two random variables X, Y that are integrable and have an integrable product we define the covariance of X and Y to be*

$$\begin{aligned}
cov\left(X, Y\right) &= \mathbb{E}\left[\left(X - \mathbb{E}X\right)\left(Y - \mathbb{E}Y\right)\right] \\
&= \mathbb{E}\left(XY\right) - \mathbb{E}X\mathbb{E}Y.
\end{aligned}$$

X and Y are uncorrelated if $cov\left(X, Y\right) = 0$.

Remark 6.4 *If X, Y are independent $cov(X, Y) = 0$.*

Theorem 6.5 *Let $X_1, \ldots, X_n$ be square integrable random variables. Then*

$$\mathbb{V}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathbb{V}(X_i) + \sum_{i \neq j} cov\left(X_i, X_j\right) \tag{6.2}$$

In particular, if $X_1, \ldots, X_n$ are uncorrelated

$$\mathbb{V}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathbb{V}\left(X_i\right). \tag{6.3}$$

Proof. We have

$$\begin{aligned}\mathbb{V}\left(\sum_{i=1}^{n} X_i\right) &= \mathbb{E}\left[\left(\sum_{i=1}^{n}\left(X_i - \mathbb{E}X_i\right)\right)^2\right] \\ &= \mathbb{E}\left[\sum_{i=1}^{n}\left(X_i - \mathbb{E}X_i\right)^2 + \sum_{i\neq j}\left(X_i - \mathbb{E}X_i\right)\left(X_j - \mathbb{E}X_j\right)\right] \\ &= \sum_{i=1}^{n}\mathbb{V}\left(X_i\right) + \sum_{i=j} cov\left(X_i, X_j\right).\end{aligned}$$

This proves (6.2). For (6.3 just note that for uncorrelated random variables X, Y one has $cov(X, Y) = 0$. ■

Eventually we turn to determining the distribution of the sum of independent random variables.

Theorem 6.6 *Let $X_1, \ldots, X_n$ be independent $\mathbb{R}^d$ valued random variables. Then the distribution of the sum $S_n := X_1 + \ldots + X_n$ is given by the convolution product of the distributions of the X_i, i.e.*

$$\mathbb{P}_{S_n} := \mathbb{P}_{X_1} * \mathbb{P}_{X_2} * \cdots * \mathbb{P}_{X_n}$$

Proof. Again let $Y := X_1 \otimes \ldots \otimes X_n : \Omega \to \left(\mathbb{R}^d\right)^n$, and vector addition

$$A_n : \mathbb{R}^d \times \ldots \times \mathbb{R}^d \to \mathbb{R}^d.$$

Then $S_n = A_n \circ Y$, hence a random variable. Now $\mathbb{P}_{S_n}$ is the image measure of $\mathbb{P}$ under $A_n \circ Y$, which we denote by $\left(A_n \circ Y\right)(\mathbb{P})$. Thus

$$\mathbb{P}_{S_n} = \left(A_n \circ Y\right)(\mathbb{P}) = A_n\left(\mathbb{P}_Y\right).$$

Now $\mathbb{P}_Y = \otimes \mathbb{P}_{X_i}$. So by the definition of the convolution product

$$\mathbb{P}_{X_1} * \ldots * \mathbb{P}_{X_n} = A_n\left(\mathbb{P}_Y\right) = \mathbb{P}_{S_n}$$

More explicitly, in the case $d = 1$, let $g(x_1, \ldots, x_n) = 1$ for $x_1 + \cdots + x_n \leq s$ and $g(x_1, \ldots, x_n) = 0$, otherwise. Then application of Fubini's theorem yields

$$\mathbb{P}(S_n \leq s) = \mathbb{E}(g(X_1, \ldots, X_n)) = \int\int g(x_1, x_2, \ldots, x_n) d\mathbb{P}_{X_1}(x_1) d\mathbb{P}_{(X_2,\ldots,X_n)}(x_2, \ldots, x_n) =$$

$$= \int \mathbb{P}(X_1 \leq s - x_2 - \cdots - x_n) d\mathbb{P}_{(X_2,\ldots,X_n)}(x_2, \ldots, x_n)$$

In the case that X_1 has a density f_{X_1} with respect to Lebesgue measure, and the order of differentiation with respect to s and integration can be exchanged, it follows that S_n has a density f_{S_n} and

$$f_{S_n}(s) = \int f_{X_1}(s - x_2 - \cdots - x_n) d\mathbb{P}_{(X_2,\ldots,X_n)}(x_2, \ldots, x_n)$$

The same formula holds in the case that $X_1, \ldots, X_n$ have a discrete distribution (that is, almost surely assume values in a fixed countable subset of $\mathbb{R}$) if densities are taken with respect to the count measure. ■

Example 6.7 *1. As we learned in Introduction to Statistics the convolution of a Binomial distribution with parameters n and p, $B(n,p)$ and a Binomial distribution $B(m,p)$ is a $B(n+m,p)$ distribution:*

$$B(n.p) * B(m,p) = B(n+m,p).$$

2. As we learned in Introduction to Statistics the convolution of a $\mathcal{P}(\lambda)$-distribution (a Poisson distribution with parameter λ) with a $\mathcal{P}(\mu)$-distribution is a $P(\mu+\lambda)$-distribution:

$$\mathcal{P}(\lambda) * \mathcal{P}(\mu) = \mathcal{P}(\lambda+\mu)$$

3. As has been communicated in Introduction to Probability:

$$\mathcal{N}\left(\mu,\sigma^2\right) * \mathcal{N}\left(\nu,\tau^2\right) = \mathcal{N}\left(\mu+\nu,\sigma^2+\tau^2\right).$$

7 Infinite product probability spaces

Many theorems in probability theory start with: "Let $X_1, X_2, \ldots, X_n, \ldots$ be a sequence of i.i.d. random variables". But how do we know that such sequences really exist? This will be shown in this section. In the last section we established the framework for some of the most important theorems from probabilities, as the Weak Law of Large Numbers or the Central Limit Theorem. Those are the theorems that assume: "Let $X_1, X_2, \ldots X_n$ be i.i.d. random variables". Others, as the Strong Law of Large Numbers ask for the behavior of a sequence of independent and identically distributed (i.i.d.) random variables; they usually start like "Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables". The natural first question to ask is: Does such a sequence exist at all?

In the same way as the existence of a finite sequence of i.i.d. random variables is related to finite product measures, the answer to the above question is related to infinite product measures. So, we assume that we are given a sequence of measure spaces $(\Omega_n, \mathcal{A}_n, \mu_n)$ of which we moreover assume that

$$\mu_n(\Omega_n) = 1 \text{ for all } n.$$

We construct $(\Omega, \mathcal{A})$ as follows: We want each $\omega \in \Omega$ to be a sequence $(\omega_n)_n$ where $\omega_n \in \Omega_n$. So we put

$$\Omega := \prod_{n=1}^{\infty} \Omega_n.$$

Moreover we have the idea that a probability measure on Ω should be defined by what happens on the first n coordinates, $n \in \mathbb{N}$. So for $A_1 \in \mathcal{A}_1, A_2 \in \mathcal{A}_2, \ldots, A_n \in \mathcal{A}_n, n \in \mathbb{N}$ we want

$$A := A_1 \times \ldots \times A_n \times \Omega_{n+1} \times \Omega_{n+2} \times \ldots \tag{7.1}$$

to be in $\mathcal{A}$. By independence we want to define a measure μ on $(\Omega, \mathcal{A})$ that assigns to A defined in (7.1) the mass

$$\mu(A) = \mu_1(A) \ldots \mu_n(A_n).$$

We will solve this problem in greater generality. Let I be an index set and $(\Omega_i, \mathcal{A}_i, \mu_i)_{i\in I}$ be measure spaces with $\mu_i(\Omega_i) = 1$. For $\emptyset \neq K \subseteq I$ define

$$\Omega_K := \prod_{i\in K} \Omega_i, \tag{7.2}$$

in particular $\Omega := \Omega_I$. Let p_J^K for $J \subseteq K$ denote the canonical projection from Ω_K to Ω_J. For $J = \{i\}$ we will also write p_i^K instead of $p_{\{i\}}^K$ and p_i in place of p_i^I. Obviously

$$p_J^L = p_J^K \circ p_K^L \qquad (J \subseteq K \subseteq L) \tag{7.3}$$

and

$$p_J := p_J^I = p_J^K \circ p_K \qquad (J \subseteq K). \tag{7.4}$$

Moreover denote by

$$\mathcal{H}(I) := \{J \subseteq I, J \neq \emptyset, |J| \text{ is finite}\}.$$

For $J \in \mathcal{H}(I)$ the σ-algebras and measures

$$\mathcal{A}_J := \otimes_{i\in J}\mathcal{A}_i \text{ and } \mu_J := \otimes_{i\in J}\mu_i$$

are defined by Fubini's theorem in measure theory.

In analogy to the finite dimensional case we define

Definition 7.1 *The product σ-algebra $\otimes_{i\in I}\mathcal{A}_i$ of the σ-algebras $(\mathcal{A}_i)_{i\in I}$ is defined as the smallest σ-algebra $\mathcal{A}$ on Ω, such that all projections $p_i : \Omega \to \Omega_i$ are $(\mathcal{A}, \mathcal{A}_i)$-measurable. Hence*

$$\otimes_{i\in I}\mathcal{A}_i := \sigma(p_i, i \in I). \tag{7.5}$$

Exercise 7.2 *Show that*

$$\otimes_{i\in I}\mathcal{A}_i := \sigma(p_J, J \in \mathcal{H}(I)). \tag{7.6}$$

According to the above we are now looking for a measure μ on $(\Omega, \mathcal{A})$, that assigns mass $\mu_1(A_1)\dots\mu_n(A_n)$ to each A as defined in (7.1). In other words

$$\mu\left(p_J^{-1}\left(\prod_{i\in J} A_i\right)\right) = \mu_J\left(\prod_{i\in J} A_i\right).$$

The question, whether such a measure exists, is solved in

Theorem 7.3 *On $\mathcal{A} := \otimes_{i\in I}\mathcal{A}_i$ there is a unique measure μ with*

$$p_J(\mu) := \mu p_J^{-1} = \mu_J \tag{7.7}$$

for all $J \in \mathcal{H}(I)$. It holds $\mu(\Omega) = 1$.

Proof. We may assume $|I| = \infty$, since otherwise the result is known from Fubini's theorem. We start with some preparatory considerations:

In Exercise 7.4 below it will be shown that p_J^K is $(\mathcal{A}_K, \mathcal{A}_J)$-measurable for $J \subseteq K$ and that $p_J^K(\mu_K) = \mu_J, (J \subseteq K, J, K \in \mathcal{H}(I))$.

Hence, if we introduce the σ-algebra of the J-cylinder sets

$$\mathcal{Z}_J := p_J^{-1}(\mathcal{A}_J) \qquad (J \in \mathcal{H}(I)) \tag{7.8}$$

the measurability of p_J^K implies $\left(p_J^K\right)^{-1}(\mathcal{A}_J) \subset \mathcal{A}_K$ and thus

$$\mathcal{Z}_J \subseteq \mathcal{Z}_K \qquad (J \subseteq K, J, K \in \mathcal{H}(I)). \tag{7.9}$$

Eventually we introduce the system of all cylinder sets

$$\mathcal{Z} := \bigcup_{J \in \mathcal{H}(I)} \mathcal{Z}_J.$$

Note that due to (7.9) for $Z_1, Z_2 \in \mathcal{Z}$ we have $Z_1, Z_2 \in \mathcal{Z}_J$, for suitably chosen $J \in \mathcal{H}(I)$. Hence $\mathcal{Z}$ is an algebra (but generally not a σ-algebra). From (7.5) and (7.6) it follows

$$\mathcal{A} = \sigma(\mathcal{Z}).$$

Now we come to the main part of the proof. This will be divided into four parts.

1. Assume $\mathcal{Z} \ni Z = p_J^{-1}(A)$, $J \in \mathcal{H}(I), A \in \mathcal{A}_J$. According to (7.7) Z must get mass $\mu(Z) = \mu_J(A)$. We have to show that this is well defined. So let

   $$Z = p_J^{-1}(A) = p_K^{-1}(B)$$

   for $J, K \in \mathcal{H}(I)$, $A \in \mathcal{A}_J, B \in \mathcal{A}_K$. If $J \subseteq K$ we obtain:

   $$p_J^{-1}(A) = p_K^{-1}\left(\left(p_J^K\right)^{-1}(A)\right)$$

   and thus

   $$p_K^{-1}(B) = p_K^{-1}(B') \quad \text{with} \quad B' := \left(p_J^K\right)^{-1}(A).$$

   Since $p_K(\Omega) = \Omega_K$ we obtain

   $$B = B' = \left(p_J^K\right)^{-1}(A).$$

   Thus by the introductory considerations

   $$\mu_K(B) = \mu_J(A).$$

   For arbitrary J, K define $L := J \cup K$. Since $J, K \subseteq L$, (7.9) implies the existence of $C \in \mathcal{A}_L$ with $p_L^{-1}(C) = p_J^{-1}(A) = p_K^{-1}(B)$. Therefore from what we have just seen:

   $$\mu_L(C) = \mu_J(A) \quad \text{and} \quad \mu_L(C) = \mu_K(B).$$

   Hence

   $$\mu_J(A) = \mu_K(B).$$

   Thus the function

   $$\mu_0\left(p_J^{-1}(A)\right) = \mu_J(A) \quad (J \in \mathcal{H}(I), A \in \mathcal{A}_J), \tag{7.10}$$

   is well-defined on $\mathcal{Z}$.

2. Now we show that μ_0 as defined in (7.10) is a volume on $\mathcal{Z}$. Trivially it holds, $\mu_0 \geq 0$ and $\mu_0(\emptyset) = 0$. Moreover, as shown above for $Y, Z \in \mathcal{Z}, Y \cap Z = \emptyset$, there is a $J \in \mathcal{H}(I), A, B \in \mathcal{A}_J$ such that $Y = p_J^{-1}(A), Z = p_J^{-1}(B)$. Now $Y \cap Z = \emptyset$ implies $A \cap B = \emptyset$ and due to

$$Y \cup Z = p_J^{-1}(A \cup B)$$

we obtain

$$\mu_0(Y \cup Z) = \mu_J(A \cup B) = \mu_J(A) + \mu_J(B) = \mu_0(Y) + \mu_0(Z)$$

hence the finite additivity of μ_0.

It remains to show that μ_0 is also σ-additive. Then the general principles from measure theory yield that μ_0 can be uniquely extended to a measure μ on $\sigma(\mathcal{Z}) = \mathcal{A}$. μ also is a probability measure, because of $\Omega = p_J^{-1}(\Omega_J)$ for all $J \in \mathcal{H}(I)$ and therefore

$$\mu(\Omega) = \mu_0(\Omega) = \mu_J(\Omega_J) = 1.$$

To prove the σ-additivity of μ_0 we first show:

3. Let $Z \in \mathcal{Z}$ and $J \in \mathcal{H}(I)$. Then for all $\omega_J \in \Omega_J$ the set

$$Z^{\omega_J} := \{\omega \in \Omega : (\omega_J, p_{I\setminus J}(\omega)) \in Z\}$$

is a cylinder set. This set consists of all $\omega \in \Omega$ with the following property: if we replace the coordinates ω_i with $i \in J$ by the corresponding coordinates of ω_J, we obtain a point in Z. Moreover

$$\mu_0(Z) = \int \mu_0(Z^{\omega_J}) d\mu_J(\omega_J). \tag{7.11}$$

This is shown by the following consideration. For $Z \in \mathcal{Z}$ there are $K \in \mathcal{H}(I)$ and $A \in \mathcal{A}_K$ such that $Z = p_K^{-1}(A)$, this means that $\mu_0(Z) = \mu_K(A)$. Since I is infinite we may assume $J \subset K$ and $J \neq K$. For the ω_J-intersection of A in Ω_K, which we call A_{ω_J}, i.e. for the set of all $\omega' \in \Omega_{K\setminus J}$ with $(\omega_J, \omega') \in A$, it holds

$$Z^{\omega_J} = p_{K\setminus J}^{-1}(A_{\omega_J}).$$

By Fubini's theorem $A_{\omega_J} \in \mathcal{A}_{K\setminus J}$ and hence $Z^{\omega_J} = p_{K\setminus J}^{-1}(A_{\omega_J})$ are $(K\setminus J)$-cylindersets. Since $\mu_K = \mu_J \otimes \mu_{K\setminus J}$ Fubini's theorem implies

$$\mu_0(Z) = \mu_K(A) = \int \mu_{K\setminus J}(A^{\omega_J}) d\mu_J(\omega_J). \tag{7.12}$$

But this is (7.11), since

$$\mu_0(Z^{\omega_J}) = \mu_{K\setminus J}(A^{\omega_J})$$

(because of $Z^{\omega_J} = p_{K\setminus J}^{-1}(A_{\omega_J})$).

4. Eventually we show that μ_0 is $\emptyset$-continuous and thus σ-additive. To this end let (Z_n) be a decreasing family of cylinder sets in $\mathcal{Z}$ with $\alpha := \inf_n \mu_0(Z_n) > 0$. We will show that

$$\bigcap_{n=1}^{\infty} Z_n \neq \emptyset. \tag{7.13}$$

Now each Z_n is of the form $Z_n = p_{J_n}^{-1}(A_n)$, $J_n \in \mathcal{H}(I)$, $A_n \in \mathcal{A}_{J_n}$. Due to (7.9) we may assume $J_1 \subseteq J_2 \subseteq J_3 \ldots$. We apply the result proved in 3. to $J = J_1$ and $Z = Z_n$. As $\omega_{J_1} \longmapsto \mu_0\left(Z_n^{\omega_{J_1}}\right)$ is $\mathcal{A}_{J_1}$-measurable

$$Q_n := \left\{\omega_{J_1} \in \Omega_{J_1} : \mu_0\left(Z_n^{\omega_{J_1}}\right) \geq \frac{\alpha}{2}\right\} \in \mathcal{A}_{J_1}.$$

Since all μ_J's have mass one we obtain from (7.11):

$$\alpha \leq \mu_0(Z_n) \leq \mu_{J_1}(Q_n) + \frac{\alpha}{2},$$

hence $\mu_{J_1}(Q_n) \geq \frac{\alpha}{2} > 0$, for all $n \in \mathbb{N}$. Together with (Z_n) also (Q_n) is decreasing. μ_{J_1} as a finite measure is $\emptyset$ - continuous, which implies $\bigcap_{n=1}^{\infty} Q_n \neq \emptyset$. Hence there is $\omega_{J_1} \in \bigcap_{n=1}^{\infty} Q_n$ with

$$\mu_0\left(Z_n^{\omega_{J_1}}\right) \geq \frac{\alpha}{2} > 0 \text{ all } n. \tag{7.14}$$

Successive application of 3. implies via induction that for each $k \in \mathbb{N}$ there is $\omega_{J_k} \in \Omega_{J_k}$ with (7.14) $\mu_0\left(Z_n^{\omega_{J_k}}\right) \geq 2^{-k}\alpha > 0$ and $p_{J_k}^{J_{k+1}}\left(\omega_{J_{k+1}}\right) = \omega_{J_k}$.

