Probability I

Foundations and Limit Theorems

Jürgen Potthoff
Universität Mannheim

Die Wahrscheinlichkeitstheorie als mathematische Disziplin soll und kann genau in demselben Sinne axiomatisiert werden wie die Geometrie oder die Algebra.

A. KOLMOGOROFF
Grundbegriffe der Wahrscheinlichkeitsrechnung (1933)

... By the same token the record of a prolonged tossing of a coin is bound to contain every conceivable book in the Morse code, from Hamlet *to eightplace logarithmic tables. It has been suggested that an army of monkeys might be trained to pound typewriters at random in the hope that ultimately great works of literature would be produced. Using a coin for the same purpose may save feeding and training expenses and free monkeys for other monkey business.*

W. FELLER
An Introduction to Probability and Its Applications, Vol. 1 (1950)

Preface

The present lecture notes are based on my previous lecture notes *Wahrscheinlichkeits-theorie I* that I used over the last decade or so at Mannheim university for the lecture of the same name (recently renamed into *Wahrscheinlichkeitstheorie I — Grundlagen und Grenzwertsätze*). As explained in the preface to my notes *Probability II*, the advent of the new Bachelor and Master programs at our faculty suggested a revision of the notes *Wahrscheinlichkeitstheorie I*, and in particular a re-writing in English. The result is the present manuscript.

The principal aim of this lecture (and therefore of these lecture notes) is to give a general, rigorous and self-contained account of the two main theorems of probability: the *(strong) law of large numbers* and the *central limit theorem*. Therefore, on the technical side this means that we shall thoroughly analyze the various notions of convergence which appear in probability theory. In fact, the discussion of convergence and of characteristic functions, which are so powerful for the treatment of weak convergence and convergence in law, make up a central part of the lecture.

When I began teaching probability in Mannheim more than twenty years ago, I based my lectures on the marvelous books [4, 5] of H. Bauer, and they still belong to the sources in which I prefer to look something up. On the other hand, over the years, the pencil notes for my courses changed continuously, and especially for the present lecture notes the book [7] by P. Billingsley became an important influence for the treatment of weak convergence.

Apart from the re-writing of my previous lecture notes in English, and a different format, the most significant change is that the first chapter, which used to be a concise course ("Steilkurs") in measure theory and Lebesgue integration, has been moved to an appendix. The rationale for this course was that our curriculum in Mannheim did not — and still does not — leave space for a genuine course in measure and integration theory (although some of it is taught in the analysis course). Since modern measure and integration theory is the backbone of probability, the lack of a course in this subject is very disturbing for teaching probability. So for many years my solution was the above mentioned first chapter in my lecture *Wahrscheinlichkeitstheorie I*. However, this "Steilkurs", even though it was usually accepted by the students, made a rather boring beginning of the lecture. [1] Therefore I decided to leave this out in the

[1] It is not without reason that D. Williams [34] calls measure theory the "most arid of subjects when done for its own sake".

future, begin right away with the Kolmogorov setup, and bring in pieces of measure theory as they become needed. At the time of writing these notes, I am quite curious whether this will work out fine.

For their comments and for pointing out my errors I owe thanks to: Johannes Berger, Thomas Deck, Oliver Falkenburg, Karen Greive, Markus Huggenberger, Stephan Knapp, Lutz Küsters, Annika Lang, Silke Lorenz, Martin Schmidt, Edina Smajlović, and all the students who participated in my lectures *Wahrscheinlichkeitstheorie I* during the last two decades. I am especially grateful to Florian Werner for his extraordinary effort in proofreading this manuscript and for his numerous helpful suggestions.

Jürgen Potthoff
Mannheim, December 2013

Contents

Chapter 1

Introduction

In his highly influential address to the second *International Congress of Mathematicians* in Paris, 1900, *David Hilbert* proposed as the sixth of his famous 23 problems the axiomatization of probability — in his own words: [1]

> Durch die Untersuchungen über die Grundlagen der Geometrie wird uns die Aufgabe nahegelegt, *nach diesem Vorbilde diejenigen physikalischen Disciplinen axiomatisch zu behandeln, in denen schon heute die Mathematik eine hervorragende Rolle spielt; dies sind in erster Linie die Wahrscheinlichkeitsrechnung und die Mechanik.*

33 years later, A.N. Kolmogorov [2] published his book *Grundbegriffe der Wahrscheinlichkeitsrechnung* [22], which nowadays is considered as the solution to Hilbert's problem, as far as probability is concerned. However, between 1900 and 1933 there was a lot of activity on the foundations of probability, which finally culminated in Kolmogorov's book. For example, around 1900 *G. Bohlmann* had a system of axioms for probability (and probably had some influence on Hilbert concerning the sixth problem — actually, in [19] Hilbert quotes a contribution of Bohlmann to a mathematics textbook), in 1909 *E. Borel* had postulated the countable additivity of probabilities, and around 1923 *F. Hausdorff* had a system of axioms for probability. Other essential contributions to the development of probability theory in this period are due to *S. Bernstein*, *M. Fréchet*, *P. Lévy*, *R. v. Mises*, *E. Slutsky*, *H. Steinhaus*, and *N. Wiener*, among others.

The interested reader is referred to the article [33], which contains a large amount of very interesting information on the development of the foundations of probability between 1900 and 1933.

---

[1] Hilbert's talk was reprinted in [19]

[2] The attentive reader will have realized that I spelled Kolmogorov's name differently on the title page. The reason is, that in his famous book [22] the name is spelled as *A. Kolmogoroff*, which leads to believe that at this time Kolmogorov himself transcribed his Russian name in that way. However, later the spelling "Kolmogorov" became the most used one — for example as the author of the English translation [23] of [22]. Therefore I also shall use this spelling.

Kolmogorov does not mention Hilbert's address in Paris explicitly in [22], but the reading of his preface, and of the introductory sentences to the first chapter, where he mentions Hilbert's book on the foundations of geometry as a role model for the formulation of the foundations of a mathematical subject, strongly suggests, that it was his aim to provide the solution to probability part of the sixth problem.

In essence the axioms of probability, as we find them in Kolmogorov's book and as they are still used today, are the following: Events are subsets of a given set, their totality forms a σ-algebra, their probabilities are given by a countably additive set function with values in $[0, 1]$, and mean values of random variables (i.e., measurable functions on the basic set) are defined by their corresponding Lebesgue integrals. [3] So far, probability would just be a sub-domain of the general theory of measure and integration. It is the definition of *independence* of events and random variables which gives probability its own life, its own methods, its own ways of thinking, and which is essential for its central results. [4]

These lecture notes then begin with an account of Kolmogorov's axioms, and independence is studied in detail. The next chapter treats the notions of convergence which are used for the formulation of laws of large numbers: Almost sure convergence and convergence in probability. In chapter 4, various forms of the law of large numbers are proved. Chapter 5 is concerned with the analysis of convergence in law, and in particular Portmanteau's theorem is proved. In chapter 6 the analysis of convergence in law is continued, and its relation to characteristic functions (Lévy's continuity lemma) is established. Finally, the central limit theorem (in the general form given by Lindeberg and Lévy) is proved in chapter 7.

[3] This was already the point of view of Borel, cf. [33].

[4] According to [24], the notion of independence in its modern form can already be found in the works of Bohlmann.

Chapter 2

Kolmogorov's Axioms, Independence

Kolmogorov's axioms [22,23] are a basis for the formulation of mathematical models of (idealized) experiments which involve randomness or uncertainty. In these lectures, we shall state this foundation as three axioms. They are formulated in the "language" of the theory of measure and integration. The interested reader finds a summary of the basic aspects of the latter in appendix A.

2.1 Events, Probabilities, and Random Variables

2.1 Axiom *A random experiment is described by a* probability space $(\Omega, \mathcal{A}, P)$, *where Ω is a non-empty set, $\mathcal{A}$ a σ–algebra over Ω, and P a probability measure on the measurable space $(\Omega, \mathcal{A})$. The elements in $\mathcal{A}$ are called the* events *of the experiment, the value $P(A)$, $A \in \mathcal{A}$, is called the* probability *of the event A.*

Suppose that $(\Omega, \mathcal{A}, P)$ is a probability space, and that $(\Omega', \mathcal{A}')$ is a measurable space.

2.2 Definition An $\mathcal{A}$–$\mathcal{A}'$–measurable mapping X from Ω into Ω' is called a *random variable with values in Ω'*.

2.3 Remark It would be more precise (but also more clumsy) to say "a random variable with values in $(\Omega', \mathcal{A}')$", since the essential point is its measurability. However, there will be no danger of confusion. In a similar vein, in the case that $(\Omega', \mathcal{A}') = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ or $(\Omega', \mathcal{A}') = (\overline{\mathbb{R}}, \mathcal{B}(\overline{\mathbb{R}}))$ we shall just say that X is a "(real valued) random variable".

Suppose that X is a random variable with values in Ω', and let P_X denote the image of P under X:

$$P_X(A) = (P \circ X^{-1})(A) = P\bigl(X^{-1}(A)\bigr), \qquad A \in \mathcal{A}'. \tag{2.1}$$

2.4 Theorem *P_X is a probability measure on* $(\Omega' \mathcal{A}')$.

The *proof* is a straightforward *exercise*.

2.5 Definition The probability measure P_X defined in (2.1) is called the *law* or the *distribution* of the random variable X.

Let us emphasize that P_X is already determined on any $\cap$–stable generator of $\mathcal{A}'$, see theorem A.21.

Properties of a random variable X which can be formulated in terms of its law are called *probabilistic*. Therefore the probabilistic properties of a random variable involve much less information than the properties of X as a mapping from Ω into Ω'. Indeed, it is known from the introductory course [28], that in general different choices of $(\Omega, \mathcal{A}, P, X)$ can lead to the same law on the image space of X, and therefore to the same probabilistic model.

Let us consider for a while the case of *real valued* random variables X: That is, we choose $(\Omega', \mathcal{A}') = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$. There are two special cases which deserve to be mentioned. The first is that X has a *discrete law*:

$$P_X = \sum_{n=1}^{\infty} p_n \, \varepsilon_{x_n}, \tag{2.2}$$

where $(p_n, n \in \mathbb{N})$ is a sequence in $[0, 1]$ such that $\sum_n p_n = 1$, and $\{x_n, n \in \mathbb{N}\}$ is a countable subset of $\mathbb{R}$. Obviously, the interpretation of X is that it assumes the value x_n, $n \in \mathbb{N}$, with probability p_n, while values not in $\{x_n, n \in \mathbb{N}\}$ appear with probability zero, i.e., are almost surely impossible. The other case is that X has a *continuous law*:

$$P_X(B) = \int_B \varphi_X(x) \, d\lambda(x), \qquad B \in \mathcal{B}(\mathbb{R}), \tag{2.3}$$

where the right hand side denotes the Lebesgue integral of the *density φ_X of X* over B with respect to Lebesgue measure λ, and φ_X is a positive measurable function on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ so that $\int_{\mathbb{R}} \varphi_X \, d\lambda = 1$.

Notation In order to simplify the notation, we shall follow the common practice to denote the Lebesgue integral of a positive, measurable or λ–integrable function f on $\mathbb{R}$ also by

$$\int_{\mathbb{R}} f(x) \, dx.$$

And we shall do so similarly for functions defined on $\mathbb{R}^n$.

2.6 Examples

(a) Let $p \in [0, 1]$, $x_1, x_2 \in \mathbb{R}$, $x_1 \neq x_2$. A random variable with law given by

$$p\varepsilon_{x_1} + (1 - p)\varepsilon_{x_2} \tag{2.4}$$

is called a *Bernoulli random variable*. Most of the time one chooses $x_1 = 1$, $x_2 = 0$ or $x_2 = -1$, and interprets the value 1 as success, and 0 or -1 as failure. We use the notation $\mathcal{B}(p)$ for the Bernoulli law with parameters p, $x_1 = 1$, $x_2 = 0$. If nothing else is said, we shall consider these values for x_1, x_2 as the standard.

(b) Let $p \in [0, 1]$, $n \in \mathbb{N}$. A random variable with the discrete law given by

$$\sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} \varepsilon_k \tag{2.5}$$

is called a *binomial random variable* with parameters p, n. Notation for the law (2.5): $\mathcal{B}(p, n)$. The interpretation is the number of successes in n independent (see below) standard Bernoulli trials with parameter p.

(c) Let $\lambda > 0$. A random variable with law given by

$$e^{-\lambda} \sum_{n=0}^{\infty} \frac{\lambda^n}{n!} \varepsilon_n \tag{2.6}$$

is said to be a *Poisson random variable* with parameter λ. The Poisson law (2.6) is denoted by $\mathcal{P}(\lambda)$. The interpretation of a Poisson random variable with parameter λ is that of the number of random events per unit volume (space, area, time,...), where the average rate of events is given by λ.

(d) A continuously distributed random variable with density given by

$$\frac{1}{b-a} 1_{[a,b]}(x), \qquad x \in \mathbb{R}, \tag{2.7}$$

where $a < b$, is said to be *uniformly distributed* on the interval $[a, b]$. The notation for the law defined by the density (2.7) is $\mathcal{U}([a, b])$.

(e) Let $\lambda > 0$, and consider the density given by

$$\lambda e^{-\lambda x} 1_{[0,+\infty)}(x), \qquad x \in \mathbb{R}. \tag{2.8}$$

A random variable with this density is called *exponentially distributed* with parameter λ. The notation for the law defined by the density (2.8) is $\mathcal{E}(\lambda)$. It is an easy exercise to show that if $X \sim \mathcal{E}(\lambda)$ with $\lambda > 0$, then X has the following property for all $s, t \in \mathbb{R}_+$ with $s \le t$:

$$P(X \ge t \mid X \ge s) = P(X \ge t - s), \tag{2.9}$$

which is interpreted as X being *memoryless*. For this reason, exponentially distributed random variables are often used as models for the time that passes by until a certain random event happens.