Due to this second property there is $\omega_0 \in \Omega$ with $p_{J_k}(\omega_0) = \omega_{J_k}$. Because of (7.14) we have $Z_n^{\omega_{J_n}} \neq \emptyset$ such that there is $\tilde{\omega}_n \in \Omega$ with $\left(\omega_{J_n}, p_{I\setminus J_n}(\tilde{\omega}_n)\right) \in Z_n$. But then also

$$\left(\omega_{J_n}, p_{I\setminus J_n}(\omega_0)\right) = \omega_0 \in Z_n.$$

Thus $\omega_0 \in \bigcap_{n=1}^{\infty} Z_n$ which proves (7.13).
Therefore μ_0 is σ-additive and hence has an extension to $\mathcal{A}$ by Carathéodory's theorem (Theorem 14.7). It is clear that μ_0 has mass one (i.e. $\mu_0(\Omega) = 1$), since for $J \in \mathcal{H}(I)$ we have $\Omega = p_J^{-1}(\Omega_J)$ and hence

$$\mu_0(\Omega) = \mu_J(\Omega_J) = 1.$$

In particular μ_0 is σ–finite, and the extension μ is unique, and it is a probability measure, that is $\mu(\Omega) = \mu_0(\Omega) = 1$.

This proves the theorem. ■

We conclude the chapter with an Exercise, that was left open during this proof.

Exercise 7.4 *With the notations of this section, in particular of Theorem 7.3 show that* p_J^K *is* $(\mathcal{A}_K, \mathcal{A}_J)$*-measurable* $(J \subseteq K, J, K \in \mathcal{H}(I))$ *and that*

$$p_J^K(\mu_K) = \mu_J.$$

8 Zero-One Laws

Already in Section 5 we encountered the prototype of a zero-one law: For a sequence of events $(A_n)_n$ that are independent we have Borel's Zero-One-Law (Theorem 5.12):

$$\mathbb{P}(\limsup A_n) \in \{0, 1\}.$$

In a first step we will now ask, when the probability in question is zero and when it is one. This leads to the following frequently used lemma:

Lemma 8.1 (Borel-Cantelli Lemma) *Let (A_n) be a sequence of events over a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Then*

$$\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty \Rightarrow \mathbb{P}(\limsup A_n) = 0 \tag{8.1}$$

If the events (A_n) are pairwise independent then also

$$\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty \Rightarrow \mathbb{P}(\limsup A_n) = 1. \tag{8.2}$$

Remark 8.2 *The Borel-Cantelli Lemma is most often used in the form of (8.1). Note that this part does not require any knowledge about the dependence structure of the A_n.*

Proof of Lemma 8.1. (8.1) is easy. Define

$$A := \limsup A_n = \bigcap_{n=1}^{\infty} \bigcup_{i=n}^{\infty} A_i.$$

This implies

$$A \subseteq \bigcup_{i=n}^{\infty} A_i \quad \text{for all } n \in \mathbb{N}.$$

and thus

$$\mathbb{P}(A) \leq \mathbb{P}\left(\bigcup_{i=n}^{\infty} A_i\right) \leq \sum_{i=n}^{\infty} \mathbb{P}(A_i) \tag{8.3}$$

Since $\sum_{i=1}^{\infty} \mathbb{P}(A_i)$ converges, $\sum_{i=n}^{\infty} \mathbb{P}(A_i)$ converges to zero as $n \to \infty$. This implies $\mathbb{P}(A) = 0$, hence (8.1).

For (8.2) again put $A := \limsup A_n$ and furthermore

$$I_n := 1_{A_n}, \quad S_n := \sum_{j=1}^{n} I_j$$

and eventually

$$S := \sum_{j=1}^{\infty} I_j.$$

Since the A_n are assumed to be pairwise independent they are pairwise uncorrelated as well. Hence

$$\begin{aligned}\mathbb{V}(S_n) &= \sum_{j=1}^{n}\mathbb{V}(I_j) = \sum_{j=1}^{n}\left[\mathbb{E}\left(I_j^2\right) - \mathbb{E}(I_j)^2\right] \\ &= \mathbb{E}(S_n) - \sum_{j=1}^{n}\mathbb{E}(I_j)^2 \leq \mathbb{E}S_n,\end{aligned}$$

where the last equality follows since $I_j^2 = I_j$. Now by assumption $\sum_{n=1}^{\infty}\mathbb{E}(I_n) = +\infty$. Since $S_n \uparrow S$ this is equivalent with

$$\lim_{n\to\infty}\mathbb{E}(S_n) = \mathbb{E}(S) = \infty \tag{8.4}$$

On the other hand $\omega \in A$, if and only if $\omega \in A_n$ for infinitely many n which is the case, if and only if $S(\omega) = +\infty$. The assertion thus is

$$\mathbb{P}(S = +\infty) = 1.$$

This can be seen as follows. By Chebyshev's inequality

$$\mathbb{P}(|S_n - \mathbb{E}(S_n)| \leq \eta) \geq 1 - \frac{\mathbb{V}(S_n)}{\eta^2}$$

for all $\eta > 0$. Because of (8.4) we may assume that $\mathbb{E}S_n > 0$ and choose $\eta = \frac{1}{2}\mathbb{E}S_n$. Hence

$$\mathbb{P}\left(S_n \geq \frac{1}{2}\mathbb{E}(S_n)\right) \geq \mathbb{P}\left(|S_n - \mathbb{E}S_n| \leq \frac{1}{2}\mathbb{E}S_n\right) \geq 1 - 4\frac{\mathbb{V}(S_n)}{\mathbb{E}(S_n)^2}$$

But $\mathbb{V}(S_n) \leq E(S_n)$ and $E(S_n) \to \infty$. Thus

$$\lim\frac{\mathbb{V}(S_n)}{\mathbb{E}(S_n)^2} = 0.$$

Therefore for all $\varepsilon > 0$ and all n large enough

$$\mathbb{P}\left(S_n \geq \frac{1}{2}\mathbb{E}S_n\right) \geq 1 - \varepsilon.$$

But now $S \geq S_n$ and hence also

$$\mathbb{P}\left(S \geq \frac{1}{2}\mathbb{E}S_n\right) \geq 1 - \varepsilon$$

for all $\varepsilon > 0$. But this implies $\mathbb{P}(S = +\infty) = 1$ which is what we wanted to show. ■

Example 8.3 *Let (X_n) be a sequence of real valued random variables which satisfies*

$$\sum_{n=1}^{\infty}\mathbb{P}(|X_n| > \varepsilon) < \infty \tag{8.5}$$

for all $\varepsilon > 0$. Then $X_n \to 0$ $\mathbb{P}$-a.s. Indeed the Borel-Cantelli Lemma says that (8.5) implies that

$$\mathbb{P}(|X_n| > \varepsilon \text{ infinitely often in } n) = 0.$$

But this is exactly the definition of almost sure convergence of X_n to 0.

Exercise 8.4 *Is (8.5) equivalent with $\mathbb{P}$-almost sure convergence of X_n to 0?*

Here is how Theorem 5.10 translates to random variables.

Theorem 8.5 (Kolmogorov's 0-1 Law) *Let $(X_n)_n$ be a sequence of independent random variables with values in arbitrary measurable spaces. Then for every tail event A, i.e. for each A with*

$$A \in \bigcap_{n=1}^{\infty} \sigma\left(X_m, m \geq n\right)$$

it holds that $\mathbb{P}(A) \in \{0,1\}$.

Exercise 8.6 *Derive Theorem 8.5 from Theorem 5.10.*

Corollary 8.7 *Let $(X_n)_{n\in\mathbb{N}}$ a sequence of independent, real-valued random variables. Define*

$$\mathcal{T}_\infty := \bigcap_{m=1}^{\infty} \sigma\left(X_i, i \geq m\right)$$

to be the tail σ-algebra. If then T is a real-valued random variable, that is measurable with respect to $\mathcal{T}_\infty$, then T is $\mathbb{P}$-almost surely constant. I.e. there is a $\alpha \in \mathbb{R}$ such that

$$\mathbb{P}(T = \alpha) = 1.$$

Such random variables $T : \Omega \to \mathbb{R}$ that are $\mathcal{T}_\infty$-measurable are called tail functions.

Proof. For each $\gamma \in \bar{\mathbb{R}}$ we have that

$$\{T \leq \gamma\} \in \mathcal{T}_\infty.$$

This implies $\mathbb{P}(T \leq \gamma) \in \{0,1\}$. On the other hand, being a distribution function we have

$$\lim_{\gamma\to-\infty} \mathbb{P}(T \leq \gamma) = 0 \text{ and } \lim_{\gamma\to+\infty} \mathbb{P}(T \leq \gamma) = 1$$

Let $C := \{\gamma \in \mathbb{R} : \mathbb{P}(T \leq \gamma) = 1\}$ and $\alpha := \inf(C) = \inf\{\gamma \in \mathbb{R} : \mathbb{P}(T \leq \gamma) = 1\}$ Then for an appropriately chosen decreasing sequence $(\gamma_n) \in C$ we have $\gamma_n \downarrow \alpha$ and since $\{T \leq \gamma_n\} \downarrow \{T \leq \alpha\}$ we have $\alpha \in C$. Hence $\alpha = \min\{\gamma \in C\}$. This implies

$$\mathbb{P}(T < \alpha) = 0$$

which implies

$$\mathbb{P}(T = \alpha) = 1.$$

■

Exercise 8.8 *A coin is tossed infinitely often. Show that every finite sequence*

$$(\omega_1, \ldots, \omega_k), \omega_i \in \{H, T\}, k \in \mathbb{N}$$

occurs infinitely often with probability one.

Exercise 8.9 *Try to prove (8.2) in the Borel-Cantelli Lemma for* ***independent*** *events* (A_n) *as follows:*

1. *For each sequence* (α_n) *of real numbers with* $0 \leq \alpha_n \leq 1$ *we have*

$$\prod_{i=1}^{n} (1 - \alpha_i) \leq \exp\left(-\sum_{i=1}^{n} \alpha_i\right) \tag{8.6}$$

This implies

$$\sum_{n=1}^{\infty} \alpha_n = \infty \Rightarrow \lim_{n\to\infty} \prod_{i=1}^{n} (1 - \alpha_i) = 0$$

2. *For* $A := \limsup A_n$ *we have*

$$\begin{aligned} 1 - \mathbb{P}(A) &= \lim_{n\to\infty} \mathbb{P}\left(\bigcap_{m=n}^{\infty} A_m^c\right) \\ &= \lim_{n\to\infty} \lim_{N\to\infty} \prod_{m=n}^{N} (1 - \mathbb{P}(A_m)). \end{aligned}$$

3. *As* $\sum \mathbb{P}(A_n)$ *diverges we have because of 1.*

$$\lim_{N\to\infty} \prod_{m=n}^{N} (1 - \mathbb{P}(A_m)) = 0$$

and hence $\mathbb{P}(A) = 1$ *because of 2. Fill in the missing details!*

9 Laws of Large Numbers

The central goal of probability theory is to describe the asymptotic behavior of a sequence of random variables. In its easiest form this has already been done for i.i.d. sequences in Introduction to Probability and Statistics. In the first theorem of this section this is slightly generalized.

Theorem 9.1 (Khintchine) *Let* $(X_n)_{n\in\mathbb{N}}$ *be a sequence of square integrable, real valued random variables, that are pairwise uncorrelated. Assume*

$$\lim_{n\to\infty} \frac{1}{n^2} \sum_{i=1}^{n} \mathbb{V}(X_i) = 0.$$

Then the ***weak law of large numbers*** *holds, i.e.*

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\mathbb{E}\sum_{i=1}^{n} X_i\right| > \varepsilon\right) = 0 \quad \text{for all } \varepsilon > 0.$$

Proof. By Chebyshev's inequality for each $\varepsilon > 0$:

$$\begin{aligned}
\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbb{E}X_i)\right| > \varepsilon\right) &\leq \frac{1}{\varepsilon^2}\mathbb{V}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbb{E}X_i)\right) \\
&= \frac{1}{\varepsilon^2}\frac{1}{n^2}\mathbb{V}\left(\sum_{i=1}^{n}(X_i - \mathbb{E}X_i)\right) \\
&= \frac{1}{\varepsilon^2}\frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}(X_i - \mathbb{E}X_i) \\
&= \frac{1}{\varepsilon^2}\frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}(X_i).
\end{aligned}$$

Here we used that the random variables are pairwise uncorrelated. By assumption the latter expression converges to zero. ■

Remark 9.2 *As we will learn in the next Theorem, for an independent sequence square integrability is even not required.*

Theorem 9.1 raises the question whether we can replace the stochastic convergence there by almost sure convergence. This will be shown in the following theorem. Such a theorem is called a **strong law of large numbers.** Its first form was proved by Kolmogorov. We will present a proof due to Etemadi from 1981.

Theorem 9.3 (Strong Law of Large Numbers – Etemadi 1981) *For each sequence $(X_n)_n$ of real-valued, pairwise independent, identically distributed (integrable) random variables the Strong Law of Large Numbers holds, i.e.*

$$\mathbb{P}\left(\limsup_{n\to\infty}\left|\frac{1}{n}\sum_{i=1}^{n}X_i - \mathbb{E}X_1\right| > \varepsilon\right) = 0 \text{ for each } \varepsilon > 0.$$

Before we prove Theorem 9.3 let us make a couple of remarks. These should reveal the structure of the proof a bit:

1. Denote $S_n = \sum_{i=1}^{n} X_i$. Then Theorem 9.3 asserts that $\frac{1}{n}S_n \to \eta := \mathbb{E}X_1$, $\mathbb{P}$-almost surely.

2. Together with X_n also X_n^+ and X_n^- (where $X_n^+ = \max(X_n, 0)$ and $X_n^- = (-X_n)^+$) satisfy the assumptions of Theorem 9.3. Since $X_n = X_n^+ - X_n^-$ it therefore suffices to prove Theorem 9.3 for positive random variables. We therefore assume $X_n \geq 0$ for the rest of the proof.

3. All proofs of the Strong Law of Large Numbers use the following trick: We truncate the random variables X_n by cutting off values that are too large. We therefore introduce
$$Y_n := X_n 1_{\{|X_n|<n\}} = X_n 1_{\{X_n<n\}}$$

Of course, if μ is the distribution of X_n and μ_n is the distribution of Y_n, then $\mu_n \neq \mu$. Indeed $\mu_n = f_n(\mu)$, where

$$f_n(x) := \begin{cases} x & \text{if } 0 \leq x < n, \\ 0 & \text{otherwise.} \end{cases}$$

The idea behind truncation is that we gain square integrability of the sequence. Indeed:

$$\mathbb{E}\left(Y_n^2\right) = \mathbb{E}\left(f_n^2 \circ X_n\right) = \int f_n^2(x) d\mu(x) = \int_0^n x^2 d\mu(x) < \infty.$$

4. Of course, after having gained information about the Y_n we need to translate these results back to the X_n. To this end we will apply the Borel-Cantelli Lemma and show that

$$\sum_{n=1}^{\infty} \mathbb{P}\left(X_n \neq Y_n\right) < \infty.$$

This implies that $X_n \neq Y_n$ only for finitely many n with probability one. In particular, if we can show that $\frac{1}{n}\sum_{i=1}^n Y_i \to \eta$ $\mathbb{P}$-a.s., we also can show that $\frac{1}{n}\sum_{i=1}^n X_i \to \eta$ $\mathbb{P}$-a.s..

5. For the purposes of the proof we remark the following: Let $\alpha > 1$ and for $n \in \mathbb{N}$ let

$$k_n := [\alpha^n]$$

denote the largest integer $\leq \alpha^n$. This means $k_n \in \mathbb{N}$ and

$$k_n \leq \alpha^n < k_n + 1.$$

Since

$$\lim_{n\to\infty} \frac{\alpha^n - 1}{\alpha^n} = 1$$

there is a number c_α, $0 < c_\alpha < 1$, such that

$$k_n > \alpha^n - 1 \geq c_\alpha \alpha^n \quad \text{for all } n \in \mathbb{N}. \tag{9.1}$$

We now turn to

Proof of Theorem 9.3.

Step 1: Without loss of generality $X_n > 0$. Define $Y_n = 1_{\{X_n < n\}} X_n$. Then Y_n are independent and square integrable. Define

$$S_n' := \sum_{i=1}^n \left(Y_i - \mathbb{E}Y_i\right).$$

Let $\varepsilon > 0$ and $\alpha > 1$. Using Chebyshev's inequality and the independence of the random variables (Y_n) we obtain

$$\mathbb{P}\left(\left|\frac{1}{n}S_n'\right| > \varepsilon\right) \leq \frac{1}{\varepsilon^2}\mathbb{V}\left(\frac{1}{n}S_n'\right) = \frac{1}{\varepsilon^2}\frac{1}{n^2}\mathbb{V}\left(S_n'\right) = \frac{1}{n^2}\frac{1}{\varepsilon^2}\sum_{i=1}^n \mathbb{V}\left(Y_i\right).$$

Observe that $\mathbb{V}(Y_i) = \mathbb{E}(Y_i^2) - (\mathbb{E}(Y_i))^2 \leq \mathbb{E}(Y_i^2)$. Thus

$$\mathbb{P}\left(\left|\frac{1}{n}S'_n\right| > \varepsilon\right) \leq \frac{1}{n^2}\frac{1}{\varepsilon^2}\sum_{i=1}^{n}\mathbb{E}\left(Y_i^2\right).$$

For $k_n = [\alpha^n]$ this gives

$$\mathbb{P}\left(\left|\frac{1}{k_n}S'_{k_n}\right| > \varepsilon\right) \leq \frac{1}{\varepsilon^2 k_n^2}\sum_{i=1}^{k_n}\mathbb{E}\left(Y_i^2\right)$$

for all $n \in \mathbb{N}$. Thus

$$\sum_{n=1}^{\infty}\mathbb{P}\left(\left|\frac{1}{k_n}S'_{k_n}\right| > \varepsilon\right) \leq \frac{1}{\varepsilon^2}\sum_{n=1}^{\infty}\sum_{i=1}^{k_n}\frac{1}{k_n^2}\mathbb{E}\left(Y_i^2\right).$$

By rearranging the order of summation we obtain

$$\sum_{n=1}^{\infty}\mathbb{P}\left(\left|\frac{1}{k_n}S'_{k_n}\right| > \varepsilon\right) \leq \frac{1}{\varepsilon^2}\sum_{j=1}^{\infty}t_j\mathbb{E}\left(Y_j^2\right)$$

where

$$t_j := \sum_{n=n_j}^{\infty}\frac{1}{k_n^2}$$

and n_j is the smallest n with $k_n \geq j$. From (9.1) we obtain

$$t_j \leq \frac{1}{c_\alpha^2}\sum_{n=n_j}^{\infty}\frac{1}{\alpha^{2n}} = \frac{1}{c_\alpha^2}\alpha^{-2n_j}\frac{1}{1-\frac{1}{\alpha^2}} = d_\alpha\alpha^{-2n_j}$$

where $d_\alpha = c_\alpha^{-2}\left(1-\alpha^{-2}\right)^{-1} > 0$. This implies

$$t_j \leq d_\alpha j^{-2}.$$

By using the above

$$\sum_{n=1}^{\infty}\mathbb{P}\left(\left|\frac{1}{k_n}S'_{k_n}\right| > \varepsilon\right) \leq \frac{d_\alpha}{\varepsilon^2}\sum_{j=1}^{\infty}\frac{1}{j^2}\sum_{k=1}^{j}\int_{k-1}^{k}x^2 d\mu(x).$$