(f) For $\mu \in \mathbb{R}$, and $\sigma^2 > 0$ consider the density given by

$$\frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-(x-\mu)^2/2\sigma^2}, \qquad x \in \mathbb{R}. \tag{2.10}$$

A random variable with this density is called *normally distributed* with *mean* μ, and *variance* σ^2. The notation for the law defined by (2.10) is $\mathcal{N}(\mu, \sigma^2)$. The interpretation of a random variable X which is distributed with the law $\mathcal{N}(\mu, \sigma^2)$ is that of an entity whose value fluctuates around the mean value μ due to a large number of small random influences. The order of magnitude of the fluctuations is given by the *standard deviation* $\sigma = \sqrt{\sigma^2}$. This interpretation is a direct consequence of the *central limit theorem* which we treat in chapter 7.

We continue to consider real valued random variables.

2.7 Definition Suppose that X is a real valued random variable with law P_X. The function F_X defined by

$$F_X(x) = P(X \le x) = P_X\big((-\infty, x]\big), \qquad x \in \mathbb{R}, \tag{2.11}$$

is called the *distribution function of* X.

2.8 Remark It is also possible (and common) to define the distribution function of $\mathbb{R}^n$–valued random variables in a similar manner, but we shall not need this for these lectures.

2.9 Theorem *Suppose that F is a real valued function defined on $\mathbb{R}$. F is the distribution function of a real valued random variable X, if and only if the following hold true:*

(a) *F is monotone increasing;*

(b) *F is continuous from the right;*

(c) $\lim_{x\to-\infty} F(x) = 0$, *and* $\lim_{x\to+\infty} F(x) = 1$.

In this case, the law P_X of X is uniquely determined by F.

Proof It is an easy *exercise* to show that the distribution function of a real-valued random variable has the properties (a)–(c). For the converse, we remark that properties (a) and (b) guarantee the existence of a Lebesgue–Stieltjes–measure μ_F on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ defined by F:

$$\mu_F((a, b]) = F(b) - F(a), \qquad a, b \in \mathbb{R},\ a < b,$$

cf. example A.26 in appendix A. The proof of uniqueness is left as an *exercise* based on theorem A.21. Property (c) shows that μ_F actually is a probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, because

$$\begin{aligned}\mu_F(\mathbb{R}) &= \lim_{n\to\infty} \mu_F\big((-n, n]\big)\\ &= \lim_{n\to\infty} \big(F(n) - F(-n)\big)\\ &= 1.\end{aligned}$$ □

In particular, all probabilistic properties of a real-valued random variable are also encoded in its distribution function.

2.2 Expectation Values

2.10 Axiom *Suppose that X is a positive, or a real or complex valued, integrable random variable on $(\Omega, \mathcal{A}, P)$. Then its Lebesgue integral*

$$\int_\Omega X\, dP \tag{2.12}$$

is called its (mathematical) expectation *or* mean value*, also denoted by $E(X)$.*

As in the formulation of axiom 2.10 we shall often have to assume that the Lebesgue integral of a random variable X with respect to the underlying probability measure P is well defined, but it is inconvenient each time to have to distinguish the case of positive and of integrable random variables. Therefore we shall use the following notations and conventions:

Notation For $p \geq 1$, the space of all (real or complex valued) p–fold integrable random variables X, that is, such that $|X|^p$ is integrable with respect to P, is — as usual — denoted by $\mathcal{L}^p(\Omega, \mathcal{A}, P)$. The cone of positive random variables on a probability space $(\Omega, \mathcal{A}, P)$ is denoted by $\mathcal{M}_+(\Omega, \mathcal{A})$. Set $\mathcal{L}(\Omega, \mathcal{A}, P) = \mathcal{M}_+(\Omega, \mathcal{A}) \cup \mathcal{L}^1(\Omega, \mathcal{A}, P)$. Whenever there is no danger of confusion, we shall also omit one or more of the qualifiers Ω, $\mathcal{A}$ or P in $\mathcal{L}(\Omega, \mathcal{A}, P)$. Therefore, we shall often simply write $X \in \mathcal{L}(P)$ in order to mean that X is positive or integrable. [1]

The transformation theorem A.53 (see appendix A) immediately gives

2.11 Theorem *Suppose that X is a real valued random variable with law P_X, and that f is a measurable, real or complex valued function on $\mathbb{R}$ such that $f \circ X$ belongs to $\mathcal{L}(P)$. Then*

$$E\big(f \circ X\big) = \int_{\mathbb{R}} f(x)\, dP_X(x) \tag{2.13}$$

holds true.

[1] Of course, strictly speaking this abbreviation is questionable, because for a positive random variable its Lebesgue integral is well defined regardless of the nature of P. However, this notation is convenient, and it will not lead to any confusion.

Of special importance are the following choices of f: $f(x) = x^n$, $n \in \mathbb{N}$, $f(x) = \exp(\lambda x)$, $f(x) = \exp(i\lambda x)$, $\lambda \in \mathbb{R}$, $x \in \mathbb{R}$:

2.12 Definition Suppose that X is a real valued random variable.

(a) If X^n, $n \in \mathbb{N}$, is P–integrable, then $E(X^n)$ is called the *n–th moment* of X. If $X \in \mathcal{L}^2(P)$, then $V(X) = E\big((X - E(X))^2\big)$ is called the *variance* of X, and $V(X)^{1/2}$ is called the *standard deviation* of X.

(b) The real valued function ψ_X defined by

$$\psi_X(\lambda) = E\big(\exp(\lambda X)\big), \qquad \lambda \in \mathbb{R}, \tag{2.14}$$

is called the *moment generating function* of X.

(c) The complex valued function φ_X defined by

$$\varphi_X(\lambda) = E\big(\exp(i\lambda X)\big), \qquad \lambda \in \mathbb{R}, \tag{2.15}$$

is called the *characteristic function* of X.

2.13 Exercise Assume that $X \in \mathcal{L}^2(P)$.

(a) Show that $V(X) = E(X^2) - E(X)^2$.

(b) Show that if $V(X) = 0$, then X is a.s. constant.

(c) Show that for every $a \in \mathbb{R}$, $V(X + a) = V(X)$.

A remark concerning the definition of the variance $V(X)$ is in order: If $X \in \mathcal{L}^p(P)$ for some $p \geq 1$, then Hölder's inequality (A.37), see theorem A.62, shows that X also belongs to $\mathcal{L}^q$ for any $q \geq 1$ with $q \leq p$. Therefore, if X is square integrable, then it is also integrable, and hence $E(X)$ is well defined. Thus under the assumption $X \in \mathcal{L}^2(P)$, $V(X)$ is indeed well defined.

The naming of the moment generating function ψ_X stems from the following fact. Suppose that we can interchange the n–th derivative with respect to λ of $\psi_X(\lambda)$ in a neighborhood of $\lambda = 0$ with the integral defining the expectation. (For example, this is true if $|X|^n \exp(\lambda X)$ is uniformly bounded in λ in a neighborhood of zero with a P–integrable bound.) Then we find

$$E(X^n) = \frac{d^n}{d\lambda^n}\, \psi_X(\lambda)\Big|_{\lambda=0}$$

so that the n–th moment of X can be computed from the moment generating function. Similarly, (under the assumption that it is finite) the n–th moment of X can be derived from the characteristic function, and we shall study this in detail in chapter 6.

Suppose that $X_1, X_2, \dots, X_n$, $n \in \mathbb{N}$, are real valued random variables defined on $(\Omega, \mathcal{A}, P)$. Then

$$\begin{aligned} X_1 \otimes X_2 \otimes \cdots \otimes X_n : \Omega &\to \mathbb{R}^n \\ \omega &\mapsto \big(X_1(\omega), \dots, X_n(\omega)\big) \end{aligned}$$

defines an $\mathbb{R}^n$–valued random variable, where $\mathbb{R}^n$ is equipped with the Borel σ–algebra $\mathcal{B}(\mathbb{R}^n)$. In order to see the measurability, one can for example use theorem A.34: $\mathcal{B}(\mathbb{R}^n)$ is generated by sets of the form $B = B_1 \times \cdots \times B_n$, with $B_i \in \mathcal{B}(\mathbb{R})$, $i = 1, \ldots, n$. But for such a set B,

$$\big(X_1 \otimes X_2 \otimes \cdots \otimes X_n\big)^{-1}(B) = \bigcap_{i=1}^{n} X_i^{-1}(B_i) \in \mathcal{A}.$$

Thus theorem A.34 implies the measurability of $X_1 \otimes X_2 \otimes \cdots \otimes X_n$. The probability measure $P_{X_1 \otimes X_2 \otimes \cdots \otimes X_n}$ on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$, defined as the image measure of P under this measurable mapping, is called the *joint law* or *joint distribution* of X_1, X_2, ..., X_n. By theorem A.21 the joint distribution is already uniquely specified by its values on a $\cap$–stable generator of $\mathcal{B}(\mathbb{R}^n)$. For example, one can choose sets like B of the form above, or Cartesian products of intervals of the form $(-\infty, x_i]$, $i = 1, \ldots, n$. Thus the joint law of $X_1, \ldots, X_n$ is completely determined by the numbers

$$P\big(X_1 \leq x_1, \ldots, X_n \leq x_n\big), \qquad x_1, \ldots, x_n \in \mathbb{R}.$$

Consider two real valued random variables $X, Y \in \mathcal{L}^2(P)$. Then X, Y, and XY belong to $\mathcal{L}^1(P)$ so that the following definition makes sense:

2.14 Definition Suppose that $X, Y \in \mathcal{L}^2(P)$. Then

$$\operatorname{Cov}(X, Y) = E\big((X - E(X))(Y - E(Y))\big) \tag{2.16}$$

is called the *covariance* of X and Y. If $\operatorname{Cov}(X, Y) = 0$, X and Y are called *uncorrelated*. If $V(X)$, $V(Y) > 0$, then

$$\rho(X, Y) = \frac{\operatorname{Cov}(X, Y)}{\sqrt{V(X)V(Y)}} \tag{2.17}$$

is called the *correlation* of X and Y.

2.15 Exercise Suppose that $X, Y \in \mathcal{L}^2(P)$.

(a) Show that $\operatorname{Cov}(X, Y) = E(XY) - E(X)E(Y)$.

(b) Show that for all $a, b \in \mathbb{R}$, $\operatorname{Cov}(X + a, Y + b) = \operatorname{Cov}(X, Y)$.

(c) Show that if $V(X)$, $V(Y) > 0$, then $|\rho(X, Y)| \leq 1$.

Consider $X_1, X_2, \ldots, X_n \in \mathcal{L}^2(P)$, $n \in \mathbb{N}$. It is an easy exercise to show that (it is of advantage to use exercises 2.13.(b), 2.15.(b))

$$V(X_1 + X_2 + \cdots + X_n) = \sum_{k=1}^{n} V(X_k) + \sum_{k,l=1,\, k \neq l}^{n} \operatorname{Cov}(X_k, X_l).$$

Therefore we immediately get

2.16 Theorem (Bienaymé) *If $X_1, X_2, \ldots, X_n \in \mathcal{L}^2(P)$, $n \in \mathbb{N}$, are pairwise uncorrelated, then*

$$V(X_1 + X_2 + \cdots + X_n) = \sum_{k=1}^{n} V(X_k). \tag{2.18}$$

2.3 Independence

In this section we consider a family $(\mathcal{E}_i,\ i \in I)$ of systems $\mathcal{E}_i \subset \mathcal{A}$, $i \in I$, of subsets of Ω, where I is a general non-empty index set.

2.17 Axiom *The family $(\mathcal{E}_i,\ i \in I)$ is called* independent*, if for every non-empty finite subset $\{i_1, i_2 \ldots, i_n\}$ of I, and every choice of $A_{i_j} \in \mathcal{E}_{i_j}$, $j = 1, \ldots, n$, the equality*

$$P\big(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_n}\big) = \prod_{j=1}^{n} P\big(A_{i_j}\big) \tag{2.19}$$

holds true.

2.18 Remark It is clear that if $(\mathcal{E}_i,\ i \in I)$ is independent, $J \subset I$, and for every $i \in J$, $\mathcal{E}'_i \subset \mathcal{E}_i$, then also $(\mathcal{E}'_i,\ i \in J)$ is independent.