Again rearranging the order of summation yields:

$$\sum_{j=1}^{\infty}\frac{1}{j^2}\sum_{k=1}^{j}\int_{k-1}^{k}x^2 d\mu(x) = \sum_{k=1}^{\infty}\left(\sum_{j=k}^{\infty}\frac{1}{j^2}\right)\int_{k-1}^{k}x^2 d\mu(x).$$

Since

$$\begin{aligned}\sum_{j=k}^{\infty}\frac{1}{j^2} &< \frac{1}{k^2} + \frac{1}{k(k+1)} + \frac{1}{(k+1)(k+2)} + \dots \\ &= \frac{1}{k^2} + \left(\frac{1}{k} - \frac{1}{k+1}\right) + \left(\frac{1}{k+1} - \frac{1}{k+2}\right) + \ldots = \frac{1}{k^2} + \frac{1}{k} \leq \frac{2}{k},\end{aligned}$$

this yields

$$\sum_{n=1}^{\infty}\mathbb{P}\left(\left|\frac{1}{k_n}S'_{k_n}\right|>\varepsilon\right)\leq\frac{2d_\alpha}{\varepsilon^2}\sum_{k=1}^{\infty}\int_{k-1}^{k}\frac{x}{k}xd\mu(x)\leq\frac{2d_\alpha}{\varepsilon^2}\sum_{k=1}^{\infty}\int_{k-1}^{k}xd\mu(x)=\frac{2d_\alpha}{\varepsilon^2}\mathbb{E}\left(X_1\right)<\infty.$$

Thus by the Borel-Cantelli Lemma

$$\mathbb{P}\left(\left|\frac{1}{k_n}S'_{k_n}\right|>\varepsilon\text{ infinitely often in }n\right)=0.$$

But this is equivalent with the almost sure convergence of

$$\lim_{n\to\infty}\frac{1}{k_n}S'_{k_n}=0\quad\mathbb{P}\text{-a.s.}\tag{9.2}$$

Step 2: Next let us see that indeed $\frac{1}{k_n}\sum_{i=1}^{k_n}Y_i$ can only converge to $\mathbb{E}\left(X_1\right)$. By definition of Y_n we have that

$$\mathbb{E}\left(Y_n\right)=\int xd\mu_n\left(x\right)=\int_0^n xd\mu\left(x\right).$$

Thus by monotone convergence

$$\mathbb{E}\left(X_1\right)=\lim_{n\to\infty}\mathbb{E}\left(Y_n\right).$$

By Exercise 9.6 below this implies

$$\mathbb{E}\left(X_1\right)=\lim_{n\to\infty}\frac{1}{n}\left(\mathbb{E}Y_1+\ldots+\mathbb{E}Y_n\right).\tag{9.3}$$

By definition of the sums S'_n we have

$$\frac{1}{k_n}S'_{k_n}=\frac{1}{k_n}\sum_{i=1}^{k_n}Y_i-\frac{1}{k_n}\sum_{i=1}^{k_n}\mathbb{E}\left(Y_i\right),$$

(9.2) and (9.3) together imply

$$\lim_{n\to\infty}\frac{1}{k_n}\sum_{i=1}^{k_n}Y_i=\lim_{n\to\infty}\frac{1}{k_n}S'_{k_n}+\lim_{n\to\infty}\frac{1}{k_n}\sum_{i=1}^{k_n}\mathbb{E}Y_i=\mathbb{E}X_1\quad\mathbb{P}\text{-a.s.},$$

which is what we wanted to show in this step.

Step 3: Now we are aiming at removing the truncation from the X_n. Consider the sum

$$\sum_{n=1}^{\infty}\mathbb{P}\left(X_n\neq Y_n\right)=\sum_{n=1}^{\infty}\mathbb{P}\left(X_n\geq n\right)$$

According to Exercise 3.4 this is smaller than $\mathbb{E}(X_1)$, so that it is bounded. Therefore

$$\mathbb{P}\left(X_n\neq Y_n\text{ infinitely often}\right)=0.$$

Hence there is a n_0 (random) such that with probability one $X_n = Y_n$ for all $n \geq n_0$. But the finitely many differences drop out when averaging, hence

$$\lim_{n\to\infty} \frac{1}{k_n} S_{k_n} = \mathbb{E}X_1 \quad \mathbb{P}\text{-a.s.}$$

Step 4: Eventually we show that the theorem holds not only for subsequences k_n chosen as above, but also for the whole sequence.

For fixed $\alpha > 1$, of course, the sequences $(k_n)_n$ are fixed and diverge to $+\infty$. Hence for every $m \in \mathbb{N}$ there exists $n \in \mathbb{N}$ such that

$$k_n < m \leq k_{n+1}.$$

Since we assumed the X_i to be non-negative this implies

$$S_{k_n} \leq S_m \leq S_{k_{n+1}}.$$

Hence

$$\frac{S_{k_n}}{k_n} \cdot \frac{k_n}{m} \leq \frac{S_m}{m} \leq \frac{S_{k_{n+1}}}{k_{n+1}} \cdot \frac{k_{n+1}}{m}.$$

The definition of k_n yields

$$k_n \leq \alpha^n < k_n + 1 \leq m \leq k_{n+1} \leq \alpha^{n+1}.$$

This gives

$$\frac{k_{n+1}}{m} < \frac{\alpha^{n+1}}{\alpha^n} = \alpha$$

as well as

$$\frac{k_n}{m} > \frac{\alpha^n - 1}{\alpha^{n+1}}.$$

Now, given α, for all $n \geq n_1 = n_1(\alpha)$ we have $\alpha^n - 1 \geq \alpha^{n-1}$. Hence, if $m \geq k_{n_1}$ and thus $n \geq n_1$ we obtain

$$\frac{k_n}{m} > \frac{\alpha^n - 1}{\alpha^{n+1}} > \frac{\alpha^{n-1}}{\alpha^{n+1}} = \frac{1}{\alpha^2}$$

Now for each α we have a set Ω_α with $\mathbb{P}(\Omega_\alpha) = 1$ with

$$\lim \frac{1}{k_n} S_{k_n}(\omega) = \mathbb{E}X_1 \text{ for all } \omega \in \Omega_\alpha.$$

Without loss of generality we may assume that X_i are not identically equal to zero $\mathbb{P}$-a.s., otherwise the assertion of the Strong Law of Large Numbers is trivially true. Therefore we may assume without loss of generality that $\mathbb{E}X_1 > 0$. Since $\alpha > 1$ we then have

$$\frac{1}{\alpha}\mathbb{E}X_1 < \frac{1}{k_n} S_{k_n}(\omega) < \alpha \mathbb{E}X_1$$

for all $\omega \in \Omega_\alpha$ and all n large enough. For such m and ω this means

$$\left(\alpha^{-3} - 1\right)\mathbb{E}X_1 < \frac{1}{m} S_m(\omega) - \mathbb{E}X_1 < \left(\alpha^2 - 1\right)\mathbb{E}X_1.$$

Define

$$\Omega_1 := \bigcap_{n=1}^{\infty} \Omega_{1+\frac{1}{n}}.$$

Then $\mathbb{P}(\Omega_1) = 1$ and

$$\lim_{m\to\infty} \frac{1}{m} S_m(\omega) = \mathbb{E}X_1$$

for all $\omega \in \Omega_1$. This proves the theorem. ■

Remark 9.4 *Theorem 9.3 in particular implies that for i.i.d. sequences of random variables (X_n) with a finite first moment the Strong Law of Large Numbers holds true. Since stochastic convergence is implied by almost sure convergence also Theorem 9.1 – the Weak Law of Large Numbers – holds true for such sequences as well. Therefore the finiteness of the second moment is not necessary for Theorem 9.1 to hold true for i.i.d. sequences.*

Remark 9.5 *One might, of course, ask whether a finite first moment is necessary for Theorem 9.3 to hold. Indeed one can prove that, if a sequence of i.i.d. random variables is such that $\frac{1}{n}\sum_{i=1}^{n} X_i$ converges almost surely to some random variable Y (necessarily a tail function as in Corollary 8.7!), then $\mathbb{E}X_1$ exists and $Y = \mathbb{E}X_1$ almost surely. This will not be shown in the context of this course.*

Exercise 9.6 *Let $(a_m)_m$ be real numbers such that $\lim_{m\to\infty} a_m = a$. Show that this implies that their Cesaro mean*

$$\lim_{n\to\infty} \frac{1}{n}(a_1 + a_2 + \ldots + a_n) = a.$$

Exercise 9.7 *Prove the Strong Law of Large Numbers for a sequence of i.i.d. random variables $(X_n)_n$ with a finite fourth moment, i.e. for random variables with $\mathbb{E}(X_1^4) < \infty$. Do **not** use the statement of Theorem 9.3 explicitly.*

Remark 9.8 *A very natural question to ask in the context of Theorem 9.3 is: how fast does $\frac{1}{n}S_n$ converge to $\mathbb{E}X_1$, i.e. given a sequence of i.i.d. random variables $(X_n)_n$ what is*

$$\mathbb{P}\left(\left|\frac{1}{n}\sum X_i - \mathbb{E}X_1\right| \geq \varepsilon\right) ?$$

If X_1 has a finite moment generating function; i.e. if

$$M(t) := \log \mathbb{E}e^{tX_1} < \infty \text{ for all } t,$$

the answer is: exponentially fast. Indeed, Cramér's theorem (which cannot be proven in the context of this course) asserts the following: let $I : \mathbb{R} \to \mathbb{R}$ be given by

$$I(x) = \sup_t [xt - M(t)].$$

Then for every closed set $A \subset \mathbb{R}$

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i \in A\right) \leq -\inf_{x\in A} I(x)$$

and for every open set $\mathcal{O} \subset \mathbb{R}$

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i \in \mathcal{O}\right) \geq -\inf_{x\in\mathcal{O}} I(x).$$

This is called a ***principle of large deviations*** *for the random variables* $\frac{1}{n}\sum_{i=1}^{n} X_i$*. In particular, one can show that the function* I *is convex and non-negative, with*

$$I(x) = 0 \Leftrightarrow x = \mathbb{E}X_1.$$

We therefore obtain

$$\forall \delta > 0 \exists N \forall n > N : \mathbb{P}\left(\left|\frac{1}{n}\sum X_i - \mathbb{E}X_1\right| \geq \varepsilon\right) \leq e^{-n\min(\bar{I}(\mathbb{E}X_1+\varepsilon),\bar{I}(\mathbb{E}X_1-\varepsilon))+n\delta}$$

where $\bar{I}$ *is the* I*-function introduced above evaluated for the random variables* $X_i - \mathbb{E}X_1$*. The speed of convergence is thus exponentially fast.*

Example 9.9 *For a Bernoulli* $B(1,p)$ *random variable* X

$$e^{M(t)} = \mathbb{E}(e^{tX}) = pe^t + (1-p); \qquad I(x) = -x\log\left(\frac{p}{x}\right) - (1-x)\log\left(\frac{1-p}{1-x}\right).$$

Exercise 9.10 *Determine the functions* M *and* I *for a normally distributed random variable.*

Exercise 9.11 *Argue that if the moment generating function of a random variable* X *is finite, all its moments are finite. In particular both Laws of Large Numbers are applicable to a sequence* $X_1, X_2, \ldots$ *of iid random variables distributed like* X*.*

At the end of this section we will turn to two applications of the Law of Large Numbers which are interesting in their own right:

The first of these two applications is in number theory. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be given by $\Omega = [0,1), \mathcal{F} = \mathcal{B}^1\big|\Omega$ and $\mathbb{P} = \lambda^1\big|\,\Omega$ (Lebesgue measure). For every number $\omega \in \Omega$ we may consider its g-adic representation

$$\omega = \sum_{n=1}^{\infty} \xi_n g^{-n} \tag{9.4}$$

Here $g \geq 2$ is a natural number and $\xi_n \in \{0, \ldots, g-1\}$. This representation is unique, if we ask that not all but finitely many of the digits ξ_n are equal to $g-1$. For each $\varepsilon \in \{0, \ldots, g-1\}$ let $S_n^{\varepsilon,g}(\omega)$ be the number of all $i \in \{1, \ldots, n\}$ with $\xi_i(\omega) = \varepsilon$ in its g-adic representation (9.4). We will call a number $\omega \in [0,1)$ g-**normal**[1], if

$$\lim_{n\to\infty} \frac{1}{n} S_n^{\varepsilon,g}(\omega) = \frac{1}{g}$$

[1]The usual meaning of g-normality is that each string of digits $\varepsilon_1\varepsilon_2\ldots\varepsilon_k$ occurs with frequency g^{-k}.

for all $\varepsilon = 0, \ldots, g-1$. Hence ω is g-normal, if in the long run all of its digits occur with the same frequency. We will call ω **absolutely normal**, if ω is g-**normal** for all $g \in \mathbb{N}, g \geq 2$. Now for a number $\omega \in [0,1)$ randomly chosen according to Lebesgue measure the $\xi_i(\omega)$ are i.i.d. random variables; they have as their distribution the uniform distribution on the set $\{0, \ldots, g-1\}$. This has to be shown in Exercise 9.13 below and is a consequence of the uniformity of Lebesgue measure. Hence the random variable

$$X_n(\omega) = \begin{cases} 1 & \text{if } \xi_n(\omega) = \varepsilon \\ 0 & \text{otherwise} \end{cases}$$

are i.i.d. random variables for each g and ε. Moreover $S_n^{\varepsilon,g}(\omega) = \sum_{i=1}^n X_i^{\varepsilon,g}(\omega)$. According to the Strong Law of Large Numbers (Theorem 9.3)

$$\frac{1}{n} S_n^{\varepsilon,g}(\omega) \to \mathbb{E}\left(X_1^{\varepsilon,g}\right) = \frac{1}{g} \quad \lambda^1\text{-a.s.}$$

for all $\varepsilon \in \{0, \ldots, g-1\}$ and all $g \geq 2$. This means λ^1-almost every number ω is g-normal, i.e. there is a set N_g with $\lambda^1(N_g) = 0$, such that ω is g-normal for all $\omega \in N_g^c$. Now

$$N := \bigcup_{g=2}^{\infty} N_g$$

is a set of Lebesgue measure zero as well. This readily implies

Theorem 9.12 *(E. Borel) λ^1-almost every $\omega \in [0,1)$ is absolutely normal.*

It is rather surprising that hardly any normal numbers (in the usual meaning, see Footnote page 33), are known. Champernowne (1933) showed that

$$\omega = 0,1234567891011121314\ldots$$

is 10-normal. Whether $\sqrt{2}, \log 2, e$ or π are normal of any kind has not been shown yet. There are no absolutely normal numbers known at all.

Exercise 9.13 *Show that for every $g \geq 2$, the random variables $\xi_n(\omega)$ introduced above are i.i.d. random variables that are uniformly distributed on $\{0, \ldots, g-1\}$.*

The second application is to derive a classical result from analysis – which in principle has nothing to do with probability theory – is related to the Strong Law of Large Numbers.

As may be well known the approximation theorem by Stone and Weierstraß asserts that every continuous function on $[a,b]$ (more generally on every compact set) can be approximated uniformly by polynomials. Obviously it suffices to prove this for $[a,b] = [0,1]$. So let $f \in \mathcal{C}([0,1])$ be a continuous function on $[0,1]$. Define the n'th Bernstein polynomial for f as

$$B_n f(x) = \sum_{k=0}^{n} \binom{n}{k} f\left(\frac{k}{n}\right) x^k (1-x)^{n-k}.$$

Theorem 9.14 *For each $f \in \mathcal{C}([0,1])$ the polynomials $B_n f$ converge to f uniformly in $[0,1]$.*

Proof. Since f is continuous and $[0,1]$ is compact f is uniformly continuous on $[0,1]$, i.e. for each $\varepsilon > 0$ there exists $\delta > 0$ such that

$$|x - y| < \delta \Rightarrow |f(x) - f(y)| < \varepsilon.$$

Now consider a sequence of i.i.d Bernoullis with parameter p, $(X_n)_n$. Call

$$S_n^* := \frac{1}{n} S_n := \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Then by Chebyshev's inequality

$$\mathbb{P}\left(|S_n^* - p| \geq \delta\right) \leq \frac{1}{\delta^2} \mathbb{V}\left(S_n^*\right) = \frac{1}{n^2 \delta^2} \mathbb{V}(S_n) = \frac{p\,(1-p)}{n\delta^2} \leq \frac{1}{4n\delta^2}. \tag{9.5}$$

This yields

$$\begin{aligned} |B_n f(p) - f(p)| &= \left|\mathbb{E}\left(f \circ S_n^*\right) - f(p)\right| = \left|\int f(x) d\mathbb{P}_{S_n^*}(x) - f(p)\right| \\ &\leq \int_{|S_n^* - p| < \delta} \left|f(S_n^*(x)) - f(p)\right| d\mathbb{P}_{S_n^*}(x) + \\ &\qquad + \int_{|S_n^* - p| \geq \delta} \left|f(S_n^*(x)) - f(p)\right| d\mathbb{P}_{S_n^*}(x) \\ &\leq \varepsilon + 2\left\|f\right\| \mathbb{P}\left(|S_n^* - p| \geq \delta\right) \leq \varepsilon + \frac{2\left\|f\right\|}{4n\delta^2.} \end{aligned}$$

Here $\|f\|$ is the sup-norm of f. Hence

$$\sup_{p \in [0,1]} |B_n f(p) - f(p)| \leq \varepsilon + \frac{2\left\|f\right\|}{4n\delta^2}$$

This can be made smaller than 2ε by choosing n large enough.
Notice that Weak Law of Large Numbers by itself would yield, instead of inequality (9.5), an inequality of the kind

$$\forall \rho > 0 \ \exists N \text{ such that } \forall n \geq N : \mathbb{P}(|S_n^* - p| \geq \delta) \leq \rho.$$

In this approach it is not clear that N can be chosen independently of p, so that we only would get pointwise convergence. ■

10 The Central Limit Theorem

In the previous section we met one of the central theorems of probability theory – the Law of Large Numbers: If $\mathbb{E}X_1$ exists, for sequence of i.i.d. random variables (X_n), their average $\frac{1}{n}\sum_{i=1}^{n} X_i$ converges to $\mathbb{E}X_1$ (a.s.). The following theorem, called the Central Limit

Theorem analyzes the fine structure in the Law of Large Numbers. Its name is due to Polya, the proof of the following theorem goes back to Charles Stein.

First of all notice that in a certain sense, in order to analyze the fine structure of $\sum_{i=1}^n X_i$ the scaling of the Weak Law of Large Numbers $\frac{1}{n}$ is already an overscaling. On this scale we just cannot see the shape of the distribution anymore, since by scaling $\sum_{i=1}^n X_i$ by a factor $\frac{1}{n}$ we have reduced its variance to a scale $\frac{1}{n}$ which converges to zero. What we see in the Law of Large Numbers is a bell shaped curve with a tiny, tiny width. Here is what we get, if we scale the variance to one:

Theorem 10.1 (Central Limit Theorem - CLT) *Let $X_1, \ldots, X_n$ be a sequence of random variables that are independent and have identical distribution (the same for all n) with $\mathbb{E}X_1^2 < \infty$. Then*

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{\sum_{i=1}^n (X_i - \mathbb{E}X_1)}{\sqrt{n\mathbb{V}X_1}} \leq a \right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^a e^{-x^2/2} dx. \tag{10.1}$$

Before proving the Central Limit Theorem let us remark that it holds under weaker assumptions as well

Remark 10.2 *Indeed, the Central Limit Theorem also holds under the following weaker assumption. Assume given for $n = 1, 2, \ldots$ an independent family of random variables X_{ni}, $i = 1, \ldots, n$. For $j = 1, \ldots n$ let*

$$\eta_{nj} = \mathbb{E}X_{nj}$$

and

$$s_n := \sqrt{\sum_{i=1}^n \mathbb{V}X_{ni}}.$$

The sequence $((X_{ni})_{i=1}^n)_n$ is said to satisfy the **Lindeberg condition**, *if*

$$L_n(\varepsilon) \to 0 \text{ as } n \to \infty$$

for all $\varepsilon > 0$. Here

$$L_n(\varepsilon) = \frac{1}{s_n^2} \sum_{j=1}^n \mathbb{E}\left[(X_{nj} - \eta_{nj})^2 ; |X_{nj} - \eta_{nj}| \geq \varepsilon s_n \right].$$

Intuitively speaking the Lindeberg condition asks that none of the variables dominates the whole sum.

The generalized form of the CLT stated above now asserts that if the sequence (X_n) satisfies the **Lindeberg condition**, *it also satisfies the CLT, i.e.*

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{\sum_{i=1}^n (X_{ni} - \eta_{ni})}{\sqrt{\sum_{i=1}^n \mathbb{V}X_{ni}}} \leq a \right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^a e^{-x^2/2} dx.$$

The proof of this more general theorem basically mimicks the proof we will give below for Theorem 10.1. We spare ourselves the additional technical work.

We will present a proof of the CLT that goes back to Charles Stein. It is based on a couple of facts:

Fact 1: It suffices to prove the CLT for i.i.d. random variables with $\mathbb{E}X_1 = 0$. Otherwise one just substracts $\mathbb{E}X_1$ from each of the X_i.

Fact 2: Define

$$S_n := \sum_{i=1}^{n} X_i \quad \text{and } \sigma^2 := \mathbb{V}(X_1)$$

Theorem 10.1 asserts the convergence in distribution of the (normalized) S_n to a Gaussian random variable. What we need to show is thus

$$\mathbb{E}\left[f\left(\frac{S_n}{\sqrt{n\sigma^2}}\right)\right] \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(x) e^{-x^2/2} dx = \mathbb{E}\left[f(Y)\right] \tag{10.2}$$

as $n \to \infty$ for all $f : \mathbb{R} \to \mathbb{R}$ that are uniformly continuous and bounded. Here Y is a standard normal random variable, i.e. it is $\mathcal{N}(0,1)$ distributed.

We prepare the proof of the CLT in two lemmata.

Lemma 10.3 *Let $f : \mathbb{R} \to \mathbb{R}$ be bounded and uniformly continuous. Define*

$$\mathcal{N}(f) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(y) e^{-y^2/2} dy$$

and

$$g(x) := e^{x^2/2} \int_{-\infty}^{x} \left(f(y) - \mathcal{N}(f)\right) e^{-y^2/2} dy.$$

Then g fulfills

$$g'(x) - xg(x) = f(x) - \mathcal{N}(f). \tag{10.3}$$

Proof. Differentiating g gives

$$\begin{aligned} g'(x) &= xe^{x^2/2} \int_{-\infty}^{x} \left(f(y) - \mathcal{N}(f)\right) e^{-y^2/2} dy + e^{x^2/2} \left(f(x) - \mathcal{N}(f)\right) e^{-x^2/2} \\ &= xg(x) + f(x) - \mathcal{N}(f). \end{aligned}$$

■

The importance of the above lemma becomes obvious, if we substitute a random variable X into (10.3) and take expectations:

$$\mathbb{E}\left[g'(X) - Xg(X)\right] = \mathbb{E}\left[f(X) - \mathcal{N}(f)\right].$$

If $X \sim \mathcal{N}(0,1)$ is standard normal, the right hand side is zero and so is the left hand side. The idea is thus that instead of showing that

$$\mathbb{E}\left[f(U_n) - \mathcal{N}(f)\right]$$

converges to zero, we may show the same for

$$\mathbb{E}\left[g'(U_n) - U_n g(U_n)\right].$$

The next step discusses the function g introduced above.

Lemma 10.4 *Let $f : \mathbb{R} \to \mathbb{R}$ be bounded and uniformly continuous and g be the solution of*

$$g'(x) - xg(x) = f(x) - \mathcal{N}(f). \tag{10.4}$$

Then $g(x), xg(x)$ and $g'(x)$ are bounded and continuous

Proof. Obviously g is even differentiable, hence continuous. But then also $xg(x)$ is continuous. Eventually

$$g'(x) = xg(x) + f(x) - \mathcal{N}(f)$$

is continuous as the sum of continuous functions. For the boundedness part first note that any continuous function on a compact set is bounded, hence we only need to check that the functions g, xg, and g' are bounded for $x \to \pm\infty$.

To this end first note that

$$\begin{aligned} g(x) &= e^{x^2/2} \int_{-\infty}^{x} (f(y) - \mathcal{N}(f))\, e^{-y^2/2} dy \\ &= -e^{x^2/2} \int_{x}^{\infty} (f(y) - \mathcal{N}(f))\, e^{-y^2/2} dy \end{aligned}$$

(this is true since the whole integral must equal zero).

For $x \leq 0$ we have

$$g(x) \leq \sup_{y \leq 0} |f(y) - \mathcal{N}(f)|\, e^{x^2/2} \int_{-\infty}^{x} e^{-y^2/2} dy$$

while for $x \geq 0$ we have

$$g(x) \leq \sup_{y \geq 0} |f(y) - \mathcal{N}(f)|\, e^{x^2/2} \int_{x}^{\infty} e^{-y^2/2} dy.$$

Now for $x \leq 0$

$$e^{x^2/2} \int_{-\infty}^{x} e^{-y^2/2} dy \leq e^{x^2/2} \int_{-\infty}^{x} \frac{-y}{|x|} e^{-y^2/2} dy = \frac{1}{|x|}$$

and similarly for $x \geq 0$

$$e^{x^2/2} \int_{x}^{\infty} e^{-y^2/2} dy \leq e^{x^2/2} \int_{x}^{\infty} \frac{y}{x} e^{-y^2/2} dy \leq \frac{1}{|x|}. \tag{10.5}$$

Thus we see that for $x \geq 1$

$$|g(x)| \leq |xg(x)| \leq \sup_{y \geq 0} |f(y) - \mathcal{N}(f)|$$

as well as for $x \leq -1$

$$|g(x)| \leq |xg(x)| \leq \sup_{y \leq 0} |f(y) - \mathcal{N}(f)|\,.$$

Hence g and xg are bounded. But then also g' is bounded, since

$$g'(x) = xg(x) + f(x) - \mathcal{N}(f).$$

■

Now we turn to proving the CLT.

Proof of Theorem 10.1. Besides assuming that $\mathbb{E}X_i = 0$ for all i we may also assume that $\sigma^2 = \mathbb{V}X_1 = 1$. Otherwise we just replace X_i by $X_i/\sqrt{\sigma^2}$. We write $S_n := \sum_{i=1}^n X_i$ and recall that in order to prove the assertion it suffices that for all f bounded and continuous

$$\mathbb{E}\left[g'\left(\frac{S_n}{\sqrt{n}}\right) - \frac{S_n}{\sqrt{n}} g\left(\frac{S_n}{\sqrt{n}}\right)\right] \to 0$$

as $n \to \infty$. Here g is defined as above.

But using the identity

$$\begin{aligned}\int_0^1 \frac{X_j}{\sqrt{n}}\left[g'\left(\frac{S_n}{\sqrt{n}} - (1-s)\frac{X_j}{\sqrt{n}}\right) - g'\left(\frac{S_n - X_j}{\sqrt{n}}\right)\right] ds \\ = g\left(\frac{S_n}{\sqrt{n}}\right) - g\left(\frac{S_n - X_j}{\sqrt{n}}\right) - \frac{X_j}{\sqrt{n}} g'\left(\frac{S_n - X_j}{\sqrt{n}}\right)\end{aligned}$$

which is to be proven in Exercise 10.5 below we arrive at

$$\begin{aligned}&\mathbb{E}\left[g'\left(\frac{S_n}{\sqrt{n}}\right) - \frac{S_n}{\sqrt{n}} g\left(\frac{S_n}{\sqrt{n}}\right)\right] \\ = &\sum_{j=1}^n \mathbb{E}\left[\frac{1}{n} g'\left(\frac{S_n}{\sqrt{n}}\right) - \frac{X_j}{\sqrt{n}} g\left(\frac{S_n}{\sqrt{n}}\right)\right] \\ = &\sum_{j=1}^n \mathbb{E}\left[\frac{1}{n} g'\left(\frac{S_n}{\sqrt{n}}\right) - \frac{X_j}{\sqrt{n}} g\left(\frac{S_n - X_j}{\sqrt{n}}\right) - \frac{X_j^2}{n} g'\left(\frac{S_n - X_j}{\sqrt{n}}\right)\right. \\ &\qquad \left. - \frac{X_j^2}{n}\int_0^1 g'\left(\frac{S_n}{\sqrt{n}} - (1-s)\frac{X_j}{\sqrt{n}}\right) - g'\left(\frac{S_n - X_j}{\sqrt{n}}\right) ds\right] \\ = &\sum_{j=1}^n \mathbb{E}\left[\frac{1}{n} g'\left(\frac{S_n}{\sqrt{n}}\right) - \frac{1}{n} g'\left(\frac{S_n - X_j}{\sqrt{n}}\right)\right. \\ &\qquad \left. - \frac{X_j^2}{n}\int_0^1 g'\left(\frac{S_n}{\sqrt{n}} - (1-s)\frac{X_j}{\sqrt{n}}\right) - g'\left(\frac{S_n - X_j}{\sqrt{n}}\right) ds\right].\end{aligned}$$

In the last step we used linearity of expectation together with $\mathbb{E}X_i = 0$ for all i as well as independence of X_i and $S_n - X_i$ for all i together with $\mathbb{E}X_i^2 = 1$. Let us define

$$\begin{aligned}\Gamma_j := \mathbb{E}\left[\frac{1}{n} g'\left(\frac{S_n}{\sqrt{n}}\right) - \frac{1}{n} g'\left(\frac{S_n - X_j}{\sqrt{n}}\right)\right. \\ \left. - \frac{X_j^2}{n}\int_0^1 g'\left(\frac{S_n}{\sqrt{n}} - (1-s)\frac{X_j}{\sqrt{n}}\right) - g'\left(\frac{S_n - X_j}{\sqrt{n}}\right) ds\right].\end{aligned}$$

The idea will now be that the continuous function g' is uniformly continuous on every compact set. So, if $\frac{X_j}{\sqrt{n}}$ is small, so are $g'\left(\frac{S_n}{\sqrt{n}}\right) - \frac{1}{n} g'\left(\frac{S_n - X_j}{\sqrt{n}}\right)$ and $g'\left(\frac{S_n}{\sqrt{n}} - (1-s)\frac{X_j}{\sqrt{n}}\right) - g'\left(\frac{S_n - X_j}{\sqrt{n}}\right)$ as long as $\frac{S_n}{\sqrt{n}}$ is inside the chosen compact set. On the other hand the probabilities that $\frac{S_n}{\sqrt{n}}$ is outside a chosen large compact set or that $\frac{X_j}{\sqrt{n}}$ is large are very small. This

together with the boundedness of g and g' will basically yield the proof. For $K > 0, \delta > 0$ we write

$$\Gamma_j^1 := \Gamma_j 1_{|\frac{X_j}{\sqrt{n}}| \le \delta} 1_{|\frac{S_n}{\sqrt{n}}| \le K}$$
$$\Gamma_j^2 := \Gamma_j 1_{|\frac{X_j}{\sqrt{n}}| \le \delta} 1_{|\frac{S_n}{\sqrt{n}}| > K}$$
$$\Gamma_j^3 := \Gamma_j 1_{|\frac{X_j}{\sqrt{n}}| > \delta}.$$

Hence

$$\sum_{j=1}^n \Gamma_j = \sum_{j=1}^n \Gamma_j^1 + \Gamma_j^2 + \Gamma_j^3 = \sum_{j=1}^n \Gamma_j^1 + \sum_{j=1}^n \Gamma_j^2 + \sum_{j=1}^n \Gamma_j^3.$$

We first consider the Γ_j^2-terms:

By Chebyshev's inequality

$$\mathbb{P}(|\frac{S_n}{\sqrt{n}}| > K) \le \frac{\mathbb{V}\left(\frac{S_n}{\sqrt{n}}\right)}{K^2} = \frac{1}{K^2} \text{ and } \mathbb{P}(|\frac{S_{n-1}}{\sqrt{n}}| > K - \delta) \le \frac{\mathbb{V}\left(\frac{S_{n-1}}{\sqrt{n}}\right)}{(K-\delta)^2} = \frac{n-1}{n(K-\delta)^2}.$$

Hence for given $\varepsilon > 0$ we can find K so large that

$$\mathbb{P}(|\frac{S_n}{\sqrt{n}}| > K) \le \varepsilon \text{ and } \mathbb{P}(|\frac{S_{n-1}}{\sqrt{n}}| > K - \delta) \le \varepsilon.$$

Since g' is bounded by $||g'|| := \sup_{x \in \mathbb{R}} |g'(x)|$ we obtain:

$$\begin{aligned}
\sum_{j=1}^n |\Gamma_j^2| &= \sum_{j=1}^n \Bigg| \mathbb{E}\left[\frac{1}{n} g'\left(\frac{S_n}{\sqrt{n}}\right) - \frac{1}{n} g'\left(\frac{S_n - X_j}{\sqrt{n}}\right) \right. \\
&\qquad \left. - \frac{X_j^2}{n} \int_0^1 g'\left(\frac{S_n}{\sqrt{n}} - (1-s)\frac{X_j}{\sqrt{n}}\right) - g'\left(\frac{S_n - X_j}{\sqrt{n}}\right) ds \right] 1_{|\frac{X_j}{\sqrt{n}}| \le \delta} 1_{|\frac{S_n}{\sqrt{n}}| > K} \Bigg| \\
&\le \sum_{j=1}^n \mathbb{E}\left[\frac{1}{n} 2||g'|| 1_{|\frac{X_j}{\sqrt{n}}| \le \delta} 1_{|\frac{S_n}{\sqrt{n}}| > K}\right] + \frac{2||g'||}{n} \mathbb{E}\left[X_j^2 1_{|\frac{X_j}{\sqrt{n}}| \le \delta} 1_{|\frac{S_n}{\sqrt{n}}| > K}\right] \\
&\le \sum_{j=1}^n \mathbb{E}\left[\frac{1}{n} 2||g'|| 1_{|\frac{X_j}{\sqrt{n}}| \le \delta} 1_{|\frac{S_n}{\sqrt{n}}| > K}\right] + \frac{2||g'||}{n} \mathbb{E}\left[X_j^2 1_{|\frac{X_j}{\sqrt{n}}| \le \delta} 1_{|\frac{S_n - X_j}{\sqrt{n}}| > K - \delta}\right] \\
&\le 2||g'|| \mathbb{E}\left[1_{|\frac{S_n}{\sqrt{n}}| > K}\right] + \frac{2||g'||}{n} \sum_j \mathbb{E}\left[X_j^2 1_{|\frac{X_j}{\sqrt{n}}| \le \delta}\right] \mathbb{E}\left[1_{|\frac{S_n - X_j}{\sqrt{n}}| > K - \delta}\right] \\
&= 2||g'|| \mathbb{P}\left[\left|\frac{S_n}{\sqrt{n}}\right| > K\right] + 2||g'|| \mathbb{P}\left[\left|\frac{S_{n-1}}{\sqrt{n}}\right| > K - \delta\right] \\
&\le 4||g'|| \, \varepsilon
\end{aligned}$$

For the Γ_j^1-terms observe that for every fixed $K > 0$ the continuous function g' is uniformly continuous on $[-K, K]$. This means that given $\varepsilon > 0$, there is $\delta > 0$ such that

$$|x - y| < \delta \Rightarrow |g'(x) - g'(y)| < \varepsilon.$$

For given $\varepsilon > 0$ we choose such δ, and K as in the first step. Then

$$\begin{aligned}
\sum_{j=1}^{n} |\Gamma_j^1| &= \sum_{j=1}^{n} \left| \mathbb{E}\left[\frac{1}{n} g'\left(\frac{S_n}{\sqrt{n}}\right) - \frac{1}{n} g'\left(\frac{S_n - X_j}{\sqrt{n}}\right) \right.\right. \\
&\qquad \left.\left. - \frac{X_j^2}{n} \int_0^1 g'\left(\frac{S_n}{\sqrt{n}} - (1-s)\frac{X_j}{\sqrt{n}}\right) - g'\left(\frac{S_n - X_j}{\sqrt{n}}\right) ds \right] 1_{|\frac{X_j}{\sqrt{n}}| \le \delta} 1_{|\frac{S_n}{\sqrt{n}}| \le K} \right| \\
&\le \sum_{j=1}^{n} \left| \mathbb{E}\left[\frac{1}{n} g'\left(\frac{S_n}{\sqrt{n}}\right) - \frac{1}{n} g'\left(\frac{S_n - X_j}{\sqrt{n}}\right) \right] 1_{|\frac{X_j}{\sqrt{n}}| \le \delta} 1_{|\frac{S_n}{\sqrt{n}}| \le K} \right| \\
&\qquad + \left| \left[\frac{X_j^2}{n} \int_0^1 g'\left(\frac{S_n}{\sqrt{n}} - (1-s)\frac{X_j}{\sqrt{n}}\right) - g'\left(\frac{S_n - X_j}{\sqrt{n}}\right) ds \right] 1_{|\frac{X_j}{\sqrt{n}}| \le \delta} 1_{|\frac{S_n}{\sqrt{n}}| \le K} \right| \\
&\le \sum_{j=1}^{n} \left(\frac{1}{n}\varepsilon + \frac{\mathbb{E}X_j^2}{n}\varepsilon \right) \\
&= n\left(\frac{2\varepsilon}{n}\right) = 2\varepsilon.
\end{aligned}$$

Eventually we turn to the Γ_j^3-terms:

Since $\mathbb{E}X_1^2 < \infty$ there exists an n_0 such that for a given $\varepsilon > 0$ and all $n \ge n_0$ and δ as above we have

$$\mathbb{E}\left(X_1^2 1_{\left|\frac{X_1}{\sqrt{n}}\right| > \delta} \right) < \varepsilon.$$

This implies

$$\begin{aligned}
\sum_{j=1}^{n} |\Gamma_j^3| &= \sum_{j=1}^{n} \left| \mathbb{E}\left[\frac{1}{n} g'\left(\frac{S_n}{\sqrt{n}}\right) - \frac{1}{n} g'\left(\frac{S_n - X_j}{\sqrt{n}}\right) \right.\right. \\
&\qquad \left.\left. - \frac{X_j^2}{n} \int_0^1 g'\left(\frac{S_n}{\sqrt{n}} - (1-s)\frac{X_j}{\sqrt{n}}\right) - g'\left(\frac{S_n - X_j}{\sqrt{n}}\right) ds \right] 1_{|\frac{X_j}{\sqrt{n}}| > \delta} \right| \\
&\le \sum_{j=1}^{n} \left(\frac{2}{n}\mathbb{E}1_{|\frac{X_j}{\sqrt{n}}| > \delta} ||g'|| + \frac{2}{n}||g'||\mathbb{E}\left[X_j^2 1_{|\frac{X_j}{\sqrt{n}}| > \delta} \right] \right) \\
&\le 2||g'||\mathbb{P}\left(\left| \frac{X_j}{\sqrt{n}} \right| > \delta \right) + 2||g'||\mathbb{E}\left[X_j^2 1_{|\frac{X_j}{\sqrt{n}}| > \delta} \right] \\
&\le 4\varepsilon||g'||
\end{aligned}$$

Hence for a given $\varepsilon > 0$ with the choice of $\delta > 0$ and K as above we obtain

$$\begin{aligned}
&\mathbb{E}\left[g'\left(\frac{S_n}{\sqrt{n}}\right) - \frac{S_n}{\sqrt{n}} g\left(\frac{S_n}{\sqrt{n}}\right) \right] \\
&= \sum_{j=1}^{n} \Gamma_j^1 + \sum_{j=1}^{n} \Gamma_j^2 + \sum_{j=1}^{n} \Gamma_j^3 \\
&\le 2\varepsilon + 8\varepsilon||g'||.
\end{aligned}$$

This can be made arbitrarily small by letting $\varepsilon \to 0$. This proves the theorem. ■

Exercise 10.5 *Let $X_1, \ldots, X_n$ be i.i.d. random variables and $S_n = \sum_{i=1}^n X_i$. Let*

$$g : \mathbf{R} \to \mathbf{R}$$

be a continuously differentiable function. Show that for all j

$$\int_0^1 [g'(S_n - (1-s)X_j) - g'(S_n - X_j)]X_j ds = g(S_n) - g(S_n - X_j) - X_j g'(S_n - X_j).$$

We conclude the section with the informal discussion of two extensions on the Central Limit Theorem. The first is of practical importance, the second is of more theoretical interest.