2.19 Lemma *Suppose that $(\mathcal{E}_i,\ i \in I)$ is independent. Then also the family $\big(d(\mathcal{E}_i),\ i \in I\big)$ of d-systems generated by the systems $\mathcal{E}_i$, $i \in I$, (see definition A.13) is independent.*

Proof Since independence is only checked for finitely many indices, it is sufficient to consider the case $I = \{1, 2, \ldots, n\}$. Set

$$\mathcal{D} = \{A \in \mathcal{A},\ (\{A\}, \mathcal{E}_2, \ldots, \mathcal{E}_n) \text{ independent}\}.$$

We show that $\mathcal{D}$ is a d-system. For all arguments below pick a subset $\{i_1, \ldots, i_k\}$ of $\{2, \ldots, n\}$ and $A_{i_1} \in \mathcal{E}_{i_1}, \ldots, A_{i_k} \in \mathcal{E}_{i_k}$
(i) $\Omega \in \mathcal{D}$, because

$$\begin{aligned} P(\Omega \cap A_{i_1} \cap \cdots \cap A_{i_k}) &= P(A_{i_1} \cap \cdots \cap A_{i_k}) \\ &= \prod_{j=1}^{k} P(A_{i_j}) \\ &= P(\Omega) \prod_{j=1}^{k} P(A_{i_j}), \end{aligned}$$

where we used the independence of $(\mathcal{E}_2, \dots, \mathcal{E}_n)$.
(ii) Suppose that $A, B \in \mathcal{D}$ with $A \subset B$. Then $B \setminus A \in \mathcal{D}$:

$$\begin{aligned}
&P\big((B \setminus A) \cap A_{i_1} \cap \dots \cap A_{i_k}\big) \\
&\qquad = P\big((B \cap A_{i_1} \cap \dots \cap A_{i_k}) \setminus (A \cap A_{i_1} \cap \dots \cap A_{i_k})\big) \\
&\qquad = P\big((B \cap A_{i_1} \cap \dots \cap A_{i_k})\big) - P\big(A \cap A_{i_1} \cap \dots \cap A_{i_k})\big) \\
&\qquad = P(B) \prod_{j=1}^{k} P(A_{i_j}) - P(A) \prod_{j=1}^{k} P(A_{i_j}) \\
&\qquad = P(B \setminus A) \prod_{j=1}^{k} P(A_{i_j}).
\end{aligned}$$

(iii) If $(A_n, n \in \mathbb{N})$ is a sequence in $\mathcal{D}$ which increases to A, then $A \in \mathcal{D}$:

$$\begin{aligned}
P(A \cap A_{i_1} \cap \dots \cap A_{i_k}) &= P\big((\cup_n A_n) \cap A_{i_1} \cap \dots \cap A_{i_k}\big) \\
&= P\big(\cup_n (A_n \cap A_{i_1} \cap \dots \cap A_{i_k})\big) \\
&= \lim_n P(A_n \cap A_{i_1} \cap \dots \cap A_{i_k}) \\
&= \lim_n P(A_n) \prod_{j=1}^{k} P(A_{i_j}) \\
&= P\big(\cup_n A_n\big) \prod_{j=1}^{k} P(A_{i_j}) \\
&= P(A) \prod_{j=1}^{k} P(A_{i_j}).
\end{aligned}$$

In the third step we used the fact that $A_n \cap A_{i_1} \cap \dots \cap A_{i_k}$ is increasing in n, and in the fourth the hypothesis that A_n belongs to $\mathcal{D}$.

We have shown that $(\mathcal{D}, \mathcal{E}_2, \dots, \mathcal{E}_n)$ is independent. Next we observe that the d-system $\mathcal{D}$ contains $\mathcal{E}_1$, and therefore it also contains the smallest d-system $d(\mathcal{E}_1)$ containing $\mathcal{E}_1$. Thus we have proved that the family $\big(d(\mathcal{E}_1), \mathcal{E}_2, \dots, \mathcal{E}_n\big)$ is independent. Finally, we only need to successively repeat the argument for $\mathcal{E}_2, \dots, \mathcal{E}_n$ to conclude the proof. □

2.20 Corollary *Assume that $(\mathcal{E}_i, i \in I)$ is an independent family of $\cap$-stable systems $\mathcal{E}_i \subset \mathcal{A}$ of events, where I is an non-empty index set.*

(a) *The system $\big(\sigma(\mathcal{E}_i), i \in I\big)$ of σ–algebras generated by the $\mathcal{E}_i$, $i \in I$, is independent.*

(b) *If* $(I_j,\ j \in J)$ *is a decomposition of* I, *then the family* $\big(\sigma(\mathcal{E}_i,\ i \in I_j),\ j \in J\big)$ *is independent.*

Proof For the proof of (a) we only need to remark that because $\mathcal{E}_i$, $i \in I$, is $\cap$–stable, the d-system $d(\mathcal{E}_i)$ generated by $\mathcal{E}_i$, coincides with the σ–algebra $\sigma(\mathcal{E}_i)$ generated by $\mathcal{E}_i$, see theorem A.15.

For the proof of (b) define $\mathcal{F}_j$, $j \in J$, to be the system of all sets of the form $E_{i_1} \cap \cdots \cap E_{i_k}$ with $i_1, \ldots, i_k \in I_j$, and $E_{i_1} \in \mathcal{E}_{i_1}, \ldots, E_{i_k} \in \mathcal{E}_{i_k}$. $\mathcal{F}_j$, $j \in J$, is a generator of $\sigma(\mathcal{E}_i,\ i \in I_j)$: Since every σ–algebra is $\cap$–stable, we have for every $j \in J$ that

$$\sigma(\mathcal{E}_i,\ i \in I_j) \supset \mathcal{F}_j \supset \bigcup_{i \in I_j} \mathcal{E}_i,$$

so that we get

$$\sigma(\mathcal{E}_i,\ i \in I_j) \supset \sigma(\mathcal{F}_j) \supset \sigma\Big(\bigcup_{i \in I_j} \mathcal{E}_i\Big) = \sigma(\mathcal{E}_i,\ i \in I_j),$$

and therefore $\sigma(\mathcal{E}_i,\ i \in I_j) = \sigma(\mathcal{F}_j)$. By construction, for every $j \in J$, $\mathcal{F}_j$ is $\cap$–stable, and the family $(\mathcal{F}_j,\ j \in J)$ is independent, so that statement (b) follows now from (a). $\square$

2.21 Definition Suppose that $(\mathcal{E}_n,\ n \in \mathbb{N})$ is a sequence of sub–σ–algebras of $\mathcal{A}$. Define

$$\mathcal{T}_n = \sigma\big(\mathcal{E}_m,\ m \geq n\big),\ n \in \mathbb{N}, \quad \mathcal{T} = \bigcap_{n \in \mathbb{N}} \mathcal{T}_n, \tag{2.20}$$

and call $\mathcal{T}$ the *tail σ–algebra* of the sequence $(\mathcal{E}_n,\ n \in \mathbb{N})$.

2.22 Theorem (Kolmogorov's 0–1–Law) *If* $(\mathcal{E}_n,\ n \in \mathbb{N})$ *is an independent sequence of sub–σ–algebras of* $\mathcal{A}$, *then for every* A *in its tail σ–algebra* $\mathcal{T}$, $P(A) = 0$ *or* $P(A) = 1$.

Proof It is sufficient to show for every $A \in \mathcal{T}$ we have that A is independent of itself, because then we get $P(A) = P(A)^2$, and hence $P(A) = 0$ or $P(A) = 1$. Let $A \in \mathcal{T}$, and set

$$\mathcal{D} = \{B \in \mathcal{A},\ \ A \text{ and } B \text{ are independent}\}.$$

We are heading for the inclusion $\mathcal{D} \supset \mathcal{T}$, which indeed implies that A is independent of itself. To this end, we first show that $\mathcal{D}$ is a d-system. That $\Omega \in \mathcal{D}$ is quite obvious. If $B, C \in \mathcal{D}$ with $B \subset C$, then $A \cap (C \setminus B) = (A \cap C) \setminus (A \cap B)$, and $A \cap C \supset A \cap B$. Thus

$$\begin{aligned} P\big(A \cap (C \setminus B)\big) &= P(A \cap C) - P(A \cap B) \\ &= P(A)\big(P(C) - P(B)\big) \\ &= P(A)P(C \setminus B), \end{aligned}$$

so that we obtain that $C \setminus B \in \mathcal{D}$. Now assume that $(B_n,\ n \in \mathbb{N})$ is a sequence in $\mathcal{D}$ which increases to B. Then $A \cap B = \cup_n (A \cap B_n) = \lim_n (A \cap B_n)$, and therefore

$$\begin{aligned} P(A \cap B) &= \lim_n P(A \cap B_n) \\ &= \lim_n P(A)P(B_n) \\ &= P(A)P(B), \end{aligned}$$

and the fact that $\mathcal{D}$ is a d-system is proved. By corollary 2.20.(b), the σ–algebra $\mathcal{T}_{n+1} = \sigma(\mathcal{E}_m,\ m \geq n+1)$ is independent of $\sigma(\mathcal{E}_1,\ \ldots,\ \mathcal{E}_n)$. Thus, since $A \in \mathcal{T}$ entails $A \in \mathcal{T}_{n+1}$ for all $n \in \mathbb{N}$, we get that for every $n \in \mathbb{N}$, $\mathcal{D} \supset \sigma(\mathcal{E}_1, \ldots, \mathcal{E}_n)$. Consequently, the family of events $\mathcal{E}$ defined by

$$\mathcal{E} = \bigcup_n \sigma(\mathcal{E}_1, \ldots, \mathcal{E}_n)$$

is a subset of $\mathcal{D}$. We show that it is $\cap$–stable: If $B, C \in \mathcal{E}$, then there exists $n \in \mathbb{N}$ so that $B, C \in \sigma(\mathcal{E}_1, \ldots, \mathcal{E}_n)$, and hence $B \cap C \in \sigma(\mathcal{E}_1, \ldots, \mathcal{E}_n) \subset \mathcal{E}$. Since $\mathcal{D}$ contains $\mathcal{E}$ it also contains the d-system generated by $\mathcal{E}$, which — because $\mathcal{E}$ is $\cap$–stable – is equal to the σ–algebra $\sigma(\mathcal{E})$ generated by $\mathcal{E}$ (see theorem A.15): $\mathcal{D} \supset \sigma(\mathcal{E})$. Now $\sigma(\mathcal{E})$ contains every σ–algebra $\mathcal{E}_n$, $n \in \mathbb{N}$, and therefore also every $\mathcal{T}_n$, $n \in \mathbb{N}$ (because $\mathcal{T}_n$, $n \in \mathbb{N}$, is the smallest σ–algebra containing every $\mathcal{E}_m$, $m \geq n$). But this means that $\mathcal{D} \supset \mathcal{T}$, and we are done. $\square$

Consider an independent sequence $(A_n,\ n \in \mathbb{N})$ of events. Set $\mathcal{E}_n = \sigma(\{A_n\})$ so that $(\mathcal{E}_n,\ n \in \mathbb{N})$ is an independent family of σ–algebras, and we define $\mathcal{T}_n$, $n \in \mathbb{N}$, and $\mathcal{T}$ as above. Let us prove that the event $A = \limsup_n A_n$ belongs to $\mathcal{T}$:

$$A = \limsup_n A_n = \bigcap_{n=1}^{\infty} B_n$$

with $B_n = \cup_{m \geq n} A_m \in \mathcal{T}_n \subset \mathcal{T}_k$ for every $k \leq n$. Clearly, the sequence $(B_n,\ n \in \mathbb{N})$ is decreasing, and hence for every $k \in \mathbb{N}$

$$A = \bigcap_{n=k}^{\infty} B_n \in \mathcal{T}_k.$$

We conclude that $A \in \cap_k \mathcal{T}_k = \mathcal{T}$. Thus we can apply theorem 2.22 which gives us the following result:

2.23 Corollary *Assume that* $(A_n,\ n \in \mathbb{N})$ *is an independent sequence in* $\mathcal{A}$. *Then* $P(\limsup_n A_n) = 0$ *or* 1.

In corollary 3.20 in the next chapter, we shall give a precise condition for when $P(\limsup_n A_n) = 0$ or 1 is true.

We turn our attention from events to random variables.

2.24 Definition Suppose that $(X_i,\ i \in I)$ is a family of random variables, where X_i, $i \in I$, takes values in a measurable space $(\Omega_i, \mathcal{A}_i)$. $(X_i,\ i \in I)$ is called *independent*, if the family of σ–algebras $\big(\sigma(X_i),\ i \in I\big)$ generated by the X_i is independent.

2.25 Theorem *Suppose that $(X_i,\ i \in I)$ is an independent family of random variables, X_i, $i \in I$, taking values in $(\Omega_i, \mathcal{A}_i)$. Assume furthermore, that for every $i \in I$, T_i is a measurable mapping from $(\Omega_i, \mathcal{A}_i)$ into a measurable space $(\Omega_i', \mathcal{A}_i')$. Then the family $(T_i \circ X_i,\ i \in I)$ is independent.*

Proof By remark 2.18, we only have to show that $\sigma(T_i \circ X_i) \subset \sigma(X_i)$ for every $i \in I$. But for $A_i' \in \mathcal{A}_i'$ we get $(T_i \circ X_i)^{-1}(A_i') = X_i^{-1}(T_i^{-1}(A_i'))$, and $T_i^{-1}(A_i') \in \mathcal{A}_i$ so that indeed $(T_i \circ X_i)^{-1}(A_i') \in \sigma(X_i)$. $\square$

Kolmogorov's 0–1–law, theorem 2.22, reads for sequences of random variables as follows:

2.26 Theorem *Suppose that $(X_n,\ n \in \mathbb{N})$ is an independent sequence of random variables. If $A \in \bigcap_n \sigma(X_m,\ m \geq n)$, then $P(A) = 0$ or 1.*

Since the independence of events or of random variables is checked by considering only finitely many, in the sequel we restrict ourselves to the case of finitely many random variables.

2.27 Theorem *Suppose that $(\Omega_i, \mathcal{A}_i)$, $i = 1, \dots, n$, $n \in \mathbb{N}$, are measurable spaces, and that $\mathcal{E}_i$ is for each $i \in \{1, 2, \dots, n\}$ a $\cap$–stable generator of $\mathcal{A}_i$. Assume furthermore that X_i, $i = 1, \dots, n$, are random variables with values in Ω_i, respectively.*

(a) *The family $(X_1, X_2, \dots, X_n)$ is independent, if and only if for each subset $\{i_1, \dots, i_k\} \subset \{1, \dots, n\}$, and every choice of $E_{i_1} \in \mathcal{E}_{i_1}, \dots, E_{i_k} \in \mathcal{E}_{i_k}$*

$$P(X_{i_1} \in E_{i_1}, \dots, X_{i_n} \in E_{i_n}) = \prod_{j=1}^{k} P(X_{i_j} \in E_{i_j}) \tag{2.21}$$

holds true.

(b) *Suppose in addition that for every $i = 1, \dots, n$, $\Omega_i \in \mathcal{E}_i$. Then the family $(X_1, X_2, \dots, X_n)$ is independent, if and only if for all $E_i \in \mathcal{E}_i$, $i = 1, \dots, n$, the equality*

$$P(X_1 \in E_1, \dots, X_n \in E_n) = \prod_{i=1}^{n} P(X_i \in E_i) \tag{2.22}$$

holds.

Proof From lemma A.33 we know that for each $i \in \{1, \dots, n\}$, $X_i^{-1}(\mathcal{E}_i)$ is a generator of $\sigma(X_i)$. Moreover, each of these generators is $\cap$–stable. By corollary 2.20, $\big(\sigma(X_1), \dots, \sigma(X_n)\big)$ is independent, if and only if $\big(X^{-1}(\mathcal{E}_1), \dots, X_n^{-1}(\mathcal{E}_n)\big)$ is independent. By definition this is the case, if and only if for each subset $\{i_1, \dots, i_k\} \subset \{1, \dots, n\}$, and every choice of $E_{i_1} \in \mathcal{E}_{i_1}, \dots, E_{i_k} \in \mathcal{E}_{i_k}$ the following equality holds true:

$$P\big(X_{i_1}^{-1}(E_{i_1}) \cap \cdots \cap X_{i_k}^{-1}(E_{i_k})\big) = \prod_{j=1}^{k} P\big(X_{i_j}^{-1}(E_{i_j})\big),$$

and this proves statement (a). For (b) set $E_i = \Omega_i \in \mathcal{E}_i$ if $i \notin \{i_1, \dots, i_k\}$. Then this equation is the same as equation (2.22). □

For the remainder of this section we only consider real valued random variables $X_1, X_2, \dots, X_n$. A convenient $\cap$-stable generator $\mathcal{E}_0$ of the Borel σ–algebra $\mathcal{B}(\mathbb{R})$ consists of all intervals of the form $(-\infty, x]$, $x \in \mathbb{R}$. (We could just as well take $(-\infty, x)$, $x \in \mathbb{R}$.) With a view towards theorem 2.27 we note that $\mathbb{R}$ does not belong to $\mathcal{E}_0$. On the other hand, if for all $x_1, x_2, \dots, x_n \in \mathbb{R}$ we have

$$P(X_1 \le x_1, X_2 \le x_2, \dots, X_n \le x_n) = \prod_{i=1}^{n} P(X_i \le x_i), \tag{2.23}$$

then we can let any number of the x_i tend trough an increasing sequence to $+\infty$, and use the continuity of the probability measure P. This way we obtain all equations of the form (2.21), and therefore the independence of the random variables $X_1, \dots, X_n$. The converse being obvious, we have proved the following

2.28 Theorem *A family $(X_1, X_2, \dots, X_n)$ of real valued random variables is independent, if and only if for every choice of $x_1, x_2, \dots, x_n \in \mathbb{R}$ equality* (2.23) *holds.*

For $x_1, x_2, \dots, x_n \in \mathbb{R}$ let us set

$$B = (-\infty, x_1] \times (-\infty, x_2] \times \cdots \times (-\infty, x_n] \in \mathcal{B}(\mathbb{R}^n). \tag{2.24}$$

Then equation (2.23) can be written in the following way

$$P_{X_1 \otimes \cdots \otimes X_n}(B) = P_{X_1} \otimes \cdots \otimes P_{X_n}(B),$$

where we have the product measure of the laws of $X_1, \dots, X_n$ on the right hand side (see section A.10 of appendix A for its definition). Since any probability measure on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ is already uniquely determined on a $\cap$–stable generator (see theorem A.21), such as $\mathcal{E}_0$, we obtain the following result.

2.29 Corollary *The family $(X_1, X_2, \dots, X_n)$ of real valued random variables is independent, if and only if*

$$P_{X_1 \otimes \cdots \otimes X_n} = P_{X_1} \otimes \cdots \otimes P_{X_n}. \tag{2.25}$$

Another important consequence for expectation values gives the following

2.30 Theorem *Suppose that $(X_1, X_2, \dots, X_n)$ are independent, real valued random variables. If every X_i, $i \in \{1, 2, \dots, n\}$, is integrable, then this also holds true for $\prod_{i=1}^n X_i$. In this case, or if all X_i, $i \in \{1, 2, \dots, n\}$, are positive, the equality*

$$E\Big(\prod_{i=1}^n X_i\Big) = \prod_{i=1}^n E(X_i) \tag{2.26}$$

is valid.

Proof Let us first assume that all X_i, $i \in \{1, 2, \dots, n\}$, are positive. The transformation theorem A.53 entails

$$\begin{aligned} E\Big(\prod_{i=1}^n X_i\Big) &= \int_\Omega \prod_{i=1}^n X_i \, dP \\ &= \int_{\mathbb{R}_+^n} \prod_{i=1}^n x_i \, dP_{X_1 \otimes \cdots \otimes X_n}(x_1, \dots, x_n) \\ &= \int_{\mathbb{R}_+^n} \prod_{i=1}^n x_i \, d(P_{X_1} \otimes \cdots \otimes P_{X_n})(x_1, \dots, x_n), \end{aligned}$$

where we used corollary 2.29 in the last step. The theorem of Fubini–Tonelli, theorem A.71, gives

$$E\Big(\prod_{i=1}^n X_i\Big) = \prod_{i=1}^n \int_{\mathbb{R}_+} x_i \, dP_{X_i}(x_i) = \prod_{i=1}^n E(X_i),$$

by another application of the transformation theorem. This proves the statement in this case. Now assume that all random variables X_i, $i \in \{1, 2, \dots, n\}$, belong to $\mathcal{L}^1(P)$. Replace every X_i by $|X_i|$ in the preceding computation, so that the computation shows that $\prod_{i=1}^n X_i$ is integrable. Therefore we may apply also in this case the theorem of Fubini–Tonelli (replacing everywhere above $\mathbb{R}_+^n$ and $\mathbb{R}_+$ by $\mathbb{R}^n$, $\mathbb{R}$, respectively), and get the result also in the integrable case. □

2.31 Corollary *Assume that X, Y are real valued random variables which are integrable and independent. Then they are uncorrelated.*

Proof Theorem 2.30 implies that $E(XY) = E(X)E(Y)$, so that $\mathrm{Cov}(X, Y) = 0$.□

2.4 Sums of Independent Random Variables

2.32 Lemma *Assume that X_1, X_2, …, X_n, $n \in \mathbb{N}$, are real valued, independent random variables. Then for the moment generating function $\psi_{X_1+\cdots+X_n}$, and for the* *characteristic function* $\varphi_{X_1+\cdots+X_n}$ *of their sum the following formulae*

$$\psi_{X_1+\cdots+X_n}(\lambda) = \prod_{i=1}^{n} \psi_{X_i}(\lambda) \tag{2.27}$$

$$\varphi_{X_1+\cdots+X_n}(\lambda) = \prod_{i=1}^{n} \varphi_{X_i}(\lambda) \tag{2.28}$$

hold true for all $\lambda \in \mathbb{R}$.

Proof We only give the simple argument for the characteristic function, the other case is similar. Obviously, for all $\lambda \in \mathbb{R}$, $i \in \{1, 2, \dots, n\}$, the random variables $\exp(i\lambda X_i)$ are integrable, and we have that

$$\exp\big(i\lambda(X_1 + X_2 + \cdots + X_n)\big) = \exp(i\lambda X_1)\ \exp(i\lambda X_2)\cdots\exp(i\lambda X_n).$$

Therefore equality (2.28) follows directly from theorem 2.30. □

Suppose that $\mu_1, \mu_2, \dots, \mu_n$, $n \in \mathbb{N}$, are σ–finite measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Then we have their well-defined product measure $\mu_1 \otimes \mu_2 \otimes \cdots \otimes \mu_n$ on $(\mathbb{R}^n \mathcal{B}(\mathbb{R}^n))$ (see section A.10 of appendix A). Consider the mapping

$$\begin{aligned} S_n : R^n &\to \mathbb{R} \\ (x_1, x_2, \dots, x_n) &\mapsto x_1 + x_2 + \cdots + x_n. \end{aligned}$$

Obviously, S_n being continuous it is measurable. Define a measure $\mu_1 * \mu_2 * \cdots * \mu_n$ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ as the image measure of $\mu_1 \otimes \mu_2 \otimes \cdots \otimes \mu_n$ under S_n.

2.33 Definition The measure $\mu_1 * \mu_2 * \cdots * \mu_n$ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is called the *convolution* of $\mu_1, \mu_2, \dots, \mu_n$.

2.34 Exercise

(a) Show that for σ–finite measures μ_1, μ_2 on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$,

$$\mu_1 * \mu_2(B) = \int_{\mathbb{R}} \mu_1(B - x)\, d\mu_2(x). \tag{2.29}$$

(b) Show that $*$ is an associative, commutative, distributive product on the positive cone of σ–finite measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

(c) Show that for $x_1, x_2 \in \mathbb{R}$, $\varepsilon_{x_1} * \varepsilon_{x_2} = \varepsilon_{x_1+x_2}$.

(d) Show that if $\mu_i = \mathcal{N}(m_i, \sigma_i^2)$, $m_i \in \mathbb{R}$, $\sigma_i^2 > 0$, $i = 1, 2$, then $\mu_1 * \mu_2 = \mathcal{N}(m_1 + m_2, \sigma_1^2 + \sigma_2^2)$.

2.35 Theorem *Suppose that $(X_1, X_2, \ldots, X_n)$ are independent, real valued random variables. Then*

$$P_{X_1+X_2+\cdots+X_n} = P_{X_1} * P_{X_2} * \cdots * P_{X_n}. \tag{2.30}$$

Proof Write $X_1 + X_2 + \cdots + X_n = S_n \circ (X_1 \otimes X_2 \otimes \cdots \otimes X_n)$. Then

$$\begin{aligned} P_{X_1+X_2+\cdots+X_n} &= P_{S_n \circ (X_1 \otimes X_2 \otimes \cdots \otimes X_n)} \\ &= S_n P_{X_1 \otimes X_2 \otimes \cdots \otimes X_n} \\ &= S_n\big(P_{X_1} \otimes P_{X_2} \otimes \cdots \otimes P_{X_n}\big) \\ &= P_{X_1} * P_{X_2} * \cdots * P_{X_n}, \end{aligned}$$

where we used corollary 2.29 and the definition of the convolution. □

2.5 Trials

Consider a trial, that is, a sequence $((\Omega_n, \mathcal{A}_n, P_n),\ n \in \mathbb{N})$ of experiments which are carried out without influencing each other. We shall develop a model for the joint experiment. Set

$$\Omega = \bigtimes_{n=1}^{\infty} \Omega_n,$$

thus $\omega \in \Omega$ is a sequence $\omega = (\omega_n,\ n \in \mathbb{N})$ such that $\omega_n \in \Omega_n$, $n \in \mathbb{N}$. Let $\mathcal{I}_0$ be the family of all finite subsets of $\mathbb{N}$, and for $I \in \mathcal{I}_0$ define

$$\Omega_I = \bigtimes_{i \in I} \Omega_i, \quad \mathcal{A}_I = \bigotimes_{i \in I} \mathcal{A}_i, \quad P_I = \bigotimes_{i \in I} P_i,$$

and the mapping

$$\begin{aligned} \pi_I : \Omega &\to \Omega_I \\ \omega &\mapsto (\omega_i,\ i \in I). \end{aligned}$$

Furthermore, consider the family $\mathcal{Z}$ of subsets Z of Ω which are of the following form:

$$Z = (\pi_I)^{-1}(A_I), \quad A_I = \bigtimes_{i \in I} A_i, \quad A_i \in \mathcal{A}_i,\ i \in I, \quad I \in \mathcal{I}_0. \tag{2.31}$$

Sets of this form will be called *cylinder sets* in Ω. Thus Z is for the form

$$Z = \bigtimes_{i \in \mathbb{N}} A_i,$$

with $A_i = \Omega_i$ for all $i \notin I$. It is clear that the family $\mathcal{Z}$ of cylinder sets does not form a σ–algebra. Define $\mathcal{A} = \sigma(\mathcal{Z})$. For $Z \in \mathcal{Z}$ of the form above we define

$$P(Z) = P_I(\pi_I Z) = P_I(A_I) = \prod_{i \in I} P_i(A_i), \tag{2.32}$$

that is, for every $I \in \mathcal{I}_0$ we have set $\pi_I P = P_I$. It is a non-trivial fact (see, e.g., [5, Satz 9.2]), that equation (2.32) defines a unique probability measure on $(\Omega, \mathcal{A})$, but we shall not give a proof here.

Let $(A_n,\, n \in \mathbb{N})$ be any sequence of events, $A_n \in \mathcal{A}_n$, $n \in \mathbb{N}$. We embed A_n into the joint experiment by defining the cylinder set

$$\tilde{A}_n = \pi_{\{n\}}^{-1}(A_n).$$

We show that the family $(\tilde{A}_n,\, n \in \mathbb{N})$ is independent. To this end, it is sufficient to show that for every $I \in \mathcal{I}_0$, we have

$$P\Big(\bigcap_{i\in I} \tilde{A}_i\Big) = \prod_{i\in I} P(\tilde{A}_i). \tag{2.33}$$

Now $\cap_{i\in I} \tilde{A}_i = \pi_I^{-1}(A_I)$, and therefore

$$\begin{aligned} P\Big(\bigcap_{i\in I} \tilde{A}_i\Big) &= P\big(\pi_I^{-1}(A_I)\big) \\ &= P_I(A_I) \\ &= \prod_{i\in I} P_i(A_i) \\ &= \prod_{i\in I} P(\tilde{A}_i), \end{aligned}$$

and equation (2.33) is verified. Thus the *product probability space* $(\Omega, \mathcal{A}, P)$ has built in the probabilistic independence of the events of the individual experiments.