When one tries to apply the Central Limit Theorem, e.g. for a sequence of i.i.d. random variables, it is of course not only important to know that

$$\tilde{X}_n := \frac{\sum_{i=1}^n (X_i - EX_1)}{\sqrt{n\mathbb{V}X_1}}$$

converges to a random variable $Z \sim \mathcal{N}(0,1)$. One also needs to know, how close the distributions of $\tilde{X}_n$ and Z are. This is stated in the following theorem due to Berry and Esseen:

Theorem 10.6 (Berry-Esseen) *Let $(X_i)_{i\in\mathbb{N}}$ be a sequence of i.i.d. random variables with $\mathbb{E}(|X_1|^3) < \infty$. Then for a $\mathcal{N}(0,1)$-distributed random variable Z it holds:*

$$\sup_{a\in\mathbb{R}} \left| \mathbb{P}\left(\frac{\sum_{i=1}^n X_i - \mathbb{E}X_1}{\sqrt{n\mathbb{V}X_1}} \le a \right) - \mathbb{P}(Z \le a) \right| \le \frac{C}{\sqrt{n}} \frac{\mathbb{E}(|X_1 - \mathbb{E}X_1|^3)}{(\mathbb{V}X_1)^{3/2}}.$$

The numerical value of C is below 6 and larger than 0.4. This is rather easy to prove.

The second extension of the Central Limit Theorem starts with the following observation: Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables with finite variance and expectation zero. Then the law of large numbers says that $\frac{1}{n}\sum_{i=1}^n X_i$ converges to $\mathbb{E}X_1 = 0$ in probability and almost surely. But it tells nothing about the size of the fluctuations. This is considered in greater detail by the Central Limit Theorem. The latter describes the asymptotic probabilities that

$$\mathbb{P}\left(\frac{\sum_{i=1}^n (X_i - \mathbb{E}X_1)}{\sqrt{n\mathbb{V}(X_1)}} \ge a \right).$$

Since these probabilities are positive for all $a \in \mathbb{R}$ according to the Central Limit Theorem, it can be shown that the fluctuations of $\sum_{i=1}^n X_i$ are larger than $\sqrt{n}$, more precisely, for each positive $a \in \mathbb{R}$ it holds with probability one that $\limsup \frac{\sum_{i=1}^n X_i}{\sqrt{n}} \ge a$.

The question for the precise size of the fluctuations, i.e. for the right scaling (a_n) such that

$$\limsup \frac{\sum_{i=1}^n X_i}{\sqrt{a_n n}}$$

is almost surely finite, is answered by the law of the iterated logarithm:

Theorem 10.7 (Law of the Iterated Logarithm by Hartmann and Winter) *Let $(X_i)_{i\in\mathbb{N}}$ be a sequence of i.i.d. random variables with $\sigma^2 := \mathbb{V}X_1 < \infty\,(\sigma > 0)$. Then for $S_n := \sum_{i=1}^n X_i$ it holds*

$$\limsup \frac{S_n}{\sqrt{2n\log\log n}} = +\sigma \qquad \mathbb{P}\text{-}a.s.$$

and

$$\liminf \frac{S_n}{\sqrt{2n\log\log n}} = -\sigma \qquad \mathbb{P}\text{-}a.s.$$

Due to the restricted time we will not be able to prove the Law of the Iterated Logarithm in the context of this course. Despite its theoretical interest its practical relevance is rather limited. To understand why, notice that the correction to the $\sqrt{n\mathbb{V}X_1}$ from the Central Limit Theorem to the Law of the Iterated Logarithm are of order $\sqrt{\log\log n}$. Even for a fantastically large number of observation, 10^{100} (which is more than one observation per atom in the universe) $\sqrt{\log\log n}$ is really small,e.g.

$$\sqrt{\log\log 10^{100}} = \sqrt{\log(100\log 10)} = \sqrt{\log 100 + \log\log 10} \simeq \sqrt{6.13} \simeq 2.47.$$

11 Conditional Expectation

To understand the concept of conditional expectation, we will start with a little example.

Example 11.1 *Let Ω be a finite population and let the random variable $X(\omega)$ denote the income of person ω. So, if we are only interested in income, X contains the full information of our experiment. Now assume we are a sociologist and want to measure the influence of a person's religion on his income. So we are not interested in the full information given by X, but only in how X behaves on each of the sets,*

$$\{catholic\}, \{protestant\}, \{islamic\}, \{jewish\}, \{atheist\},$$

etc. This leads to the concept of conditional expectation.

The basic idea of conditional expectation will be that given a random variable

$$X : (\Omega, \mathcal{F}) \to \mathbb{R}$$

and a sub-σ-algebra $\mathcal{A}$ of $\mathcal{F}$ to introduce a new random variable called $\mathbb{E}[X \mid \mathcal{A}] =: X_0$ such that X_0 is $\mathcal{A}$-measurable and

$$\int_C X_0 d\mathbb{P} = \int_C X d\mathbb{P}$$

for all $C \in \mathcal{A}$. So X_0 contains all information necessary when we only consider events in $\mathcal{A}$. First we need to see that such a X_0 can be found in a unique way.

Theorem 11.2 *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and X an integrable random variable. Let $\mathcal{C} \subseteq \mathcal{F}$ be a sub-σ-algebra. Then (up to $\mathbb{P}$-a.s. equality) there is a unique random variable X_0, which is $\mathcal{C}$-measurable and satisfies*

$$\int_C X_0 d\mathbb{P} = \int_C X d\mathbb{P} \quad \text{for all } C \in \mathcal{C}. \tag{11.1}$$

If $X \geq 0$, then $X_0 \geq 0$ $\mathbb{P}$-a.s.

Proof. First we treat the case $X \geq 0$. Denote $P_0 := \mathbb{P}\,|\mathcal{C}$ and $Q = X\mathbb{P}\,|\mathcal{C}$. Both, P_0 and Q are measures on $\mathcal{C}$, P_0 even is a probability measure. By definition

$$Q(C) = \int_C X d\mathbb{P}.$$

Hence $Q(C) = 0$ for all C with $\mathbb{P}(C) = 0 = \mathbb{P}_0(C)$. Hence $Q \ll P_0$. By the theorem of Radon-Nikodym there is a $\mathcal{C}$-measurable function $X_0 \geq 0$ on Ω such that $Q = X_0 P_0$. Thus

$$\int_C X_0 dP_0 = \int_C X d\mathbb{P} \text{ for all } C \in \mathcal{C}.$$

Hence

$$\int_C X_0 d\mathbb{P} = \int_C X d\mathbb{P} \text{ for all } C \in \mathcal{C}.$$

Hence X_0 satisfies (11.1). For $\bar{X}_0$ that is $\mathcal{C}$-measurable and satisfies (11.1) the set $C = \left\{\bar{X}_0 < X_0\right\}$ is in $\mathcal{C}$ and $\int_C \bar{X}_0 d\mathbb{P} = \int_C X_0 d\mathbb{P}$, whence $\mathbb{P}(C) = 0$. In the same way $\mathbb{P}(\left\{\bar{X}_0 > X_0\right\} = 0$. Therefore $\bar{X}_0$ is $\mathbb{P}$-a.s. equal to X_0.

The proof for arbitrary, integrable X is left to the reader. ■

Exercise 11.3 *Prove Theorem 11.2 for arbitrary, integrable $X : \Omega \to \mathbb{R}$.*

Definition 11.4 *Under the conditions of Theorem 11.2 the random variable X_0 (which is $\mathbb{P}$-a.s. unique) is called the conditional expectation of X given $\mathcal{C}$. It is denoted by*

$$X_0 =: \mathbb{E}[X \mid \mathcal{C}] =: \mathbb{E}^{\mathcal{C}}[X].$$

If $\mathcal{C}$ is generated by a sequence of random variable $(Y_i)_{i \in I}$ such that $\mathcal{C} = \sigma(Y_i, i \in I)$ we write

$$\mathbb{E}\left[X \mid (Y_i)_{i \in I}\right] = \mathbb{E}[X \mid \mathcal{C}].$$

If $I = \{1, \ldots, n\}$ we also write $\mathbb{E}[X \mid Y_1, \ldots, Y_n]$.

Note that, in order to check whether Y (Y $\mathcal{C}$-measurable) is a conditional expectation of X given the sub-σ-algebra $\mathcal{C}$ we need to check

$$\int_C Y d\mathbb{P} = \int_C X d\mathbb{P}$$

for all $C \in \mathcal{C}$. This determines $\mathbb{E}[X \mid \mathcal{C}]$ only $\mathbb{P}$-a.s. on sets $C \in \mathcal{C}$. We therefore also speak about different **versions** of conditional expectation.

Example 11.5 *1. If $\mathcal{C} = \{\emptyset, \Omega\}$, then the constant random variable $\mathbb{E}X$ is a version of $\mathbb{E}[X \mid \mathcal{C}]$. Indeed if $C = \emptyset$, then any variable does the job. If $C = \Omega$*

$$\int_C X d\mathbb{P} = \mathbb{E}X = \int \mathbb{E}X d\mathbb{P}.$$

2. If $\mathcal{C}$ is generated by the family $(B_i)_{i\in I}$ of mutually disjoint sets (i.e. $B_i \cap B_j = \emptyset$ if $i \neq j$), where I is countable and $B_i \in \mathcal{A}$ (the original space being $(\Omega, \mathcal{A}, \mathbb{P})$) and $\mathbb{P}(B_i) > 0$ then

$$\mathbb{E}[X \mid \mathcal{C}] = \sum_{i\in I} \frac{1}{\mathbb{P}(B_i)} 1_{B_i} \int_{B_i} X d\mathbb{P} \quad \mathbb{P}\text{-a.s.}$$

be checked in the following exercise.

Exercise 11.6 *Show that the assertion of Example 11.5.2. is true.*

Exercise 11.7 *Show that the following assertions for the conditional expectation $\mathbb{E}[X \mid \mathcal{C}]$ of random variables*

$$X, Y : (\Omega, \mathcal{A}) \to \left(\mathbb{R}, \mathcal{B}^1\right)$$

($\mathcal{C} \subset \mathcal{A}$) are true:

1. $\mathbb{E}[\mathbb{E}[X \mid \mathcal{C}]] = \mathbb{E}X$
2. *If X is $\mathcal{C}$-measurable then $\mathbb{E}[X \mid \mathcal{C}] = X$ $\mathbb{P}$-a.s.*
3. *If $X = Y$ $\mathbb{P}$-a.s., then $\mathbb{E}[X \mid \mathcal{C}] = \mathbb{E}[Y \mid \mathcal{C}]$ $\mathbb{P}$-a.s.*
4. *If $X \equiv \alpha$, then $\mathbb{E}[X \mid \mathcal{C}] = \alpha$ $\mathbb{P}$-a.s.*
5. $\mathbb{E}[\alpha X + \beta Y \mid \mathcal{C}] = \alpha\mathbb{E}[X \mid \mathcal{C}] + \beta\mathbb{E}[Y \mid \mathcal{C}]$ $\mathbb{P}$-a.s. *Here $\alpha, \beta \in \mathbb{R}$.*
6. $X \leq Y$ $\mathbb{P}$-a.s. *implies* $\mathbb{E}[X \mid \mathcal{C}] \leq \mathbb{E}[Y \mid \mathcal{C}]$ $\mathbb{P}$-a.s.

The following theorems have proofs that are almost identical with the proofs of the theorems for expectations:

Theorem 11.8 (monotone convergence) *Let (X_n) be an increasing sequence of positive random variables with $X = \sup X_n$, X integrable , then*

$$\sup_n \mathbb{E}[X_n \mid \mathcal{C}] = \lim_{n\to\infty} \mathbb{E}[X_n \mid \mathcal{C}] = \mathbb{E}\left[\lim_{n\to\infty} X_n \mid \mathcal{C}\right] = \mathbb{E}[X \mid \mathcal{C}].$$

Theorem 11.9 (dominated convergence) *Let (X_n) be a sequence of random variables converging pointwise to an (integrable) random variable X, such that there is an integrable random variable Y with $Y \geq |X_n|$, then*

$$\lim_{n\to\infty} \mathbb{E}[X_n \mid \mathcal{C}] = \mathbb{E}[X \mid \mathcal{C}].$$

Also Jensen's inequality has a generalization to conditional expectations:

Theorem 11.10 (Jensen's inequality) *Let X be an integrable random variable taking values in an open interval $I \subset \mathbb{R}$ and let*

$$q : I \to \mathbb{R}$$

be a convex function. Then for each $\mathcal{C} \subset \mathcal{A}$ it holds

$$\mathbb{E}[X \mid \mathcal{C}] : \Omega \to I$$

and

$$q(\mathbb{E}[X \mid \mathcal{C}]) \leq \mathbb{E}[q \circ X \mid \mathcal{C}].$$

An immediate consequence of Theorem 11.10 is the following (for $p \geq 1$):

$$|\mathbb{E}[X \mid \mathcal{C}]|^p \leq \mathbb{E}[|X|^p \mid \mathcal{C}]$$

which implies

$$\mathbb{E}(|\mathbb{E}[X \mid \mathcal{C}]|^p) \leq \mathbb{E}(|X|^p).$$

Denoting by

$$N_p(f) = \left(\int |f|^p \, d\mathbb{P} \right)^{1/p}$$

this means

$$N_p(\mathbb{E}[X \mid \mathcal{C}]) \leq N_p(X), \quad X \in \mathcal{L}^p(\mathbb{P}).$$

This holds for $1 \leq p < \infty$. $N_p(f)$ is called the $\mathcal{L}^p$-norm of f. The case $p = \infty$, which means that if X is bounded $\mathbb{P}$-a.s. by some $M \geq 0$, then so is $\mathbb{E}[X \mid \mathcal{C}]$, follows from Exercise 11.7.

We slightly reformulate the definition of conditional expectation to discuss its further properties.

Lemma 11.11 *Let X be a positive integrable function. Let $X_0 : (\Omega, \mathcal{A}) \to (\mathbb{R}, \mathcal{B}^1)$ a positive $\mathcal{C}$-measurable integrable random variable that is a version of $\mathbb{E}[X \mid \mathcal{C}]$ (X integrable), then*

$$\int Z X_0 \, d\mathbb{P} = \int Z X \, d\mathbb{P} \tag{11.2}$$

for all $\mathcal{C}$-measurable, positive random variables Z.

Proof. From (11.1) we obtain (11.2) for step functions. The general result follows from monotone convergence. ■

We are now prepared to show a number of properties of conditional expectations which we will call smoothing properties

Theorem 11.12 (Smoothing properties of conditional expectations) *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be probability space and $X \in \mathcal{L}^p(\mathbb{P})$ and $Y \in \mathcal{L}^q(\mathbb{P})$, $1 \leq p \leq \infty$, $\frac{1}{p} + \frac{1}{q} = 1$.*

1. *If $\mathcal{C} \subseteq \mathcal{F}$ and X is $\mathcal{C}$-measurable then*

$$\mathbb{E}[XY \mid \mathcal{C}] = X\mathbb{E}[Y \mid \mathcal{C}] \tag{11.3}$$

2. *If* $\mathcal{C}_1, \mathcal{C}_2 \subseteq \mathcal{F}$ *with* $\mathcal{C}_1 \subseteq \mathcal{C}_2$ *then*

$$\mathbb{E}\left[\mathbb{E}\left[X \mid \mathcal{C}_2\right] \mid \mathcal{C}_1\right] = \mathbb{E}\left[\mathbb{E}\left[X \mid \mathcal{C}_1\right] \mid \mathcal{C}_2\right] = \mathbb{E}\left[X \mid \mathcal{C}_1\right].$$

Proof.

1. First assume that $X, Y \geq 0$. Let X be $\mathcal{C}$-measurable and $C \in \mathcal{C}$. Then

$$\int_C XY d\mathbb{P} = \int 1_C XY d\mathbb{P} = \int 1_C X \mathbb{E}\left[Y \mid \mathcal{C}\right] d\mathbb{P} = \int_C X \mathbb{E}\left[Y \mid \mathcal{C}\right] d\mathbb{P}.$$

   Indeed, this follows immediately from lemma 11.11 since $1_C X$ is $\mathcal{C}$-measurable. On the other hand, we also have $XY \in \mathcal{L}^1(\mathbb{P})$ and

$$\int_C XY d\mathbb{P} = \int_C \mathbb{E}\left[XY \mid \mathcal{C}\right] d\mathbb{P}.$$

   Since $X\mathbb{E}\left[Y \mid \mathcal{C}\right]$ is $\mathcal{C}$-measurable we obtain

$$\mathbb{E}\left[XY \mid \mathcal{C}\right] = X\mathbb{E}\left[Y \mid \mathcal{C}\right] \quad \mathbb{P}\text{-a.s.}$$

   In the case $X \in \mathcal{L}^p\left(\mathbb{P}\right), Y \in \mathcal{L}^q\left(\mathbb{P}\right)$ we observe that then $XY \in \mathcal{L}^1\left(\mathbb{P}\right)$ and conclude as above.