2.36 Exercise Suppose that $(X_n,\, n \in \mathbb{N})$ is a sequence of mappings, such that for each $n \in \mathbb{N}$, X_n is a real valued random variable on $(\Omega_n, \mathcal{A}_n, P_n)$. For $n \in \mathbb{N}$ define a real valued random variable $\tilde{X}_n$ on the product space $(\Omega, \mathcal{A}, P)$ by setting $\tilde{X}_n = X_n \circ \pi_{\{n\}}$. Show that $(\tilde{X}_n,\, n \in \mathbb{N})$ is an independent family of random variables, such that $\tilde{X}_n$, $n \in \mathbb{N}$, has under P the same law as X_n under P_n.

Chapter 3

Almost Sure Convergence, Convergence in Probability

Various kinds of convergence are used in probability theory, and in this chapter we study two of them, *almost sure convergence* and *convergence in probability* (sometimes also called *stochastic convergence*), and their relation. Sometimes the beginner is irritated by the use of so many notions of convergence (we shall use altogether six in these lectures). However, one should consider these as *tools*, and having various tools at hand, one can choose the appropriate or "natural" one in order to solve a given problem.

Throughout we base our discussion on an underlying probability space $(\Omega, \mathcal{A}, P)$, and consider real valued random variables, unless otherwise stated. Most of the time, the generalization to complex or $\mathbb{R}^n$–valued random variables is obvious, but we will not need this here.

3.1 Almost Sure Convergence

3.1 Definition A subset N of Ω is called *P–negligible* or a *P–null set*, if $N \in \mathcal{A}$ and $P(N) = 0$.

3.2 Remark If the probability measure P is understood from the context — as it is the case in this chapter — we shall simply say "negligible" or "null set". When comparing with other literature the reader should be careful: Sometimes these notions mean that N is a subset of a set in $\mathcal{A}$ having probability zero.

3.3 Definition Suppose that $(X_n, n \in \mathbb{N})$ is a sequence of real valued random variables.

(a) $(X_n, n \in \mathbb{N})$ is called *almost surely Cauchy* (*a.s. Cauchy*), if there exists a null set N such for every $\omega \in \complement N$ the real sequence $(X_n(\omega), n \in \mathbb{N})$ is a Cauchy sequence.

(b) $(X_n,\, n \in \mathbb{N})$ *converges almost surely* (*converges a.s.*) to a real valued random variable X, if there exists a null set N such for every $\omega \in \complement N$ the real sequence $(X_n(\omega),\, n \in \mathbb{N})$ converges to $X(\omega)$.

(c) $(X_n,\, n \in \mathbb{N})$ is *almost surely convergent* (*a.s convergent*), if there exists a real valued random variable X, such that $(X_n,\, n \in \mathbb{N})$ converges a.s. to X.

Notation That $(X_n,\, n \in \mathbb{N})$ converges almost surely to X we also write as

$$X_n \to X, \text{ a.s,} \quad \text{or as} \quad X_n \xrightarrow{\text{a.s.}} X.$$

3.4 Exercise Let $(X_n,\, n \in \mathbb{N})$ be a sequence of real valued random variables.

(a) Show that $(X_n,\, n \in \mathbb{N})$ is a.s. convergent, if and only if it is a.s. Cauchy.

(b) Show that a.s. limits are in general not unique, but they are *almost surely unique*, i.e., if $(X_n,\, n \in \mathbb{N})$ converges a.s. to X and to Y, then $P(X = Y) = 1$.

The next lemma provides a convenient criterion for a.s. convergence. Since we can consider instead of $(X_n,\, n \in \mathbb{N})$ the sequence $(X_n - X,\, n \in \mathbb{N})$, it is sufficient to treat a.s. convergence to zero. The reader is asked to prepare the proof with the following

3.5 Exercise Suppose that $(X_n,\, n \in \mathbb{N})$ is a sequence of real valued random variables, and that $\varepsilon > 0$. Show that for every $n \in \mathbb{N}$,

$$\big\{\sup_{m\geq n} |X_m| > \varepsilon\big\} = \bigcup_{m\geq n} \{|X_m| > \varepsilon\}, \tag{3.1}$$

and conclude that

$$\lim_{n\to\infty} \big\{\sup_{m\geq n} |X_m| > \varepsilon\big\} = \limsup_{n\to\infty} \{|X_n| > \varepsilon\}. \tag{3.2}$$

(*Hint*: It is convenient to consider the complements of the events in (3.1).)

3.6 Lemma *The following are equivalent:*

(a) $(X_n,\, n \in \mathbb{N})$ *converges a.s. to zero.*

(b) *For every* $\varepsilon > 0$, $\lim_{n\to\infty} P(\sup_{m\geq n} |X_m| > \varepsilon) = 0$.

(c) *For every* $\varepsilon > 0$, $P(\limsup_n \{|X_n| > \varepsilon\}) = 0$.

Proof From equation (3.2) and the continuity of the probability measure P we get

$$\lim_{n\to\infty} P\big(\sup_{m\geq n} |X_m| > \varepsilon\big) = P\big(\limsup_{n\to\infty}\{|X_n| > \varepsilon\}\big),$$

which proves the equivalence of statements (b) and (c).

For $\varepsilon > 0$ set

$$N(\varepsilon) = \limsup_{n\to\infty}\{|X_n| > \varepsilon\}.$$

Now we prove "(a) $\Rightarrow$ (c)". Assume that (a) holds true, but (c) is false. Then there exists $\varepsilon > 0$ so that $P(N(\varepsilon)) > 0$. Let $\omega \in N(\varepsilon)$. Then this means that $|X_n(\omega)| > \varepsilon$ for infinitely many $n \in \mathbb{N}$. Thus $(X_n(\omega),\, n \in \mathbb{N})$ does not converge to zero for all ω in a set of strictly positive probability, which contradicts (a).

For the proof of "(c) $\Rightarrow$ (a)" we set

$$N = \bigcup_{k\in\mathbb{N}} N\Big(\frac{1}{k}\Big).$$

(c) implies that $N(1/k)$ is a P–null set for every $k \in \mathbb{N}$, and therefore also N is a P–null set. Observe that

$$\complement N = \bigcap_{k\in\mathbb{N}} \bigcup_{n\in\mathbb{N}} \bigcap_{m\geq n} \Big\{|X_m| \leq \frac{1}{k}\Big\}.$$

So for an $\omega \in \complement N$ we find

$$\forall\, k \in \mathbb{N}\, \exists\, n \in \mathbb{N}\, \forall\, m \geq n : |X_m(\omega)| \leq \frac{1}{k}.$$

In other words, $(X_n(\omega),\, n \in \mathbb{N})$ converges to zero. Thus (a) holds true. □

3.7 Remark Since for every real valued random variable Z and $\varepsilon > 0$ we have the inclusions

$$\{Z > \varepsilon\} \subset \{Z \geq \varepsilon\} \subset \{Z > \varepsilon/2\}, \tag{3.3}$$

one equivalently can replace ">" in the statements (b) and (c) of lemma 3.6 by "≥".

For the a.s. Cauchy property we have the following result:

3.8 Lemma *The following statements are equivalent:*

(a) $(X_n,\, n \in \mathbb{N})$ *is a.s. convergent.*

(b) $(X_n,\, n \in \mathbb{N})$ *is a.s. Cauchy.*

(c) *For every* $\varepsilon > 0$*,*

$$\lim_{n\to\infty} P\Big(\sup_{k,m\geq n} |X_k - X_m| > \varepsilon\Big) = 0.$$

(d) *For every $\varepsilon > 0$,*

$$\lim_{n\to\infty} P\Big(\sup_{m\geq n} |X_n - X_m| > \varepsilon\Big) = 0.$$

Moreover, in statements (c) and (d), "$>$" can equivalently be replaced by "$\geq$".

Proof The equivalence of (a) and (b) has already been shown in exercise 3.4. The equivalence of (c) and (d) follows from the inclusions

$$\Big\{\sup_{m\geq n} |X_n - X_m| > \varepsilon\Big\} \subset \Big\{\sup_{k,m\geq n} |X_k - X_m| > \varepsilon\Big\},$$

$$\Big\{\sup_{k,m\geq n} |X_k - X_m| > \varepsilon\Big\} \subset \Big\{\sup_{m\geq n} |X_n - X_m| > \frac{\varepsilon}{2}\Big\},$$

of which the first one is trivial. The second one is equivalent to

$$\Big\{\sup_{k,m\geq n} |X_k - X_m| \leq \varepsilon\Big\} \supset \Big\{\sup_{m\geq n} |X_n - X_m| \leq \frac{\varepsilon}{2}\Big\},$$

but for an ω belonging to the set on the right hand side, and for $k, m \geq n$ we have

$$|X_k(\omega) - X_m(\omega)| \leq |X_k(\omega) - X_n(\omega)| + |X_n(\omega) - X_m(\omega)| \leq \varepsilon,$$

so that this inclusion is proved, too.

It remains to show the equivalence of (b) and (c). This is similar to the proof of the equivalences in lemma 3.6, and the proof is only sketched. For $\varepsilon > 0$ set

$$N'(\varepsilon) = \bigcap_{n\in\mathbb{N}} \bigcup_{k,m\geq n} \{|X_k - X_m| > \varepsilon\} = \lim_{n\to\infty} \bigcup_{k,m\geq n} \{|X_k - X_m| > \varepsilon\},$$

and

$$N' = \bigcup_{j\in\mathbb{N}} N'\Big(\frac{1}{j}\Big).$$

Assume that (b) holds true, and that (c) does not. Then there exists $\varepsilon > 0$ so that $N'(\varepsilon)$ has strictly positive probability, and for every $\omega \in N'(\varepsilon)$ there are infinitely many $k, m \in \mathbb{N}$ so that $|X_k(\omega) - X_m(\omega)| > \varepsilon$, contradicting (b). On the other hand, (c) implies that N' is a P–null set, and for $\omega \in \complement N'$, we have that $(X_n(\omega),\ n \in \mathbb{N})$ is a real Cauchy sequence, i.e., (c) entails (b).

Finally, remark 3.7 applies again and gives the last statement of the lemma. □

3.2 Convergence in Probability

3.9 Definition Let $(X_n,\ n \in \mathbb{N})$ be a sequence of real valued random variables.

(a) $(X_n,\ n \in \mathbb{N})$ is said to be *Cauchy in probability*, if for every $\varepsilon > 0$,

$$\lim_{n,m\to\infty} P(|X_n - X_m| > \varepsilon) = 0. \tag{3.4}$$

(b) $(X_n,\ n \in \mathbb{N})$ is called to be *convergent in probability to a real valued random variable* X, if for every $\varepsilon > 0$,

$$\lim_{n\to\infty} P(|X_n - X| > \varepsilon) = 0. \tag{3.5}$$

(c) $(X_n,\ n \in \mathbb{N})$ is called to be *convergent in probability* if there exists a real valued random variable X, so that $(X_n,\ n \in \mathbb{N})$ converges in probability to X.

Notation If $(X_n,\ n \in \mathbb{N})$ converges in probability to X, we also write $X_n \xrightarrow{\mathrm{P}} X$.

3.10 Remarks Again remark 3.7 applies, so that one can equivalently replace above "$>$" by "$\geq$". Sometimes convergence in probability is also called *stochastic convergence*.

3.11 Exercise Prove that if $(X_n,\ n \in \mathbb{N})$ converges in probability to X, then also every subsequence converges in probability to X.

3.12 Exercise Show that $(X_n,\ n \in \mathbb{N})$ converges in probability to X, if and only if $\big(E(|X_n - X| \wedge 1),\ n \in \mathbb{N}\big)$ converges to zero.

3.13 Lemma *Suppose that X, Y, Z a real valued random variables, and that $\alpha > 0$. Then*

$$\big\{|X - Y| > \alpha\big\} \subset \Big(\Big\{|X - Z| > \frac{\alpha}{2}\Big\} \cup \Big\{|Y - Z| > \frac{\alpha}{2}\Big\}\Big) \tag{3.6}$$

holds true.

Proof Suppose that ω does not belong to the set on the right hand side. Then

$$|X(\omega) - Y(\omega)| \leq |X(\omega) - Z(\omega)| + |Y(\omega) - Z(\omega)| \leq \alpha$$

shows that ω does not belong to the left hand side. $\square$

3.14 Exercise Prove that if $(X_n,\ n \in \mathbb{N})$ is convergent in probability, then it is Cauchy in probability. (*Hint:* Use (3.6).)