2. Observe that, of course, $\mathbb{E}\left[X \mid \mathcal{C}_1\right]$ is $\mathcal{C}_1$-measurable and, since $\mathcal{C}_1 \subseteq \mathcal{C}_2$, also $\mathcal{C}_2$-measurable. Property 2 in Exercise 11.7 than implies

$$\mathbb{E}\left[\mathbb{E}\left[X \mid \mathcal{C}_1\right] \mid \mathcal{C}_2\right] = \mathbb{E}\left[X \mid \mathcal{C}_1\right], \quad \mathbb{P}\text{-a.s.}$$

   Moreover for all $C \in \mathcal{C}_1$

$$\int_C \mathbb{E}\left[X \mid \mathcal{C}_1\right] d\mathbb{P} = \int_C X d\mathbb{P}.$$

   Hence for all $C \in \mathcal{C}_1$

$$\int_C \mathbb{E}\left[X \mid \mathcal{C}_1\right] d\mathbb{P} = \int_C \mathbb{E}\left[X \mid \mathcal{C}_2\right] d\mathbb{P}.$$

   But this means

$$\mathbb{E}\left[\mathbb{E}\left[X \mid \mathcal{C}_2\right] \mid \mathcal{C}_1\right] = \mathbb{E}\left[X \mid \mathcal{C}_1\right] \quad \mathbb{P}\text{-a.s.}$$

■

The previous theorem leads to yet another characterization of the conditional expectation. To this end take $X \in \mathcal{L}^2\left(\mathbb{P}\right)$ and denote $X_0 := \mathbb{E}\left[X \mid \mathcal{C}\right]$ for a $\mathcal{C} \subseteq \mathcal{F}$. Let $Z \in \mathcal{L}^2(\mathbb{P})$ be $\mathcal{C}$-measurable. Then $X_0 \in \mathcal{L}^2\left(\mathbb{P}\right)$ and by (11.3)

$$\mathbb{E}\left[Z \cdot (X - X_0) \mid \mathcal{C}\right] = Z\mathbb{E}\left[X - X_0 \mid \mathcal{C}\right] = Z \cdot \left(\mathbb{E}\left[X \mid \mathcal{C}\right] - X_0\right) = Z \cdot (X_0 - X_0) = 0.$$

Theorem 11.13 *For all* $X \in \mathcal{L}^2\left(\mathbb{P}\right)$ *and each* $\mathcal{C} \subseteq \mathcal{F}$ *the conditional expectation* $\mathbb{E}\left[X \mid \mathcal{C}\right]$ *is (up to a.s. equality) the unique* $\mathcal{C}$*-measurable random variable* $X_0 \in \mathcal{L}^2\left(\mathbb{P}\right)$ *with*

$$\mathbb{E}\left[\left(X - X_0\right)^2\right] = \min \ \left\{\mathbb{E}\left[\left(X - Y\right)^2\right]; Y \in \mathcal{L}^2\left(\mathbb{P}\right), Y \ \mathcal{C}\text{-measurable}\right\}$$

Proof. Let $Y \in \mathcal{L}^2(\mathbb{P})$ be $\mathcal{C}$-measurable. Put $X_0 := \mathbb{E}[X \mid \mathcal{C}]$. Then

$$\mathbb{E}((X-Y)^2) = \mathbb{E}((X-X_0+X_0-Y)^2) = \mathbb{E}((X-X_0)^2)+\mathbb{E}((X_0-Y)^2)+2\mathbb{E}((X-X_0)(X_0-Y))$$

But $\mathbb{E}((X - X_0)(X_0 - Y)) = 0$, since $X_0 - Y$ is $\mathcal{C}$-measurable.
This gives

$$\mathbb{E}\left[(X-Y)^2\right] - \mathbb{E}\left[(X-X_0)^2\right] = \mathbb{E}\left[(X_0-Y)^2\right]. \tag{11.4}$$

Due to positivity of squares we hence obtain

$$\mathbb{E}\left[(X-X_0)^2\right] \leq \mathbb{E}\left[(X-Y)^2\right].$$

If, on the other hand

$$\mathbb{E}\left[(X-X_0)^2\right] = \mathbb{E}\left[(X-Y)^2\right]$$

then

$$\mathbb{E}\left[(X_0-Y)^2\right] = 0$$

which implies $Y = X_0 = \mathbb{E}[X \mid \mathcal{C}]$ $\mathbb{P}$-a.s. ■

The last theorem states that $\mathbb{E}[X \mid \mathcal{C}]$ for $X \in \mathcal{L}^2(\mathbb{P})$ is the "best approximation" of X in the $\mathcal{C}$-measurable function space "in the sense of a least squares approximation". It is the projection of X onto the space of square integrable, $\mathcal{C}$-measurable functions.

Exercise 11.14 *Prove that for $X \in \mathcal{L}^2(\mathbb{P})$, $\mu = \mathbb{E}(X)$ is the number that minimizes $\mathbb{E}((X-\mu)^2)$.*

With the help of conditional expectation we can also give a new definition of conditional probability

Definition 11.15 *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $\mathcal{C} \subset \mathcal{F}$ be a sub-σ-algebra. For $A \in \mathcal{F}$*

$$\mathbb{P}[A \mid \mathcal{C}] := \mathbb{E}[1_A \mid \mathcal{C}]$$

is called the conditional probability of A given $\mathcal{C}$.

Example 11.16 *In the situation of Example 11.5.2 the conditional expectation of $A \in \mathcal{F}$ is given by*

$$\mathbb{P}(A \mid \mathcal{C}) = \sum_{i \in I} \mathbb{P}(A \mid B_i)\, 1_{B_i} := \sum_{i \in I} \frac{\mathbb{P}(A \cap B_i)\, 1_{B_i}}{\mathbb{P}(B_i)}.$$

In a last step we will only introduce (but not prove) conditional expectations on events with zero probability. Of course, in general this will just give nonsense but in the case of a conditional expectation $\mathbb{E}[X \mid Y = y]$ where X, Y are random variables such that (X, Y) has a Lebesgue density we can give this expression a meaning.

Theorem 11.17 *Let X, Y be real valued random variables such that (X, Y) has a density $f : \mathbb{R}^2 \to \mathbb{R}^+_{\{0\}}$ with respect to two dimensional Lebesgue measure λ^2. Assume that X is integrable and that*

$$f_0(y) := \int f(x, y)\, dx > 0 \quad \text{for all } y \in \mathbb{R}.$$

Then the function $\mathbb{E}(X|Y)$ will be denoted by

$$y \mapsto \mathbb{E}(X \mid Y = y)$$

and one has

$$\mathbb{E}(X \mid Y = y) = \frac{1}{f_0(y)} \int x f(x, y)\, dx \qquad \text{for } \mathbb{P}_Y\text{-a.e.} \quad y \in \mathbb{R}.$$

In particular

$$\mathbb{E}(X \mid Y) = \frac{1}{f_0(Y)} \int x f(x, Y)\, dx \qquad \mathbb{P}\text{-a.s.}$$

We will also need the following relationship between conditional expectation and independence, which is a generalization of Example 11.5, case 1.

Lemma 11.18 *Let X be an integrable real valued random variable and $\mathcal{C} \subset \mathcal{F}$ a sub-σ-algebra such that X is independent of $\mathcal{C}$, that is $\sigma(X)$ and $\mathcal{C}$ are independent, then*

$$\mathbb{E}(X \mid \mathcal{C}) = \mathbb{E}(X), \quad \mathbb{P}\text{-a.s.}$$

Proof. Suppose $X \geq 0$. Then an increasing sequence of step functions X_n can be constructed by $X_n = [2^n X]/2^n$. Then X_n converges monotonically to X. Notice that X_n is a linear combination of indicator functions 1_A with $A \in \sigma(X)$. And $\int_C 1_A d\mathbb{P} = \mathbb{P}(C \cap A) = \mathbb{P}(C)\mathbb{P}(A) = \int_C \mathbb{E}(1_A) d\mathbb{P}$. Thus $\mathbb{E}(1_A \mid \mathcal{C}) = \mathbb{E}(1_A)$, and by linearity $\mathbb{E}(X_n \mid \mathcal{C}) = \mathbb{E}(X_n)$ and by the monotone convergence theorem $\mathbb{E}(X \mid \mathcal{C}) = \mathbb{E}(X)$. The general case follows by linearity, $X = X^+ - X^-$. ∎

Exercise 11.19 *Let X and Y be as in Theorem 11.17, such that X and Y are independent. Then X is independent of $\sigma(Y)$, and by Lemma 11.18 we have $\mathbb{E}(X \mid Y) = \mathbb{E}(X \mid \sigma(Y)) = \mathbb{E}(X)$. Apply Theorem 11.17 to give an alternative derivation of this fact.*

12 Martingales

In this section we are going to define a notion, that will turn out to be of central interest in all of so called stochastic analysis and mathematical finance. A key role in its definition will be taken by conditional expectation. In this section we will just give the definition and a couple of examples. There is a rich theory of martingales. Parts of this theory we will meet in a class on Stochastic Calculus.

Definition 12.1 *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and I be an ordered set (linearly ordered), i.e. for $s, t \in I$ either $s \leq t$ or $t \leq s$, with $s \leq t$ and $t \leq s$ implies $s = t$ and $s \leq t$, $t \leq u$ implies $s \leq u$. For $t \in I$ let $\mathcal{F}_t \subset \mathcal{F}$ be a σ-algebra. $(\mathcal{F}_t)_{t\in I}$ is called a filtration, if $s \leq t$ implies $\mathcal{F}_s \subset \mathcal{F}_t$. A sequence of random variables $(X_t)_{t\in I}$ is called $(\mathcal{F}_t)_{t\in I}$ - adapted if X_t is $\mathcal{F}_t$-measurable for all $t \in I$.*

Exercise 12.2 *Construct a filtration on a probability space with $|I| \geq 3$.*

Example 12.3 *Let $(X_t)_{t\in I}$ be a family of random variables, and I a linearly ordered set, then*

$$\mathcal{F}_t = \sigma\left\{X_s, s \leq t\right\}$$

is a filtration and (X_t) is adapted with respect to $(\mathcal{F}_t)$. $(F_t)_{t\in I}$ is called the canonical filtration with respect to $(X_t)_{t\in I}$.

Definition 12.4 *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and I a linearly ordered set. let $(\mathcal{F}_t)_{t\in I}$ be a filtration and $(X_t)_{t\in I}$ be an $(\mathcal{F}_t)$-adapted sequence of random variables. (X_t) is called an (F_t)-**supermartingale**, if*

$$\mathbb{E}\left[X_t \mid \mathcal{F}_s\right] \leq X_s \quad \mathbb{P}\text{-a.s.} \tag{12.1}$$

for all $s \leq t$. (12.1) is equivalent with

$$\int_C X_t d\mathbb{P} \leq \int_C X_s d\mathbb{P}, \quad \text{for all } C \in \mathcal{F}_s. \tag{12.2}$$

*(X_t) is called a $(\mathcal{F}_t)$-**submartingale**, if $(-X_t)$ is a (F_t)-supermartingale. Eventually (X_t) is called a martingale, if it is both a submartingale and a supermartingale. This means that*

$$\mathbb{E}\left[X_t \mid \mathcal{F}_s\right] = X_s \quad \mathbb{P}\text{-a.s.}$$

for $s \leq t$ or, equivalently,

$$\int_C X_t d\mathbb{P} = \int_C X_s d\mathbb{P}, \quad C \in \mathcal{F}_s.$$

Exercise 12.5 *Show that the conditions (12.1) and (12.2) are equivalent.*

Remark 12.6 *1. If $(\mathcal{F}_t)$ is the canonical filtration with respect to $(X_t)_{t\in I}$, then often (X_t) simply called a supermartingale, submartingale, or a martingale.*

2. *(12.1) and (12.2) are evidently correct for $s = t$ (with equality). Hence these properties only need to be checked for $s < t$.*

3. *Putting $C = \Omega$ in (12.2) we obtain for a supermartingale $(X_t)_t$*

$$s \leq t \Rightarrow \mathbb{E}(X_s) \geq \mathbb{E}(X_t).$$

Hence for supermartingales $(\mathbb{E}(X_s))_s$ is a decreasing sequence, while for a submartingale $(\mathbb{E}(X_s))$ is an increasing sequence.

4. *In particular, if each of the random variables X_s is almost surely constant, e.g. if Ω is a singleton (a set with just one element) then (X_s) is a decreasing sequence, if (X_s) is a supermartingale. And it is an increasing sequence, if (X_s) is a submartingale. Hence martingales are (in a certain sense) the stochastic generalization of constant sequences.*

Exercise 12.7 *Let (X_t), (Y_t) be adapted to the same filtration and $\alpha, \beta \in \mathbb{R}$. Show the following*

1. *If (X_t) and (Y_t) are martingales, then $(\alpha X_t + \beta Y_t)$ is a martingale.*

2. *If (X_t) and (Y_t) are supermartingales, then so is $(X_t \wedge Y_t) = (\min(X_t, Y_t))$*

3. *If $(X_t))$ is a submartingale, so is $\left(X_t^+, \mathcal{F}_t\right)$.*

4. *If (X_t) is a martingale taking values in an open set $J \leq \mathbb{R}$ and*

$$q : J \to \mathbb{R}$$

is convex then $(q \circ X_t, \mathcal{F}_t)$ is a submartingale, if $q(X_t)$ is integrable for all t.

Of course, at first glance the definition of a martingale may look a bit weird. We will therefore give a couple of examples to show that it is not as strange as expected.

Example 12.8 *Let (X_n) be an i.i.d. sequence of $\mathbb{R}$-valued random variables. Put $S_n = X_1 + \ldots + X_n$ and consider the canonical filtration $\mathcal{F}_n = \sigma(S_m, m \leq n)$. By Lemma 11.18 we have*

$$\mathbb{E}[X_{n+1} \mid S_1, \ldots, S_n] = \mathbb{E}[X_{n+1}] \quad \mathbb{P}\text{-a.s.}$$

and by part 2. of Exercise 11.7

$$\mathbb{E}[X_i \mid S_1, \ldots, S_n] = X_i \quad \mathbb{P}\text{-a.s.}$$

for all $i = 1, \ldots, n$. Adding these $n + 1$ equations gives

$$\mathbb{E}[S_{n+1} \mid \mathcal{F}_n] = S_n + \mathbb{E}[X_{n+1}] \quad \mathbb{P}\text{-a.s.}$$

If $\mathbb{E}X_i = 0$ for all i, then

$$\mathbb{E}[S_{n+1} \mid \mathcal{F}_n] = S_n$$

i.e. (S_n) is a martingale. If $\mathbb{E}[X_i] \leq 0$ then

$$\mathbb{E}[S_{n+1} \mid \mathcal{F}_n] \leq S_n,$$

i.e. (S_n) is a supermartingale. In the same way (S_n) is a submartingale if $\mathbb{E}X_i \geq 0$.

Example 12.9 *Consider the following game. For each $n \in \mathbb{N}$ a coin with probability p for heads is tossed. If it shows heads ($X_n = +1$) our player receives money otherwise he ($X_n = -1$) looses money. The way he wins or looses is determined in the following way. Before the game starts he determines a sequence $(\varrho_n)_n$ of functions*

$$\varrho_n : \{H, T\}^n \to \mathbb{R}^+.$$

In round number $n+1$ he plays for $\varrho_n(X_1, \ldots, X_n)$ Euros depending on how the first n games ended. If we denote by S_n his capital at time n, then

$$S_1 = X_1 \quad \text{and} \quad S_{n+1} = S_n + \varrho_n(X_1, \ldots, X_n) X_{n+1}.$$

Hence

$$\begin{aligned} \mathbb{E}[S_{n+1} \mid X_1, \ldots, X_n] &= S_n + \varrho_n(X_1, \ldots, X_n) \cdot \mathbb{E}[X_{n+1} \mid X_1, \ldots, X_n] \\ &= S_n + \varrho_n(X_1, \ldots, X_n)\, \mathbb{E}(X_{n+1}) \\ &= S_n + (2p-1)\, \varrho_n(X_1, \ldots, X_n), \end{aligned}$$

since X_{n+1} is independent of $X_1, \ldots, X_n$ and $\mathbb{E}(X_{n+1}) = 2p - 1$. Hence for $p = \frac{1}{2}$

$$\mathbb{E}[S_{n+1} \mid X_1, \ldots, X_n] = S_n$$

so (S_n) is a martingale while for $p > \frac{1}{2}$

$$\mathbb{E}[S_{n+1} \mid X_1, \ldots, X_n] \geq S_n,$$

hence (S_n) is a submartingale and for $p < \frac{1}{2}$

$$\mathbb{E}[S_{n+1} \mid X_1, \ldots X_n] \leq S_n,$$

so (S_n) is a supermartingale. This explains the idea that "martingales are generalizations of fair games".

Exercise 12.10 *Let $X_1, X_2, \ldots$ be a sequence of independent random variables with finite variance $\mathbb{V}(X_i) = \sigma_i^2$. Then $\{\sum_{i=1}^n (X_i - \mathbb{E}(X_i))\}^2 - \sum_{i=1}^n \sigma_i^2$ is a martingale with respect to the filtration $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$.*

Exercise 12.11 *Consider the gambler's martingale. Consider an i.i.d. sequence $(X_n)_{n=1}^\infty$ of Bernoulli variables with values –1 and 1, each with probability 1/2. Consider the sequence (Y_n) such that $Y_n = 2^{n-1}$ if $X_1 = \cdots = X_{n-1} = -1$, and $Y_n = 0$ if $X_i = 1$ for some $i \leq n-1$. Show that $S_n = \sum_{i=1}^n X_i Y_i$ is a martingale. Show that S_n almost surely converges and determine its limit S_∞. Observe that $S_n \neq \mathbb{E}(S_\infty \mid \mathcal{F}_n)$.*

Example 12.12 *In a sense Example 12.8 is both, a special case and a generalization of the following example. To this end let $X_1, \ldots, X_n, \ldots$ denote an i.i.d. sequence of $\mathbb{R}^d$-valued random variables. Assume*

$$\mathbb{P}(X_i = +\varrho_k) = \mathbb{P}(X_i = -\varrho_k) = \frac{1}{2d}$$

for all $i = 1, 2, \ldots$ and all $k = 1, \ldots, d$. Here ϱ_k denotes the k-th unit vector. Define the stochastic process S_n by

$$S_0 = 0,$$

and

$$S_n = \sum_{i=1}^{n} X_i.$$

This process is called a random walk in d directions. Some of its properties will be discussed below. First we will see that indeed (S_n) is a martingale. Indeed,

$$\mathbb{E}\left[S_{n+1} \mid X_1, \ldots, X_n\right] = \mathbb{E}\left[X_{n+1}\right] + S_n = S_n.$$

*As a matter of fact, not only is (S_n) a martingale, but, in a certain sense it is **the** discrete time martingale.*

Since the random walk in d dimensions is **the** model for a discrete time martingale (the standard model of a continuous time martingale will be introduced in the following section) it is worth while studying some of its properties. This has been done in thousands of research papers in the past 50 years. We will just mention one interesting property here, that reveals a dichotomy in the random walk's behavior for dimensions $d = 1, 2$ or $d \geq 3$.

Definition 12.13 *Let (S_n) be a stochastic process in $\mathbb{Z}^d$, i.e. for each $n \in \mathbb{N}$, S_n is a random variable with values in $\mathbb{Z}^d$. (S_n) is called recurrent in a state $x \in \mathbb{Z}^d$, if*

$$\mathbb{P}\left(S_n = x \quad \textit{infinitely often in } n\right) = 1.$$

It is called transient in x, if

$$\mathbb{P}\left(S_n = x \quad \textit{infinitely often in } n\right) < 1.$$

(S_n) is called recurrent (transient), if each $x \in \mathbb{Z}^d$ is recurrent (transient).