3.15 Lemma *If $(X_n,\ n \in \mathbb{N})$ is convergent in probability, then its limit is a.s. unique.*

Proof Assume that $(X_n,\ n \in \mathbb{N})$ converges in probability to X and to Y. For all m, $n \in \mathbb{N}$ we get from lemma 3.13

$$\Big\{|X - Y| > \frac{1}{m}\Big\} \subset \Big\{|X - X_n| > \frac{1}{2m}\Big\} \cup \Big\{|Y - X_n| > \frac{1}{2m}\Big\}.$$

Theorem A.20.(a) yields

$$P\Big(|X - Y| > \frac{1}{m}\Big) \leq P\Big(|X - X_n| > \frac{1}{2m}\Big) + P\Big(|Y - X_n| > \frac{1}{2m}\Big).$$

Letting n tend to infinity, the right hand side converges to zero, so that we obtain that for all $m \in \mathbb{N}$, $P(|X - Y| > 1/m) = 0$. Then

$$\begin{aligned} P(X \neq Y) &= P(|X - Y| > 0) \\ &= P\Big(\bigcup_{m \in \mathbb{N}} \Big\{|X - Y| > \frac{1}{m}\Big\}\Big) \\ &= \lim_{m \to \infty} P\Big(|X - Y| > \frac{1}{m}\Big) \\ &= 0. \end{aligned}$$ □

3.16 Lemma *If $(X_n, n \in \mathbb{N})$ converges almost surely to X, then it converges in probability to X.*

Proof Without loss of generality we may assume that $X = 0$. Hence suppose that $X_n \xrightarrow{\text{a.s.}} 0$. From lemma 3.6 we have that for every $\varepsilon > 0$,

$$\lim_{n \to \infty} P\Big(\sup_{m \geq n} |X_m| > \varepsilon\Big) = 0.$$

Clearly

$$\{|X_n| > \varepsilon\} \subset \Big\{\sup_{m \geq n} |X_m| > \varepsilon\Big\},$$

and therefore

$$P(|X_n| > \varepsilon) \leq P\Big(\sup_{m \geq n} |X_m| > \varepsilon\Big) \to 0,$$

finishing the proof. □

3.17 Remark The converse of the statement of lemma 3.16 does in general *not* hold true — a counterexample is given by example 4.4 in the next chapter.

3.3 Borel–Cantelli–Lemma

In this section we prove the *Borel–Cantelli–Lemma*. It is a technical device which — in spite of its relatively simple proof — is of central importance in probability theory.

3.18 Lemma (Borel–Cantelli) *Let $(A_n, n \in \mathbb{N})$ be a sequence in $\mathcal{A}$.*

(a) *If $\sum_{n \in \mathbb{N}} P(A_n) < +\infty$, then $P(\limsup_n A_n) = 0$.*

(b) *If the sequence $(A_n, n \in \mathbb{N})$ is independent, and if $\sum_{n \in \mathbb{N}} P(A_n) = +\infty$, then $P(\limsup_n A_n) = 1$.*

Proof For the proof of (a) we remark that

$$\limsup_n A_n = \bigcap_{n\in\mathbb{N}} \bigcup_{m\geq n} A_m = \lim_{n\to\infty} \bigcup_{m\geq n} A_m,$$

and therefore by the continuity of probability measures (see theorem A.23) we obtain

$$P\big(\limsup_n A_n\big) = \lim_{n\to\infty} P\Big(\bigcup_{m\geq n} A_m\Big) \leq \limsup_{n\to\infty} \sum_{m=n}^{\infty} P(A_m) = 0$$

where we used theorem A.20.(b), and the hypothesis.

For the proof of (b) we first show that for every $n \in \mathbb{N}$,

$$P\Big(\bigcap_{m=n}^{\infty} \complement A_m\Big) = 0.$$

To this end, write

$$\bigcap_{m=n}^{\infty} \complement A_m = \lim_{N\to\infty} \bigcap_{m=n}^{N} \complement A_m,$$

use the independence of $\complement A_n, \ldots, \complement A_n$, and the continuity of P (see theorem A.23) to calculate as follows:

$$\begin{aligned}
P\Big(\bigcap_{m=n}^{\infty} \complement A_m\Big) &= P\Big(\lim_{N\to\infty} \bigcap_{m=n}^{N} \complement A_m\Big) \\
&= \lim_{N\to\infty} P\Big(\bigcap_{m=n}^{N} \complement A_m\Big) \\
&= \lim_{N\to\infty} \prod_{m=n}^{N} P\big(\complement A_m\big) \\
&= \lim_{N\to\infty} \prod_{m=n}^{N} \big(1 - P(A_m)\big).
\end{aligned}$$

Since $1 - x \leq e^{-x}$ for all $x \geq 0$ (as the reader will check), we find

$$\begin{aligned}
P\Big(\bigcap_{m=n}^{\infty} \complement A_m\Big) &\leq \limsup_{N\to\infty} \prod_{m=n}^{N} \exp(-P(A_m)) \\
&= \limsup_{N\to\infty} \exp\Big(-\sum_{m=n}^{N} P(A_m)\Big) \\
&= 0,
\end{aligned}$$

due the hypothesis that the series $\sum_n P(A_n)$ diverges. To conclude the proof, we compute in the following way:

$$\begin{aligned} P\big(\limsup_n A_n\big) &= 1 - P\big(\complement \limsup_n A_n\big) \\ &= 1 - P\big(\liminf_n \complement A_n\big) \\ &= 1 - P\Big(\bigcup_{n=1}^{\infty} \bigcap_{m=n}^{\infty} \complement A_m\Big) \\ &= 1 - \lim_{n\to\infty} P\Big(\bigcap_{m=n}^{\infty} \complement A_m\Big) \\ &= 1. \end{aligned}$$ □

3.19 Remark For the proof of statement (b) it is actually sufficient to assume only that the events are *pairwise independent*, e.g., [11, Theorem 4.2.5] or [5, Lemma 11.1].

As a corollary to the Borel–Cantelli–lemma, lemma 3.18, we can now sharpen the result of corollary 2.23 of the previous chapter:

3.20 Corollary *Assume that $(A_n,\ n \in \mathbb{N})$ is an independent sequence of events. Then $P(\limsup_n A_n) = 0$ or 1 according to whether $\sum_n P(A_n)$ is convergent or not.*

3.21 Example Consider a sequence of (independent) coin tossing. We ask for the probability that infinitely often heads show up twice one after another. Let $p \in (0, 1)$ be the probability that head shows up in an individual trial, and let $(X_n,\ n \in \mathbb{N})$ be an independent sequence of 0–1–Bernoulli random variables with parameter p, modeling the experiment, where "1" stands for heads, "0" for tails. Define $A_n = \{X_n = 1, X_{n+1} = 1\}$. Obviously, the sequence $(A_n,\ n \in \mathbb{N})$ is not independent, but just as obviously $(A_{2n},\ n \in \mathbb{N})$ is. We have that $P(A_{2n}) = p^2 > 0$, and the Borel–Cantelli–lemma, lemma 3.18, implies that $P(\limsup_n A_{2n}) = 1$. Since the event we asked for, namely "infinitely often heads show up twice one after another" contains the event $\limsup_n A_{2n}$, we find that the event in question has probability one.

Clearly we can modify this example to prove: For every $n \in \mathbb{N}$, the probability that infinitely often heads show up n times in a row (or any other pattern of length n one may think of) is equal to one. ◇

Next we apply the Borel–Cantelli–lemma to convergence in probability to prove

3.22 Lemma *Assume that the sequence $(X_n,\ n \in \mathbb{N})$ of real valued random variables is Cauchy in probability. Then it has a subsequence which is a.s. Cauchy.*

Proof By hypothesis, for every choice of $\varepsilon, \delta > 0$ there exists $n_0 \in \mathbb{N}$ so that for all $m, n \geq n_0$ we have that $P(|X_m - X_n| > \varepsilon) \leq \delta$. Choose for $k \in \mathbb{N}$, $\varepsilon = \delta = 2^{-k}$, and rename the index n_0 into n_k. Thus for every $k \in \mathbb{N}$ and all $m \geq n_k$ we get

$$P\big(|X_m - X_{n_k}| > 2^{-k}\big) \leq 2^{-k}.$$

Consider the subsequence $(X_{n_k}, k \in \mathbb{N})$. Then we have for all $k \in \mathbb{N}$ (by choosing $m = n_{k+1}$),

$$P\big(|X_{n_k} - X_{n_{k+1}}| > 2^{-k}\big) \leq 2^{-k}.$$

Set $Y_k = X_{n_k}$, $k \in \mathbb{N}$. We prove that $(Y_k, k \in \mathbb{N})$ is a.s. Cauchy. To this end, consider the events

$$A_k = \big\{|Y_{k+1} - Y_k| > 2^{-k}\big\},\ k \in \mathbb{N}, \quad A = \limsup_k A_k.$$

Clearly, $\sum_k P(A_k)$ is convergent, so that the Borel–Cantelli–lemma, lemma 3.18, implies that $P(A) = 0$. Consider $\omega \in \complement A$, that is $\omega \in \liminf_k \complement A_k$. Then there exists an $k_0(\omega) \in \mathbb{N}$, so that for all $k \geq k_0(\omega)$, $\omega \in \complement A_k$:

$$|Y_{k+1}(\omega) - Y_k(\omega)| \leq 2^{-k}, \qquad k \geq k_0(\omega).$$

This entails that

$$\sum_{k=1}^{\infty} |Y_{k+1}(\omega) - Y_k(\omega)| < +\infty,$$

which in turn — by a well-known result from analysis — implies that $(Y_k(\omega), k \in \mathbb{N})$ is a Cauchy sequence. In other words, $(Y_k = X_{n_k}, k \in \mathbb{N})$ is a.s. Cauchy, and the proof is done.

For convenience of the reader, we include here a proof of the above mentioned result from analysis: Suppose that $(a_n, n \in \mathbb{N})$ is a real sequence such that the series $\sum_n |a_{n+1} - a_n|$ is convergent. Given $\varepsilon > 0$ choose $n_0 \in \mathbb{N}$ such that $\sum_{n \geq n_0} |a_{n+1} - a_n| \leq \varepsilon$. Then we can estimate for all $m, n \geq n_0$, say $m \geq n+1$, as follows:

$$|a_m - a_n| = \Big|\sum_{l=n}^{m-1} (a_{l+1} - a_l)\Big| \leq \sum_{l=n_0}^{\infty} |a_{l+1} - a_l| \leq \varepsilon. \qquad \square$$

3.23 Corollary

(a) *Suppose that* $(X_n, n \in \mathbb{N})$ *converges to zero in probability. Then a subsequence converges a.s. to zero.*

(b) $(X_n, n \in \mathbb{N})$ *converges in probability, if and only if it is Cauchy in probability.*

Proof (a): Since $(X_n, n \in \mathbb{N})$ converges in probability, it is Cauchy in probability by exercise 3.14. It follows from lemma 3.22 that there exists a subsequence $(X_{n_k}, k \in \mathbb{N})$ which is a.s. Cauchy. Lemma 3.8 implies that $(X_{n_k}, k \in \mathbb{N})$ is a.s. convergent, that is, there exists a real valued random variable X so that $(X_{n_k},\, k \in \mathbb{N})$ converges a.s. to X. Lemma 3.16 entails that $(X_{n_k},\, k \in \mathbb{N})$ converges to X in probability. On the other hand, exercise 3.11 and the hypothesis give that $(X_{n_k},\, k \in \mathbb{N})$ converges to zero in probability. Now lemma 3.15 implies that $X = 0$ a.s., and (a) is proved.

(b): Due to exercise 3.14 we only have to show that “Cauchy in probability” implies convergence in probability. Hence assume that $(X_n,\, n \in \mathbb{N})$ is Cauchy in probability. As in the proof of (a) above, this entails that a subsequence $(X_{n_k},\, k \in \mathbb{N})$ is a.s. convergent to some random variable X. By lemma 3.16 this subsequence converges also in probability to X. For given $\varepsilon > 0$ consider $P(|X_n - X| > \varepsilon)$. With (3.6) we find for every $k \in \mathbb{N}$,

$$\big\{|X_n - X| > \varepsilon\big\} \subset \Big\{|X_n - X_{n_k}| > \frac{\varepsilon}{2}\Big\} \cup \Big\{|X - X_{n_k}| > \frac{\varepsilon}{2}\Big\}.$$

Thus

$$P\big(|X_n - X| > \varepsilon\big) \le P\Big(|X_n - X_{n_k}| > \frac{\varepsilon}{2}\Big) + P\Big(|X - X_{n_k}| > \frac{\varepsilon}{2}\Big).$$

The first term on the right hand side vanishes for $n, k \to \infty$ because $(X_n,\, n \in \mathbb{N})$ is Cauchy in probability. The second term vanishes for $k \to \infty$, as has already been remarked before. Thus $(X_n,\, n \in \mathbb{N})$ converges in probability to X. □

3.4 $\mathcal{L}^p$–Convergence

For $p \ge 1$ consider the spaces $\mathcal{L}^p(\Omega, \mathcal{A}, P)$ of random variables, see also section A.9 in appendix A: $X \in \mathcal{L}^p(\Omega, \mathcal{A}, P)$, if and only if $|X|^p$ is P–integrable. On $\mathcal{L}^p(\Omega, \mathcal{A}, P)$ define a seminorm $\|\cdot\|_p$ by

$$\|X\|_p = \big(E(|X|^p)\big)^{1/p}.$$

As in appendix A, we shall often simply write $\mathcal{L}^p(P)$ instead of $\mathcal{L}^p(\Omega, \mathcal{A}, P)$.

3.24 Definition Let $p \ge 1$, and suppose that $(X_n,\, n \in \mathbb{N})$ is a sequence of real valued random variables in $\mathcal{L}(P)$.

(a) $(X_n,\, n \in \mathbb{N})$ is called a *Cauchy sequence in* $\mathcal{L}(P)$, if $\|X_n - X_m\|_p$ converges to zero as $m, n \to \infty$.

(b) $(X_n,\, n \in \mathbb{N})$ is said to *converge in* $\mathcal{L}^p(P)$ *to* $X \in \mathcal{L}(P)$, if $\|X_n - X\|_p$ converges to zero as $n \to \infty$.