Proposition 12.14 *In the situation of Example 12.12, if $x \in \mathbb{Z}^d$ is recurrent, then all $y \in \mathbb{Z}^d$ are recurrent.*

Exercise 12.15 *Show proposition 12.14.*

We will show a variant of the following

Theorem 12.16 *The random walk (S_n) introduced in Example 12.12 is recurrent in dimensions $d = 1, 2$ and transient for $d \geq 3$.*

To prove a version of Theorem 12.16 we will first discuss the property of recurrence:

Lemma 12.17 *Let f_k denote probability that the random walk return to the origin after k steps for the first time, and let p_k denote probability that the random walk return to the origin after k steps. Then a random walk (S_n) is recurrent if and only if $\sum_k f_k = 1$ and this is the case if and only if $\sum_k p_k = \infty$.*

Proof. The first equivalence is easy. Denote by Ω_k the set of all realizations of the random walk returning to the origin for the first time after k steps. Then if the random walk (S_n) is recurrent, with probability one there exists a $k > 0$ such that $S_k = 0$ and $S_l \neq 0$ for all $0 < l < k$. Hence $\sum_k f_k = 1$. On the other hand, if $\sum_k f_k = 1$, then $\mathbb{P}(\bigcup_k \Omega_k) = 1$. Hence with probability one there exists a $k > 0$ such that $S_k = 0$ and $S_l \neq 0$ for all $0 < l < k$. But then the situation at times 0 and k is completely the same and hence there exists $k' > k$ such that $S_{k'} = 0$ and $S_l \neq 0$ for all $k < l < k'$. Iterating this gives that $S_k = 0$ for infinitely many k's with probability one.

In order to relate f_k and p_k we derive the following recursion

$$p_k = f_k + f_{k-1}p_1 + \cdots + f_0 p_k \tag{12.3}$$

(the last summand ist just added for completeness, we have $f_0 = 0$). Indeed this is again easy to see. The left hand side is just the probability to be at the origin at time k. This event is the disjoint uinion of the events to be at 0 for the first time after $1 \leq l \leq k$ steps and to walk from zero to zero in the remaining steps. Hence we obtain.

$$p_k = \sum_{i=1}^{k} f_i p_{k-i} \quad \text{and } p_0 = 1. \tag{12.4}$$

Define the generating functions

$$F(z) = \sum_{k\geq 0} f_k z^k \quad \text{and } P(z) = \sum_{k\geq 0} p_k z^k.$$

Multiplying the left and right sides in (12.4) with z^k and summing them from $k = 0$ to infinity gives

$$P(z) = 1 + P(z)F(z)$$

i.e.

$$F(z) = 1 - 1/P(z).$$

By Abel's theorem

$$\sum_{k=1}^{\infty} f_k = F(1) = \lim_{z\uparrow 1} F(z) = 1 - \lim_{z\uparrow 1} \frac{1}{P(z)}.$$

First assume that $\sum_k p_k < \infty$. Then

$$\lim_{z\uparrow 1} P(z) = P(1) = \sum_k p_k < \infty$$

and thus

$$\lim_{z\uparrow 1} \frac{1}{P(z)} = 1/\sum_k p_k > 0.$$

Hence $\sum_{k=1}^{\infty} f_k < 1$ and the random walk (S_n) is transient.

Next assume that $\sum_k p_k = \infty$ and fix $\varepsilon > 0$. Then we find N such that

$$\sum_{k=0}^{N} p_k \geq \frac{2}{\varepsilon}.$$

Then for z sufficiently close to one we have $\sum_{k=0}^{N} p_k z^k \geq \frac{1}{\varepsilon}$ and consequently for such z

$$\frac{1}{P(z)} \leq \frac{1}{\sum_{k=0}^{N} p_k z^k} \leq \varepsilon.$$

But this implies that

$$\lim_{z \uparrow 1} \frac{1}{P(z)} = 1/\sum_k p_k = 0$$

and therefore $\sum_{k=1}^{\infty} f_k = 1$ and the random walk (S_n) is transient. ■

Exercise 12.18 *What has the Borel-Cantelli Lemma 8.1 to say in the above situation? Don't overlook that the events $\{S_n = x\}$ $(n \in \mathbb{N})$ may be dependent.*

We will now apply this criterion to analyze recurrence and transience for a random walk similar to the one defined in Example 12.12.

To this end define the following random walk (R_n) in d dimensions. For $k \in \mathbb{N}$ let $Y_1^k, \ldots, Y_d^k$ be i.i.d. random variables, taking values in $\{-1, +1\}$ with $\mathbb{P}(Y_1^k = 1) = \mathbb{P}(Y_1^k = -1) = 1/2$. Let X_k be the random vector $X_k = (Y_1^k, \ldots, Y_d^k)$. Define $R_0 \equiv 0$ and for $k \geq 1$

$$R_n = \sum_{k=1}^{n} X_k.$$

Theorem 12.19 *The random walk (R_n) defined above is recurrent in dimensions $d = 1, 2$ and transient for $d \geq 3$.*

Proof. Consider a sequence of i.i.d. random variables (Z_k) taking values in $\{-1, +1\}$ with $\mathbb{P}(Z_k = 1) = \mathbb{P}(Z_k = -1) = 1/2$. Write $q_k = \mathbb{P}(\sum_{i=1}^{2k} Z_i = 0)$. Then we apply Stirling's formula

$$\lim_{n \to \infty} n!/(\sqrt{2\pi} n^{n+1/2} e^{-n}) = 1.$$

to obtain

$$\begin{aligned} q_k &= \binom{2k}{k} 2^{-2k} = \frac{2k!}{k!k!} 2^{-2k} \\ &\sim \frac{\sqrt{4\pi k} \left(\frac{2k}{e}\right)^{2k}}{2\pi k \left(\frac{k}{e}\right)^{2k}} 2^{-2k} \\ &= \sqrt{\frac{1}{\pi k}}. \end{aligned}$$

Hence the probability of a single coordinate of R_{2n} to be zero (R_n cannot be zero if n is odd) asymptotically behaves like $\sqrt{\frac{1}{\pi n}}$. Hence

$$\mathbb{P}(R_n = 0) \sim \left(\frac{1}{\pi n}\right)^{\frac{d}{2}}.$$

But

$$\sum_n \left(\frac{1}{\pi n}\right)^{\frac{d}{2}} = \infty$$

for $d = 1$ and $d = 2$, while

$$\sum_n \left(\frac{1}{\pi n}\right)^{\frac{d}{2}} < \infty$$

for $d \geq 3$. This proves the theorem. ■

We will give two results about martingales. The first one is inspired by the fact that given a random variable M and a filtration $\mathcal{F}_t$ of $\mathcal{F}$, the family of conditional expectations, $\mathbb{E}(M \mid \mathcal{F}_t)$, yields a martingale. Can a martingale always be described in this way? That is the content of the Martingale Limit Theorem.

Theorem 12.20 *Suppose M_t is a martingale with respect to the filtration $\mathcal{F}_t$ and that the martingale is (uniformly) square integrable, that is $\limsup_t \mathbb{E}(M_t^2) < \infty$. Then there is a square integrable random variable M_∞ such that $M_t = \mathbb{E}(M_\infty \mid \mathcal{F}_t)$ a.s.. Moreover $\lim M_t = M_\infty$ in L^2 sense.*

Proof. The basic property that we will use, is that the space of square integrable random variables $L^2(\Omega, \mathcal{F}, \mathbb{P})$ with the L^2 inner product is a complete vectorspace. In other words, a Cauchy sequence converges. Recall that for any $t < s$ it holds that $M_t = \mathbb{E}(M_s \mid \mathcal{F}_t)$, and we have seen in Theorem 11.13 that then $M_s - M_t$ is perpendicular to M_t. In particular we have the Pythagoras formula

$$\mathbb{E}(M_s^2) = \mathbb{E}(M_t^2) + \mathbb{E}((M_s - M_t)^2).$$

This implies that $\mathbb{E}(M_s^2)$ is increasing in s, and therefore its limit exists and equals $\limsup \mathbb{E}(M_t^2)$ which is finite. Therefore given $\varepsilon > 0$ there is a u such that for $t > s > u$ we have $\mathbb{E}((M_s - M_t)^2) < \varepsilon$. That means that $\{M_s\}_s$ is a Cauchy sequence. Let M_∞ be its limit. In particular M_∞ is a random variable. Since orthogonal projection onto the suspace of $\mathcal{F}_t$ measurable functions is a continuous map, it holds that $\mathbb{E}(M_\infty \mid \mathcal{F}_t) = \lim \mathbb{E}(M_s \mid \mathcal{F}_t) = M_t$. ■

With some extra effort one may show that M_∞ is also the limit in the sense of almost sure convergence. The Martingale Limit Theorem is valid under more general circumstances, for example it is sufficient that only $\limsup_t \mathbb{E}(|M_t|) < \infty$, in which case M_∞ is the limit in L^1 sense (as well as almost surely).

An important concept for random processes is the concept of a stopping time.

Definition 12.21 *A stopping time is a random variable $\tau : \Omega \to I \cup \{\infty\}$, such that for all $t \in I$, $\{\omega\,;\, \tau(\omega) \leq t\}$ is $\mathcal{F}_t$ measurable. Here $I \cup \{\infty\}$ is ordered such that $t < \infty$ for all $t \in I$. Given a process M_t, the stopped process $M_{\tau \wedge t}$ is given by $M_{\tau \wedge t}(\omega) = M_s(\omega)$ where $s = \tau \wedge t = \min(\tau, t)$.*

Example 12.22 *Given $A \in \mathcal{F}_u$ a stopping time is constructed by $\tau = \infty \cdot 1_{A^c} + u \cdot 1_A$, that is, $\tau_A(\omega) = \infty$ if $\omega \notin A$, and $\tau(\omega) = u$ if $\omega \in A$.*

Exercise 12.23 *If $T : \Omega \to \mathbb{R}$ is constant, then T is a stopping time. If S and T are stopping times then $\max(S,T)$ and $\min(S,T)$ are stopping times.*

A very important property of martingales is the following Martingale Stopping Theorem.

Theorem 12.24 *Let $\{M_t\}_t$ be a martingale, and τ a stopping time. Then the stopped process $\{M_{\tau\wedge t}\}_t$ is a martingale.*

Proof. It is easy to see that an adapted process stopped at a stopping time is again an adapted process. We will give a proof of the martingale property for the simple stopping time τ_A given above. If $s > t \geq u$, let $B \in \mathcal{F}_t$, then

$$\begin{aligned}\int_B \mathbb{E}(M_{\tau\wedge s} \mid \mathcal{F}_t)d\mathbb{P} &= \int_B M_{\tau\wedge s}d\mathbb{P} = \int_{B\cap A} M_{\tau\wedge s} + \int_{B\cap A^c} M_{\tau\wedge s} \\ &= \int_{B\cap A} M_u + \int_{B\cap A^c} M_s = \int_{B\cap A} M_u + \int_{B\cap A^c} \mathbb{E}(M_s \mid \mathcal{F}_t) \\ &= \int_{B\cap A} M_u + \int_{B\cap A^c} M_t = \int_B M_{\tau\wedge t}.\end{aligned}$$

If $u \geq t$ and $s > t$, then one can apply Theorem 11.12

$$\mathbb{E}(M_{\tau\wedge s} \mid \mathcal{F}_t) = \mathbb{E}(\mathbb{E}(M_{\tau\wedge s} \mid \mathcal{F}_u) \mid \mathcal{F}_t) = \mathbb{E}(M_{u\wedge s} \mid \mathcal{F}_t) = M_t = M_{\tau\wedge t}.$$

■

Exercise 12.25 *Modify the proof of Theorem 12.24 to show that a stopped supermartingale is a supermartingale.*

Exercise 12.26 *Consider the roulette game. There are several possibilities for a bet, given by a number $p \in (0,1)$ such that with probability $36/37 \cdot p$ the return is p^{-1} times the stake and the return is zero with probability $1-36/37\cdot p$. The probabilities p are such that $p^{-1} \in \mathbb{N}$. Suppose you start with an initial fortune $X_0 \in \mathbb{N}$, and perform a sequence of bets until this fortune is reduced to zero. We are interested in the expected value of the total sum of stakes. To determine this consider the sequence of subsequent fortunes X_i, and consider the sequence of stakes Y_i, meaning that the stake in bet i is $Y_i = Y_i(X_1, \ldots, X_{i-1})$ $(Y_i \leq X_{i-1})$. In particular, if for this stake the probability p is chosen, either $X_i = X_{i-1} - Y_i + p^{-1} \cdot Y_i$ (with probability $36/37 \cdot p$) or $X_i = X_{i-1} - Y_i$ (with probability $1 - 36/37 \cdot p$). Show that $(X_i + 1/37 \cdot \sum_{j=1}^{i} Y_j)_i$ is a martingale with respect to the filtration $\mathcal{F}_i = \sigma(X_1, \ldots, X_i)$. The stopping time N is the first time i such that $X_i = 0$. Show that $\mathbb{E}(\sum_{j=1}^{N} Y_j) = 37 \cdot X_0$.*

13 Brownian motion

In this section we will construct *the* continuous time martingale, Brownian motion. Besides this, Brownian motion is also a building block of stochastic calculus and stochastic analysis.

In stochastic analysis one studies random functions of one variable and various kinds of integrals and derivatives thereof. The argument of these functions is usually interpreted as 'time', so the functions themselves can be thought of as the path of a random process.

Here, like in other areas of mathematics, going from the discrete to the continuous yields a pay-off in simplicity and smoothness, at the price of a formally more complicated analysis. Compare, to make an analogy, the integral $\int_0^n x^3 dx$ with the sum $\sum_{k=1}^n k^3$. The integral requires a more refined analysis for its definition and its properties, but once this has been done the integral is easier to calculate. Similarly, in stochastic analysis you will become acquainted with a convenient differential calculus as a reward for some hard work in analysis.

Stochastic analysis can be applied in a wide variety of situations. We sketch a few examples below.

1. Some differential equations become more realistic when we allow some randomness in their coefficients. Consider for example the following *growth equation*, used among other places in population biology:

$$\frac{d}{dt}S_t = (r + \text{“}N_t\text{”})S_t. \tag{13.1}$$

   Here, S_t is the size of the population at time t, r is the average growth rate of the population, and the “noise” N_t models random fluctuations in the growth rate.

2. At time $t = 0$ an investor buys stocks and bonds on the financial market, i.e., he divides his initial capital C_0 into A_0 shares of stock and B_0 shares of bonds. The bonds will yield a guaranteed interest rate r'. If we assume that the stock price S_t satisfies the growth equation (13.1), then his capital C_t at time t is

$$C_t = A_t S_t + B_t e^{r't}, \tag{13.2}$$

   where A_t and B_t are the amounts of stocks and bonds held at time t. With a keen eye on the market the investor sells stocks to buy bonds and vice versa. If his tradings are ‘self-financing’, then $dC_t = A_t dS_t + B_t d(e^{r't})$. An interesting question is:

   - What would he be prepared to pay for a so-called *European call option*, i.e., the right (bought at time 0) to purchase at time $T > 0$ a share of stock at a predetermined price K?

   The rational answer, q say, was found by Black and Scholes (1973) through an analysis of the possible strategies leading from an initial investment q to a payoff C_T. Their formula is being used on the stock markets all over the world.

3. The *Langevin equation* describes the behaviour of a dust particle suspended in a fluid:

$$m\frac{d}{dt}V_t = -\eta V_t + \text{“}N_t\text{”}. \tag{13.3}$$

   Here, V_t is the velocity at time t of the dust particle, the friction exerted on the particle due to the viscosity η of the fluid is $-\eta V_t$, and the “noise” N_t stands for the disturbance due to the thermal motion of the surrounding fluid molecules colliding with the particle.

4. The path of the dust particle in example 3 is observed with some inaccuracy. One measures the perturbed signal $Z(t)$ given by

$$Z_t = V_t + \text{“}\tilde{N}_t\text{”}. \tag{13.4}$$

Here $\tilde{N}_t$ is again a “noise”. One is interested in the best guess for the actual value of V_t, given the observation Z_s for $0 \leq s \leq t$. This is called a *filtering problem*: how to filter away the noise $\tilde{N}_t$. Kalman and Bucy (1961) found a linear algorithm, which was almost immediately applied in aerospace engineering. Filtering theory is now a flourishing and extremely useful discipline.

5. Stochastic analysis can help solve boundary value problems such as the Dirichlet problem. If the value of a harmonic function f on the boundary of some bounded regular region $D \subset \mathbb{R}^n$ is known, then one can express the value of f in the interior of D as follows:

$$\mathbb{E}\left(f\left(B_\tau^x\right)\right) = f(x), \tag{13.5}$$

where $B_t^x := x + \int_0^t N_t dt$ is an “integrated noise” or *Brownian motion*, starting at x, and τ denotes the time when this Brownian motion first reaches the boundary. (A harmonic function f is a function satisfying $\Delta f = 0$ with Δ the Laplacian.)

The goal of the course Stochastic Analysis is to make sense of the above equations, and to work with them.

In all the above examples the unexplained symbol N_t occurs, which is to be thought of as a “completely random” function of t, in other words, the continuous time analogue of a sequence of independent identically distributed random variables. In a first attempt to catch this concept, let us formulate the following requirements:

1. N_t is independent of N_s for $t \neq s$;
2. The random variables N_t $(t \geq 0)$ all have the same probability distribution μ;
3. $\mathbb{E}\left(N_t\right) = 0$.

However, when taken literally these requirements do not produce what we want. This is seen by the following argument. By requirement 1 we have for every point in time an independent value of N_t. We shall show that such a “continuous i.i.d. sequence” N_t is not measurable in t, unless it is identically 0.

Let μ denote the probability distribution of N_t, which by requirement 2 does not depend on t, i.e., $\mu([a,b]) := \mathbb{P}[a \leq N_t \leq b]$. Divide $\mathbb{R}$ into two half lines, one extending from a to $-\infty$ and the other extending from a to ∞. If N_t is not a constant function of t, then there must be a value of a such that each of the half lines has positive measure. So

$$p := \mathbb{P}(N_t \leq a) = \mu\left((-\infty, a]\right) \in (0,1). \tag{13.6}$$

Now consider the set of time points where the noise N_t is low: $E := \{\, t \geq 0 \colon N_t \leq a \,\}$. It can be shown that with probability 1 the set E is not Lebesgue measurable. Without

giving a full proof we can understand this as follows. Let λ denote the Lebesgue measure on $\mathbb{R}$. If E were measurable, then by requirement 1 and Eq. (13.6) it would be reasonable to expect its relative share in any interval (c, d) to be p, i.e.,

$$\lambda\left(E \cap (c, d)\right) = p\,(d - c)\,. \tag{13.7}$$

On the other hand, it is known from measure theory that every measurable set E is arbitrarily thick somewhere with respect to the Lebesgue measure λ, i.e., for all $\alpha < 1$ an interval (c, d) can be found such that

$$\lambda\left(E \cap (c, d)\right) > \alpha(d - c)$$

(cf. Halmos (1974) Th. III.16.A). This clearly contradicts Eq. (13.7), so E is not measurable. This is a bad property of N_t: for, in view of (13.1), (13.3), (13.4) and (13.5), we would like to integrate N_t.

For this reason, let us approach the problem from another angle. Instead of N_t, let us consider the *integral* of N_t, and give it a name:

$$B_t := \int_0^t N_s ds.$$

The three requirements on the evasive object N_t then translate into three quite sensible requirements for B_t.

BM1. For $0 = t_0 \leq t_1 \leq \cdots \leq t_n$ the random variables $B_{t_{j+1}} - B_{t_j}$ $(j = 0, \ldots, n-1)$ are independent;

BM2. B_t has stationary increments, i.e., the joint probability distribution of

$$(B_{t_1+s} - B_{u_1+s}, B_{t_2+s} - B_{u_2+s}, \ldots, B_{t_n+s} - B_{u_n+s})$$

does not depend on $s \geq 0$, where $t_i > u_i$ for $i = 1, 2, \cdots, n$ are arbitrary.