The *Riesz–Fischer–theorem* (see theorem A.69) states that $\mathcal{L}^p(P)$, $p \ge 1$, is complete relative to $\|\cdot\|_p$.

In order to relate convergence in $\mathcal{L}^p(P)$ to the notions of convergence we discussed before, we prove

3.25 Theorem (Chebyshev's Inequality) *Suppose that μ is a probability measure on* $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$*, and that f is a positive, strictly increasing function on $\mathbb{R}_+$. Then for all $\alpha > 0$, the inequality*

$$\mu\big(\{x \in \mathbb{R},\ |x| \geq \alpha\}\big) \leq f(\alpha)^{-1} \int_{\mathbb{R}} f(|x|)\, d\mu(x) \tag{3.7}$$

holds true. In particular, if X is a real valued random variable, and f and α are as above, then

$$P(|X| \geq \alpha) \leq f(\alpha)^{-1} E(f \circ |X|). \tag{3.8}$$

3.26 Remark Observe that if f is an increasing function on $\mathbb{R}_+$, then it is automatically measurable (*exercise!*), so that we did not need to require this in the hypothesis of theorem 3.25.

Proof (of theorem 3.25) Since $f \geq 0$, and f is strictly increasing we have $f(\alpha) > 0$, and $\{|x| \geq \alpha\} = \{f(|x|) \geq f(\alpha)\}$. So we can estimate as follows:

$$\begin{aligned}
\int_{\mathbb{R}} f(|x|)\, d\mu(x) &\geq f(\alpha) \int_{\{f(|x|) \geq f(\alpha)\}} f(\alpha)^{-1}\, f(|x|)\, d\mu(x) \\
&\geq f(\alpha)\, \mu\big(\{f(|x|) \geq f(\alpha)\}\big) \\
&= f(\alpha)\, \mu\big(\{|x| \geq \alpha\}\big).
\end{aligned}$$

The last statement of the lemma follows directly from the choice $\mu = P_X$ and the transformation theorem. □

The most common choice for f is $f(x) = x^p$, $x \in \mathbb{R}_+$, for $p \geq 1$, so that one gets estimates of the form

$$P(|X| \geq \alpha) \leq \alpha^{-p} E(|X|^p). \tag{3.9}$$

Another typical choice is $f(x) = \exp(\lambda x)$, $x \in \mathbb{R}_+$, with $\lambda > 0$, yielding

$$P(|X| \geq \alpha) \leq e^{-\lambda\alpha}\, E\big(e^{\lambda |X|}\big). \tag{3.10}$$

A further interesting possibility is to use the function

$$f_\alpha(x) = 1 - \exp\Big(-\frac{x^2}{2\alpha^2}\Big), \qquad x \in \mathbb{R}_+, \tag{3.11}$$

which leads to the following result:

3.27 Corollary *Suppose that μ is a probability measure on* $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$*, with* characteristic function $\hat{\mu}$*:*

$$\hat{\mu}(t) = \int_{\mathbb{R}} e^{itx}\, d\mu(x), \qquad t \in \mathbb{R}. \tag{3.12}$$

Let γ denote the standard Gaussian measure on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$:

$$\gamma(B) = \frac{1}{\sqrt{2\pi}} \int_B e^{-x^2/2}\, dx, \qquad B \in \mathscr{B}(\mathbb{R}). \tag{3.13}$$

Then for all $\alpha > 0$, the inequality

$$\mu\Big(\complement(-\alpha,\alpha)\Big) \le \frac{\sqrt{e}}{\sqrt{e}-1}\Big(1 - \int_{\mathbb{R}} \hat{\mu}\Big(\frac{t}{\alpha}\Big)\, d\gamma(t)\Big) \tag{3.14}$$

holds true. If X is a real valued random variable with characteristic function φ_X, then for all $\alpha > 0$,

$$P(|X| \ge \alpha) \le \frac{\sqrt{e}}{\sqrt{e}-1}\Big(1 - \int_{\mathbb{R}} \varphi_X\Big(\frac{t}{\alpha}\Big)\, d\gamma(t)\Big). \tag{3.15}$$

3.28 Remark At first sight the inequalities (3.14), (3.15) might look strange, because the characteristic functions appearing under the integrals on the right hand sides are complex valued functions. However, since they are integrated with respect to the Gaussian measure γ, which is symmetric about the origin of $\mathbb{R}$, one actually only has the real part of the characteristic functions appearing there. Nevertheless, the author prefers to write the inequalities using the characteristic functions themselves instead of their real parts.

Proof (of corollary 3.27) We use the function f_α defined in (3.11) above, and remark that $K = f_\alpha(\alpha)^{-1}$ is given by

$$K = \frac{\sqrt{e}}{\sqrt{e}-1}.$$

With Chebyshev's inequality (3.7) we obtain (note that $f_\alpha(|x|) = f_\alpha(x)$, $x \in \mathbb{R}$)

$$\begin{aligned}\mu\Big(\complement(-\alpha,\alpha)\Big) &\le K \int_{\mathbb{R}} f_\alpha(x)\, d\mu(x)\\ &= K\Big(1 - \int_{\mathbb{R}} \exp\Big(-\frac{x^2}{2\alpha^2}\Big)\, d\mu(x)\Big).\end{aligned}$$

Now write (*exercise!*)

$$\exp\Big(-\frac{x^2}{2\alpha^2}\Big) = \int_{\mathbb{R}} e^{itx/\alpha}\, d\gamma(t), \qquad x \in \mathbb{R}. \tag{3.16}$$

Insert this above, and use Fubini's theorem (the hypotheses of which are readily checked), so that

$$\begin{aligned}\mu\Big(\complement(-\alpha,\alpha)\Big) &\le K\Big(1 - \int_{\mathbb{R}}\int_{\mathbb{R}} e^{itx/\alpha}\, d\mu(x)\, d\gamma(t)\Big)\\ &= K\Big(1 - \int_{\mathbb{R}} \hat{\mu}\Big(\frac{t}{\alpha}\Big)\, d\gamma(t)\Big).\end{aligned}$$

The second statement follows again directly by the choice $\mu = P_X$, combined with an application of the transformation theorem. $\square$

3.29 Corollary *If $(X_n, n \in \mathbb{N})$ converges in $\mathcal{L}^p(P)$, $p \geq 1$, to X, then it converges to X in probability.*

Proof For given $\varepsilon > 0$, apply Chebyshev's inequality in the form (3.9) to estimate $P(|X_n - X| \geq \varepsilon)$ from above by $\varepsilon^{-p} \|X_n - X\|_p^p$. $\square$

Chapter 4

Law of Large Numbers

Let X be a random variable in $\mathcal{L}^1(P)$, and consider a sequence of independent trials, in each of which one measures the value of X. That is, we actually consider an independent, identically distributed (*iid*) sequence $(X_n,\, n \in \mathbb{N})$ of random variables such that for every $n \in \mathbb{N}$ we have $P_X = P_{X_n}$. For example, such a sequence can be constructed as in section 2.5 on the product space of $(\mathbb{R}, \mathcal{B}(\mathbb{R}), P_X)$ over the index set $\mathbb{N}$. Intuitively one expects that the statistical mean value

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

should approach for large $n \in \mathbb{N}$ in some sense the mathematical mean value $E(X)$ of X. In fact, everyday experience tells that this will happen for each — at least for each "typical" — sequence of trials, i.e., for each "typical" $\omega \in \Omega$. [1] Slightly more generally one considers the question of convergence of

$$Z_n = \frac{1}{n}\sum_{k=1}^{n}\big(X_k - E(X_k)\big), \qquad n \in \mathbb{N}, \tag{4.1}$$

as n tends to infinity:

4.1 Definition Let $(X_n,\, n \in \mathbb{N})$ be a sequence of real valued, integrable random variables. It is said to be subject to the *strong (weak) law of large numbers*, if the sequence $(Z_n,\, n \in \mathbb{N})$ defined by (4.1) converges almost surely (in probability, resp.) to zero, as n tends to infinity.

[1] Note that each $\omega \in \Omega$ defines a whole sequence $(X_n(\omega),\, n \in \mathbb{N})$ of values, viz. the values of one sequence of trials.

4.1 Weak Law of Large Numbers

4.2 Lemma *Suppose that* $(X_n,\ n \in \mathbb{N})$ *is a sequence of real valued random variables in* $\mathcal{L}^2(P)$ *which are pairwise uncorrelated, and such that*

$$\lim_{n\to\infty} \frac{1}{n^2} \sum_{k=1}^{n} V(X_k) = 0.$$

Then $(Z_n,\ n \in \mathbb{N})$ *defined by* (4.1) *converges in* $\mathcal{L}^2(P)$ *to zero, as n tends to infinity.*

Proof Z_n has expectation zero, and therefore we get with Bienaymé's theorem, theorem 2.16,

$$\begin{aligned}
\|Z_n\|_2^2 &= V(Z_n) \\
&= \frac{1}{n^2} V\Big(\sum_{k=1}^{n} \big(X_k - E(X_k)\big)\Big) \\
&= \frac{1}{n^2} V\Big(\sum_{k=1}^{n} X_k\Big) \\
&= \frac{1}{n^2} \sum_{k=1}^{n} V(X_k).
\end{aligned}$$

By hypothesis, the last term converges to zero as $n \to \infty$. □

An application of corollary 3.29 and of Chebyshev's inequality in the form (3.9) with $p = 2$ provides the

4.3 Theorem (Khintchine) *Suppose that* $(X_n,\ n \in \mathbb{N})$ *is as in lemma 4.2. Then it is subject to the weak law of large numbers. In particular, if* $(X_n,\ n \in \mathbb{N})$ *is a sequence of pairwise uncorrelated, identically distributed random variables in* $\mathcal{L}^2(P)$*, then for every* $\varepsilon > 0$

$$P\Big(\Big|\frac{1}{n}\sum_{k=1}^{n} X_k - E(X_1)\Big| \geq \varepsilon\Big) \leq \frac{1}{\varepsilon^2 n} V(X_1), \tag{4.2}$$

and therefore is subject to the weak law of large numbers.

From a heuristic point of view, one cannot be satisfied with the law of large numbers in its weak form: It only makes a statement about the probability of deviation from zero of Z_n in sequences (*plural!*) of experiments, but *not* — as we would expect — in *one* sequence of trials. In fact, this would not be disturbing if convergence in probability would be equivalent to a.s. convergence. But his is not the case, as the following example shows:

4.4 Example Let $(X_n,\ n \in \mathbb{N})$ be an independent sequence of random variables, such that X_n, $n \in \mathbb{N}$, only assumes the values $\pm n$ and 0, namely

$$P(X_n = \pm n) = \frac{1}{2}\,\frac{1}{n\log(n+1)}, \quad P(X_n = 0) = 1 - \frac{1}{n\log(n+1)}.$$

Obviously we have $E(X_n) = 0$ for all $n \in \mathbb{N}$, and

$$V(X_n) = \frac{n}{\log(n+1)}.$$

An elementary calculation shows that $x \mapsto x/\log(x+1)$ is monotone increasing on $(0, +\infty)$. Thus we get

$$\frac{1}{n^2}\sum_{k=1}^{n} V(X_k) \le \frac{1}{n^2}\, n \frac{n}{\log(n+1)} = \frac{1}{\log(n+1)},$$

which tends to zero with $n \to \infty$. Therefore, Khintchine's theorem, theorem 4.3, entails that $(X_n,\ n \in \mathbb{N})$ is subject to the weak law of large numbers.

Next we show that $(X_n,\ n \in \mathbb{N})$ is *not* subject to the strong law of large numbers. First we show that $\sum_n P(|X_n| \ge n) = +\infty$. Indeed, $P(|X_n| \ge n) = (n\log(n+1))^{-1}$, and

$$\begin{aligned}
\int_1^\infty \frac{1}{x\log(x+1)}\,dx &\ge \int_1^\infty \frac{1}{(x+1)\log(x+1)}\,dx \\
&= \int_2^\infty \frac{1}{x\log(x)}\,dx \\
&= \log\bigl(\log(x)\bigr)\Big|_2^\infty \\
&= +\infty.
\end{aligned}$$

The Borel–Cantelli–lemma, lemma 3.18, implies that with probability 1, for infinitely many $n \in \mathbb{N}$ we have $|X_n| \ge n$. For such $n \in \mathbb{N}$, $n \ge 2$, we find

$$\begin{aligned}
\Bigl|\frac{1}{n}\sum_{k=1}^{n} X_k - \frac{1}{n-1}\sum_{k=1}^{n-1} X_k\Bigr| &= \Bigl|\frac{1}{n} X_n - \Bigl(\frac{1}{n-1} - \frac{1}{n}\Bigr)\sum_{k=1}^{n-1} X_k\Bigr| \\
&= \Bigl|\frac{1}{n} X_n - \frac{1}{n(n-1)}\sum_{k=1}^{n-1} X_k\Bigr| \\
&\ge \frac{1}{n}|X_n| - \frac{1}{n(n-1)}\sum_{k=1}^{n-1} |X_k|,
\end{aligned}$$

and since we consider $n \in \mathbb{N}$ such that $|X_n| \geq n$, we get

$$\begin{aligned}\Big|\frac{1}{n}\sum_{k=1}^{n} X_k - \frac{1}{n-1}\sum_{k=1}^{n-1} X_k\Big| &\geq 1 - \frac{1}{n(n-1)}\sum_{k=1}^{n-1}|X_k| \\ &\geq 1 - \frac{1}{n(n-1)}\sum_{k=1}^{n-1} k \\ &\geq 1 - \frac{1}{n(n-1)}\frac{n(n-1)}{2} \\ &= \frac{1}{2}.\end{aligned}$$

But then the sequence $(Z_n,\, n \in \mathbb{N})$ defined by (4.1) for this choice of X_n cannot be a.s. Cauchy. $\diamond$

Even though we just argued that for the description of sequences of independent trials one cannot be satisfied with the weak law of large numbers, it has interesting applications, and we end this section with an application found by S. N. Bernstein. Namely, we shall use Khintchine's theorem to prove the well-known theorem of Weierstraß about the uniform approximation of continuous functions by polynomials on compact intervals. Without loss of generality we only consider the interval $[0, 1]$.