BM3. $\mathbb{E}\,(B_t - B_0) = 0$ for all t.

We add a normalisation:

BM4. $B_0 = 0$ and $\mathbb{E}(B_1^2) = 1$.

Still, these four requirements do not determine B_t. For example, the compensated Poisson jump process also satisfies them. Our fifth requirement fixes the process B_t uniquely:

BM5. $t \mapsto B_t$ continuous a.s.

The object B_t so defined is called the *Wiener process*, or (by a slight abuse of physical terminology) *Brownian motion*. In the next section we shall give a rigorous and explicit construction of this process.

Before we go into details we remark the following

Exercise 13.1 *Show that* ***BM5****, together with* ***BM1*** *and* ***BM2****, implies the following: For any* $\varepsilon > 0$

$$nP(|B_{t+\frac{1}{n}} - B_t| > \varepsilon) \to 0 \tag{13.8}$$

as $n \to \infty$*. Hint: compare with inequality (8.6).*

Exercise 13.1 helps us to specify the increments of Brownian motion in the following way[2].

Exercise 13.2 *Suppose* ***BM1****,* ***BM2****,* ***BM4*** *and (13.8) hold. Apply the Central Limit Theorem (Lindeberg's condition, page 36) to*

$$X_{n,k} := B_{\frac{kt}{n}} - B_{\frac{(k-1)t}{n}}$$

and conclude that $B_{s+t} - B_s$*,* $t > 0$ *has a normal distribution with variance* t*, i.e.*

$$P(B_{s+t} - B_s \in A) = \frac{1}{\sqrt{2\pi t}} \int_A e^{-\frac{x^2}{2t}} dx.$$

As a matter of fact, **BM1** and **BM5** already imply that the increments $B_{s+t} - B_s$ are normally distributed[3].

BM 2'. If $s \geq 0$ and $t > 0$, then

$$P(B_{s+t} - B_s \in A) = \frac{1}{\sqrt{2\pi t}} \int_A e^{-\frac{x^2}{2t}} dx.$$

we can now define Brownian motion as follows

Definition 13.3 *A one-dimensional Brownian motion is a real-valued process* $B_t, t \geq 0$ *with the properties* ***BM1****,* ***BM2'****, and* ***BM5****.*

13.1 Construction of Brownian Motion

Whenever a stochastic process with certain porperties is defined, the most natural question to ask is, does such a process exist? Of course, the answer is yes, otherwise these lecture notes would not have been written.

In this section we shall construct Brownian motion on $[0, T]$. For the sake of simplicity we will take $T = 1$, the construction for general T can be carried out along the same lines, or, by just concatenating independent Brownian motions.

The construction we shall use was given by P. Lévy in 1948. Since we saw that the increments of Brownian motion are independent Gaussian random variables, the idea is to construct Brownian motion from these Gaussian increments.

[2]See R. Durrett (1991), *Probability: Theory and Examples*, Section 7.1, Exercise 1.1, p. 334. Unfortunately, there is something wrong with this exercise. See the 3rd edition (2005) for a correct treatment.

[3]See e.g. I. Gihman, A. Skorohod, *The Theory of Stochastic Processes I*, Ch. III, § 5, Theorem 5, p. 189. For a high-tech approach, see N. Ikeda, S. Watanabe, *Stochastic Differential Equations and Diffusion Processes*, Ch. II, Theorem 6.1, p. 74.

More precisely, we start with the following observation. Suppose we already had constructed Brownian motion, say $(B_t)_{0\le t\le T}$. Take two times $0 \le s < t \le T$, put $\theta := \frac{s+t}{2}$, and let

$$p(\tau, x, y) := \frac{1}{\sqrt{2\pi\tau}} e^{-(y-x)^2/2\tau}, \quad \tau > 0, x, y, \in \mathbb{R}$$

be the Gaussian kernel centered in x with variance τ. Then, conditioned on $B_s = x$ and $B_t = z$, the random variable B_θ is normal with mean $\mu := \frac{x+z}{2}$ and variance $\sigma^2 := \frac{t-s}{4}$. Indeed, since B_s,$B_\theta - B_s$, and $B_t - B_\theta$ are independent we obtain

$$\begin{aligned} P[B_s \in dx, B_\theta \in dy, B_t \in dz] &= p(s,0,x)p(\frac{t-s}{2}, x, y)p(\frac{t-s}{2}, y, z)dx\,dy\,dz \\ &= p(s,0,x)p(t-s,x,z) \cdot \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} dx\,dy\,dz \end{aligned}$$

(which is just a bit of algebra). Dividing by

$$P[B_s \in dx, B_t \in dz] = p(s,0,x)p(t-s,x,z)dx\,dz$$

we obtain

$$P[B_\theta \in dy | B_s \in dx, B_t \in dz] = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}\, dy,$$

which is our claim.

This suggests that we might be able to construct Brownian motion on $[0,1]$ by interpolation.

To carry out this program, we begin with a sequence $\{\xi_k^{(n)}, k \in I(n), n \in \mathbb{N}_0\}$ of independent, standard normal random variables on some probability space $(\Omega, \mathcal{F}, P)$. Here

$$I(n) := \{k \in \mathbb{N}, k \le 2^n, k = 2l+1 \text{ for some } l \in \mathbb{N}\}$$

denotes the set of odd, positive integers less than 2^n. For each $n \in \mathbb{N}_0$ we define a process $B^{(n)} := \{B_t^{(n)} : 0 \le t \le 1\}$ by recursion and linear interpolation of the preceeding process, as follows. For $n \in \mathbb{N}$, $B^{(n)}_{k/2^{n-1}}$ will agree with $B^{(n-1)}_{k/2^{n-1}}$, for all $k = 0, 1, \ldots, 2^{n-1}$. Thus for each n we only need to specify the values of $B^{(n)}_{k/2^n}$ for $k \in I(n)$. We start with

$$B_0^{(0)} = 0 \text{ and } \quad B_1^{(1)} = \xi_1^{(0)}.$$

If the values of $B^{(n-1)}_{k/2^{n-1}}$, $k = 0, 1 \ldots 2^{n-1}$ have been defined (an thus $B_t^{(n-1)}$, $k/2^{n-1} \le t \le (k+1)/2^{n-1}$ is the linear interpolation between $B^{(n-1)}_{k/2^{n-1}}$ and $B^{(n-1)}_{(k+1)/2^{n-1}}$) and $k \in I(n)$, we denote $s = (k-1)/2^n$, $t = (k+1)/2^n$, $\mu = \frac{1}{2}(B_s^{(n-1)} + B_t^{(n-1)})$ and $\sigma^2 = \frac{t-s}{4} = 2^{-n+1}$ and set in accordance with the above observations

$$B^{(n)}_{k/2^n} := B^{(n)}_{(t+s)/2} := \mu + \sigma\xi_k^{(n)}.$$

We shall show that, almost surely, $B_t^{(n)}$ converges uniformly in t to a continuous function B_t (as $n \to \infty$) and that B_t is a Brownian motion.

We start with giving a more convenient representation of the processes $B^{(n)}, n = 0, 1, \ldots$. We define the following *Haar functions* by $H_1^0(t) \equiv 1$, and for $n \in \mathbb{N}, k \in I(n)$

$$H_k^{(n)}(t) := \begin{cases} 2^{(n-1)/2}, & \frac{k-1}{2^n} \le t < \frac{k}{2^n} \\ -2^{(n-1)/2}, & \frac{k}{2^n} \le t < \frac{k+1}{2^n} \\ 0 & \text{otherwise.} \end{cases}$$

The *Schauder functions* are defined by

$$S_k^{(n)}(t) := \int_0^t H_k^{(n)}(u)du, \quad 0 \le t \le 1,\, n \in \mathbb{N}_0,\, k \in I(n).$$

Note that $S_1^{(0)}(t) = t$, and that for $n \ge 1$ the graphs of $S_k^{(n)}$ are little tents of height $2^{-(n+1)/2}$ centered at $k/2^n$ and non overlapping for different values of $k \in I(n)$. Clearly, $B_t^{(0)} = \xi_1^{(0)} S_1^{(0)}(t)$, and by induction on n, it is readily verified that

$$B_t^{(n)}(\omega) = \sum_{m=0}^{n} \sum_{k \in I(m)} \xi_k^{(m)}(\omega) S_k^{(m)}(t), \quad 0 \le t \le 1,\, n \in \mathbb{N}. \tag{13.9}$$

Lemma 13.4 *As $n \to \infty$, the sequence of functions $\{B_t^{(n)}(\omega), 0 \le t \le 1\}, n \in N_0$, given by (13.9) converges uniformly in t to a continuous function $\{B_t(\omega), 0 \le t \le 1\}$ for almost every $\omega \in \Omega$.*

Proof. Let $b_n := \max_{k \in I(n)} |\xi_k^{(n)}|$. Oberserve that for $x > 0$ and each n, k

$$\begin{aligned} P(|\xi_k^{(n)}| > x) &= \sqrt{\frac{2}{\pi}} \int_x^\infty e^{-u^2/2} du \\ &\le \sqrt{\frac{2}{\pi}} \int_x^\infty \frac{u}{x} e^{-u^2/2} du = \sqrt{\frac{2}{\pi}} \frac{1}{x} e^{-x^2/2}, \end{aligned}$$

which gives

$$P(b_n > n) = P(\bigcup_{k \in I(n)} \{|\xi_k^{(n)}| > n\}) \le 2^n P(|\xi_1^{(n)}| > n) \le \sqrt{\frac{2}{\pi}} \frac{2^n}{n} e^{-n^2/2},$$

for all $n \in \mathbb{N}$. Since

$$\sum_n \sqrt{\frac{2}{\pi}} \frac{2^n}{n} e^{-n^2/2} < \infty,$$

the Borel-Cantelli Lemma implies that there is a set $\tilde{\Omega}$ with $P(\tilde{\Omega}) = 1$ such that for $\omega \in \tilde{\Omega}$ there is an $n_0(\omega)$ such that for all $n \ge n_0(\omega)$ it holds true that $b_n(\omega) \le n$. But then

$$\sum_{n \ge n_0(\omega)} \sum_{k \in I(n)} |\xi_k^{(n)}(\omega) S_k^{(n)}(t)| \le \sum_{n \ge n_0(\omega)} n 2^{-(n+1)/2} < \infty;$$

so for $\omega \in \tilde{\Omega}$, $B_t^{(n)}(\omega)$ converges uniformly in t to a limit B_t. The uniformity of the convergence implies the continuity of the limit B_t. ■

The following exercise facilitates the construction of Brownian motion substantially:

Exercise 13.5 *Check the following in a textbook of functional analysis:*

The inner product

$$\langle f, g\rangle := \int_0^1 f(t)g(t)dt$$

turns $L^2[0,1]$ *into a Hilbert space, and the Haar functions* $\{H_k^{(n)}; k \in I(n), n \in \mathbb{N}_0\}$ *form a complete, orthonormal system.*

Thus the Parseval equality

$$\langle f, g\rangle = \sum_{n=0}^{\infty} \sum_{k\in I(n)} \langle f, H_k^{(n)}\rangle\langle g, H_k^{(n)}\rangle \tag{13.10}$$

holds true.

Applying (13.10) to $f = 1_{[0,t]}$ and $g = 1_{[0,s]}$ yields

$$\sum_{n=0}^{\infty} \sum_{k\in I(n)} S_k^{(n)}(t)S_k^{(n)}(s) = s \wedge t. \tag{13.11}$$

Now we are able to prove

Theorem 13.6 *With the above notations*

$$B_t := \lim_{n\to\infty} B_t^{(n)}$$

is a Brownian motion in $[0,1]$.

Proof. In view of our definition of Brownian motion it suffices to prove that for $0 = t_0 < t_1 \ldots < t_n \leq 1$, the increments $(B_{t_j} - B_{t_{j-1}})_{j=1,\ldots,n}$ are independent, normally distributed with mean zero and variance $(t_j - t_{j-1})$. For this we will show that the Fourier transforms satisfy the appropriate condition, namely that for $\lambda_j \in \mathbb{R}$ (and as usual $i := \sqrt{-1}$)

$$\mathbb{E}\Big[\exp\big(i\sum_{j=1}^{n} \lambda_j(B_{t_j} - B_{t_{j-1}})\big)\Big] = \prod_{j=1}^{n} \exp\big(-\frac{1}{2}\lambda_j^2(t_j - t_{j-1})\big). \tag{13.12}$$

To derive (13.12) it is most natural to exploit the construction of B_t form Gaussian random variables. Set $\lambda_{n+1} = 0$ and use the independence and normality of the $\xi_k^{(n)}$ to compute for

$M \in \mathbb{N}$

$$\begin{aligned}
&\mathbb{E}\Big[\exp\big(-i\sum_{j=1}^{n}(\lambda_{j+1}-\lambda_j)B_{t_j}^{(M)}\big)\Big] \\
&= \mathbb{E}\Big[\exp\big(-i\sum_{m=0}^{M}\sum_{k\in I(m)}\xi_k^{(m)}\sum_{j=1}^{n}(\lambda_{j+1}-\lambda_j)S_k^{(m)}(t_j)\big)\Big] \\
&= \prod_{m=0}^{M}\prod_{k\in I(m)}\mathbb{E}\Big[\exp\big(-i\xi_k^{(m)}\sum_{j=1}^{n}(\lambda_{j+1}-\lambda_j)S_k^{(m)}(t_j)\big)\Big] \\
&= \prod_{m=0}^{M}\prod_{k\in I(m)}\exp\big(-\frac{1}{2}\big(\sum_{j=1}^{n}(\lambda_{j+1}-\lambda_j)S_k^{(m)}(t_j)\big)^2\big) \\
&= \exp\Big(-\frac{1}{2}\sum_{j=1}^{n}\sum_{l=1}^{n}(\lambda_{j+1}-\lambda_j)(\lambda_{l+1}-\lambda_l)\sum_{m=0}^{M}\sum_{k\in I(m)}S_k^{(m)}(t_j)S_k^{(m)}(t_l)\Big)
\end{aligned}$$

Now we send $M \to \infty$ and apply (13.11) to obtain

$$\begin{aligned}
&\mathbb{E}\Big[\exp\big(i\sum_{j=1}^{n}\lambda_j(B_{t_j}-B_{t_{j-1}})\big)\Big] = \mathbb{E}\Big[\exp\big(-i\sum_{j=1}^{n}(\lambda_{j+1}-\lambda_j)B_{t_j}\big)\Big] \\
&= \exp\Big(-\sum_{j=1}^{n-1}\sum_{l=j+1}^{n}(\lambda_{j+1}-\lambda_j)(\lambda_{l+1}-\lambda_l)t_j - \frac{1}{2}\sum_{j=1}^{n}(\lambda_{j+1}-\lambda_j)^2 t_j\Big) \\
&= \exp\Big(-\sum_{j=1}^{n-1}(\lambda_{j+1}-\lambda_j)(-\lambda_{j+1})t_j - \frac{1}{2}\sum_{j=1}^{n}(\lambda_{j+1}-\lambda_j)^2 t_j\Big) \\
&= \exp\Big(\frac{1}{2}\sum_{j=1}^{n-1}(\lambda_{j+1}^2-\lambda_j^2)t_j - \frac{1}{2}\lambda_n^2 t_n\Big) \\
&= \prod_{j=1}^{n}\exp\big(-\frac{1}{2}\lambda_j^2(t_j-t_{j-1})\big).
\end{aligned}$$

■

14 Appendix

Let Ω be a set and $\mathcal{P}(\Omega)$ the collection of subsets of Ω.

Definition 14.1 *A system of sets $\mathcal{R} \subset \mathcal{P}(\Omega)$ is called a ring if it satisfies*

$$\emptyset \in \mathcal{R}$$

$$A, B \in \mathcal{R} \Rightarrow A \setminus B \in \mathcal{R}$$

$$A, B \in \mathcal{R} \Rightarrow A \cup B \in \mathcal{R}$$

If additionally

$$\Omega \in \mathcal{R}$$

then $\mathcal{R}$ is called an algebra.

Note that for $A, B \subset \Omega$ their intersection $A \cap B = A \setminus (B \setminus A)$.

Definition 14.2 *A system $\mathcal{D} \subset \mathcal{P}(\Omega)$ is called a Dynkin system if it satisfies*

$$\Omega \in \mathcal{D}$$

$$D \in \mathcal{D} \Rightarrow D^c \in \mathcal{D}$$

For every sequence $(D_n)_{n\in\mathbb{N}}$ of pairwise disjoint sets $D_n \in \mathcal{D}$, their union $\cup_n D_n$ is also in $\mathcal{D}$.

The following theorem holds:

Theorem 14.3 *A Dynkin system is a σ-algebra if and only of for any two $A, B \in \mathcal{D}$ we have*

$$A \cap B \in \mathcal{D}$$

Similar to the case of σ-algebras for every system of sets $\mathcal{E} \subset \mathcal{P}(\Omega)$ there is a smallest Dynkin system $\mathcal{D}(\mathcal{E})$ generated by (and containing) $\mathcal{E}$. The importance of Dynkin systems mainly is due to the following

Theorem 14.4 *For every $\mathcal{E} \subset \mathcal{P}(\Omega)$ with*

$$A, B \in \mathcal{E} \Rightarrow A \cap B \in \mathcal{E}$$

we have

$$\mathcal{D}(\mathcal{E}) = \sigma(\mathcal{E}).$$

Definition 14.5 *Let $\mathcal{R}$ be a ring. A function*

$$\mu : \mathcal{R} \to [0, \infty]$$

is called a volume, if it satisfies

$$\mu(\emptyset) = 0$$

and

$$\mu(\cup_{i=1}^{n} A_i) = \sum_{i=1}^{n} \mu(A_i) \tag{14.1}$$

for all pairwise disjoint sets $A_1, \ldots, A_n \in \mathcal{R}$ *and all* $n \in \mathbb{N}$. *A volume* μ *is called a pre-measure if*

$$\mu(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i) \tag{14.2}$$

for all pairwise disjoint sequence of sets $(A_i)_{i\in\mathbb{N}} \in \mathcal{R}$. *We will call (14.1) finite additivity and (14.2)* σ*-additivity.*

A pre-measure μ on a σ-algebra $\mathcal{A}$ is called a measure.

Theorem 14.6 *Let* $\mathcal{R}$ *be a ring and* μ *be a volume on* $\mathcal{R}$. *If* μ *is a pre-measure, then it is* $\emptyset$*-continuous, i.e. for all* $(A_n)_n$, $A_n \in \mathcal{R}$, *with* $\mu(A_n) < \infty$ *and* $A_n \downarrow \emptyset$ *it holds* $\lim_{n\to\infty} \mu(A_n) = 0$. *If* $\mathcal{R}$ *is an algebra and* $\mu(\Omega) < \infty$, *then the reverse also holds: an* $\emptyset$*-continuous volume is a pre-measure.*

Theorem 14.7 (Carathéodory) *For every pre-measure* μ *on a ring* $\mathcal{R}$ *over* Ω *there is at least one way to extend* μ *to a measure on* $\sigma(\mathcal{R})$.

In the case that $\mathcal{R}$ is an algebra and μ is σ-finite (i.e. Ω is the countable union of subsets of finite measure), this extension is unique.