Let f be a continuous function on $[0, 1]$. For $n \in \mathbb{N}$ define its *n–th Bernstein polynomial* B_n^f by

$$B_n^f(p) = \sum_{k=0}^{n} \binom{n}{k} f\Big(\frac{k}{n}\Big) p^k (1-p)^{n-k}, \qquad p \in [0, 1].$$

Now let $(X_n,\, n \in \mathbb{N})$ be an independent sequence of Bernoulli random variables with parameter $p \in [0, 1]$. Note that $Y_n = X_1 + \cdots + X_n$, $n \in \mathbb{N}$, is a binomial random variable with parameters n and p. Thus

$$E\Big(f\big(n^{-1}Y_n\big)\Big) = B_n^f(p).$$

Estimate (4.2) in theorem 4.3, tells us that for every choice of $\delta > 0$,

$$P\big(|n^{-1}Y_n - p| \geq \delta\big) \leq \frac{1}{\delta^2 n} p(1-p) \leq \frac{1}{4\delta^2 n}.$$

f being continuous on $[0, 1]$ it is uniformly continuous. Thus for given $\varepsilon > 0$ there exists $\delta > 0$ so that $|f(x) - f(y)| < \varepsilon/2$ whenever $|x - y| < \delta$. Then, for these choices of ε, δ, and any $p \in [0, 1]$,

$$\begin{aligned}
|B_n^f(p) - f(p)| &= \Big|E\big(f(n^{-1}Y_n) - f(p)\big)\Big| \\
&= \Big|\int_{[0,1]} \big(f(x) - f(p)\big)\, dP_{n^{-1}Y_n}(x)\Big| \\
&= \Big|\int_{|x-p|<\delta} \big(f(x) - f(p)\big)\, dP_{n^{-1}Y_n}(x)\Big| \\
&\qquad + \Big|\int_{|x-p|\geq\delta} \big(f(x) - f(p)\big)\, dP_{n^{-1}Y_n}(x)\Big| \\
&\leq \frac{\varepsilon}{2} + 2 \max_{x\in[0,1]} |f(x)|\, P\big(|n^{-1}Y_n - p| \geq \delta\big) \\
&\leq \frac{\varepsilon}{2} + \max_{x\in[0,1]} |f(x)|\, \frac{1}{2\delta^2 n}
\end{aligned}$$

Choosing n large enough (independently of p!), we can make the second term smaller than $\varepsilon/2$, so that we obtain the following result:

4.5 Theorem (Weierstraß–Bernstein) *For any continuous function f on $[0, 1]$, the sequence $(B_n^f, n \in \mathbb{N})$ of Bernstein polynomials of f converges uniformly on $[0, 1]$ to f.*

4.2 Strong Law of Large Numbers

In this section we prove the main result of this chapter, the strong law of large numbers in the form it has been given 1981 by N. Etemadi. The strong law of large numbers of Kolmogorov will then be a special case.

Before we move towards Etemadi's theorem, we shortly discuss the question of validity of the strong law of large numbers for an independent sequence $(X_n, n \in \mathbb{N})$ of integrable random variables from the point of view of Kolmogorov's 0–1–law, theorem 2.22. For $n \in \mathbb{N}$ set $\mathcal{T}_m = \sigma(X_n, n \geq m)$, and $\mathcal{T} = \cap_m \mathcal{T}_m$, so that theorem 2.22 states that $A \in \mathcal{T}$ has either probability one or zero. Define the sequence $(Z_n, n \in \mathbb{N})$ as in (4.1). Consider the event

$$A = \{\lim_n Z_n = 0\}.$$

(That A indeed belongs to $\mathcal{A}$ follows from theorem A.41.(a).) For every $\omega \in \Omega$ and every $m \in \mathbb{N}$ we have $\lim_n n^{-1} \sum_{k=1}^{m-1} X_k(\omega) = 0$. Therefore, for every $m \in \mathbb{N}$, A is equal to the event

$$\Big\{\lim_n \frac{1}{n} \sum_{k=m}^{n} \big(X_k - E(X_k)\big) = 0\Big\},$$

which belongs to $\mathcal{T}_m$. In other words, for every $m \in \mathbb{N}$ the event A belongs to $\mathcal{T}_m$, and therefore it belongs to the σ–algebra of terminal events $\mathcal{T}$. Kolmogorov's 0–1–law implies $P(A) = 1$ or $P(A) = 0$. Thus either with probability one $(Z_n, n \in \mathbb{N})$ converges to zero, and $(X_n, n \in \mathbb{N})$ is subject to the strong law of large numbers, or with probability one the sequence $(Z_n, n \in \mathbb{N})$ does not converge to zero.

From now on we assume that $(X_n, n \in \mathbb{N})$ is a sequence of real valued, integrable and identically distributed random variables which are pairwise independent.

Consider the sequences $(X_n^\pm, n \in \mathbb{N})$ positive and negative parts of the X_n, $n \in \mathbb{N}$. Clearly, they have the same properties as the sequence $(X_n, n \in \mathbb{N})$ (for the pairwise independence see theorem 2.25). Moreover, if $(X_n^\pm, n \in \mathbb{N})$ are subject to the strong law of large numbers then so is $(X_n, n \in \mathbb{N})$. Therefore from now on we may and will assume without loss of generality that in addition to the assumptions above we have that $X_n \geq 0$ for all $n \in \mathbb{N}$.

For the next steps we fix an $\alpha > 1$, and set

$$k_n = \lfloor \alpha^n \rfloor, \qquad n \in \mathbb{N}, \tag{4.3}$$

where for $x \in \mathbb{R}$, $\lfloor x \rfloor$ is the greatest integer less or equal to x.

4.6 Lemma *For each $\alpha > 1$ there exists $c_\alpha \in (0, 1)$ such that*

$$c_\alpha \alpha^n \leq \alpha^n - 1 < k_n \leq \alpha^n, \qquad n \in \mathbb{N}, \tag{4.4}$$

holds true.

Proof We only have to choose $c_\alpha = 1 - 1/\alpha$, then for $n \in \mathbb{N}$

$$c_\alpha \alpha^n = \alpha^n - \alpha^{n-1} \leq \alpha^n - 1,$$

because $\alpha^{n-1} \geq 1$. The rest follows by definition of k_n, $n \in \mathbb{N}$. □

For later purposes we prove now the following somewhat technical result:

4.7 Lemma *For every $\alpha > 1$ there exists a constant $K_\alpha > 0$ so that for every $m \in \mathbb{N}$*

$$\sum_{n \in \mathbb{N},\, k_n \geq m} \frac{1}{k_n^2} \leq K_\alpha \frac{1}{m^2}. \tag{4.5}$$

holds.

Proof Let $m \in \mathbb{N}$ be given, and let N_m denote the smallest natural number so that $k_{N_m} \geq m$. Then we can estimate with (4.4) as follows:

$$\begin{aligned}
\sum_{n \in \mathbb{N},\, k_n \geq m} \frac{1}{k_n^2} &= \sum_{n=N_m}^{\infty} \frac{1}{k_n^2} \\
&\leq \frac{1}{c_\alpha^2} \sum_{n=N_m}^{\infty} \frac{1}{\alpha^{2n}} \\
&= \frac{1}{c_\alpha^2} \frac{\alpha^2}{\alpha^2 - 1} \frac{1}{\alpha^{2N_m}} \\
&= K_\alpha \frac{1}{\alpha^{2N_m}},
\end{aligned}$$

where we have set $K_\alpha = c_\alpha^{-2}(\alpha^2 - 1)^{-1}\alpha^2$. Next we apply again (4.4) to estimate $\alpha_{N_m} \geq k_{N_m} \geq m$, where we used the definition of N_m. Thus we arrive at the claimed inequality (4.5). □

The first step in our strategy to prove the strong law of large numbers for the sequence $(X_n,\ n \in \mathbb{N})$, is to cut-off the random variables at large values, so that they become square integrable, and to consider subsequences of the resulting random variables indexed by k_n. Then one shows a strong law of large numbers for these subsequences of truncated random variables. For $n \in \mathbb{N}$ define

$$Y_n = X_n\, 1_{\{X_n < n\}}, \tag{4.6}$$

$$T_n = \sum_{k=1}^{n} \big(Y_k - E(Y_k)\big). \tag{4.7}$$

4.8 Lemma *For every* $\alpha > 0$*,*

$$\frac{1}{k_n}\, T_{k_n} \xrightarrow{a.s.} o.$$

Proof Our aim is to show that for every $\varepsilon > 0$,

$$\sum_{n=1}^{\infty} P\Big(\frac{1}{k_n}\,\big|T_{k_n}\big| > \varepsilon\Big) < +\infty. \tag{4.8}$$

Since then we can apply the Borel–Cantelli–lemma 3.18, and it implies that

$$P\Big(\limsup_n \Big\{\frac{1}{k_n}\,\big|T_{k_n}\big| > \varepsilon\Big\}\Big) = 0.$$

With (c) in lemma 3.6, the statement of the present lemma follows.

First we remark that the sequence $(Y_n,\ n \in \mathbb{N})$ is pairwise independent, because we may write $Y_n = \iota_n \circ X_n$, where $\iota(x) = x$ on $[0, n)$, and $\iota(x) = 0$ on $[n, +\infty)$, and the pairwise independence follows with theorem 2.25. Moreover, since $0 \leq Y_n \leq n$, $n \in \mathbb{N}$, all moments of Y_n exist, and in particular it belongs to $\mathscr{L}^2(P)$. Let $\varepsilon > 0$, and use Chebyshev's inequality in the form (3.9) with $p = 2$ to estimate for $k \in \mathbb{N}$ as follows:

$$\begin{aligned}
P\Big(\frac{1}{k}\big|T_k\big| > \varepsilon\Big) &\leq \frac{1}{\varepsilon^2}\, V\Big(\frac{1}{k}\, T_k\Big) \\
&= \frac{1}{k^2\varepsilon^2}\, V\Big(\sum_{m=1}^{k}\big(Y_m - E(Y_m)\big)\Big) \\
&= \frac{1}{k^2\varepsilon^2} \sum_{m=1}^{k} V(Y_m) \\
&\leq \frac{1}{k^2\varepsilon^2} \sum_{m=1}^{k} E(Y_m^2),
\end{aligned}$$

where we used Bienaymé's theorem, theorem 2.16, in the third step. Thus we obtain

$$\sum_{n=1}^{\infty} P\Big(\frac{1}{k_n}|T_{k_n}| > \varepsilon\Big) \le \frac{1}{\varepsilon^2}\sum_{n=1}^{\infty}\frac{1}{k_n^2}\sum_{m=1}^{k_n} E(Y_m^2).$$

We want to interchange the last two summations. This can be done as follows:

$$\begin{aligned}\sum_{n=1}^{\infty}\frac{1}{k_n^2}\sum_{m=1}^{k_n} E(Y_m^2) &= \sum_{n=1}^{\infty}\sum_{m=1}^{\infty}\frac{1}{k_n^2}\,1_{\{m\le k_n\}}\,E(Y_m^2)\\ &= \sum_{m=1}^{\infty}\sum_{n=1}^{\infty}\frac{1}{k_n^2}\,1_{\{m\le k_n\}}\,E(Y_m^2)\\ &= \sum_{m=1}^{\infty}\sum_{n\in\mathbb{N},\,k_n\ge m}\frac{1}{k_n^2}\,E(Y_m^2),\end{aligned}$$

where the interchange of the two sums in the second step is readily justified, because all terms are positive (e.g., one can apply the theorem of Fubini–Tonelli A.71). Hence we get with lemma 4.7

$$\begin{aligned}\sum_{n=1}^{\infty} P\Big(\frac{1}{k_n}|T_{k_n}| > \varepsilon\Big) &\le \frac{1}{\varepsilon^2}\sum_{m=1}^{\infty}\sum_{n\in\mathbb{N},\,k_n\ge m}\frac{1}{k_n^2}\,E(Y_m^2)\\ &\le K_\alpha\,\frac{1}{\varepsilon^2}\sum_{m=1}^{\infty}\frac{1}{m^2}\,E(Y_m^2).\end{aligned}$$

With the definition of Y_m write

$$E(Y_m^2) = \int_{\{x<m\}} x^2\,dP_{X_1}(x) = \sum_{l=1}^{m}\int_{[l-1,l)} x^2\,dP_{X_1}(x).$$

Insert this above:

$$\begin{aligned}\sum_{n=1}^{\infty} P\Big(\frac{1}{k_n}|T_{k_n}| > \varepsilon\Big) &\le K_\alpha\,\frac{1}{\varepsilon^2}\sum_{m=1}^{\infty}\frac{1}{m^2}\sum_{l=1}^{m}\int_{[l-1,l)} x^2\,dP_{X_1}(x)\\ &= K_\alpha\,\frac{1}{\varepsilon^2}\sum_{l=1}^{\infty}\Big(\sum_{m=l}^{\infty}\frac{1}{m^2}\Big)\int_{[l-1,l)} x^2\,dP_{X_1}(x),\end{aligned}$$

where we used the same trick as above to interchange the two sums. The series in brackets can be estimated in the following way

$$\sum_{m=l}^{\infty}\frac{1}{m^2} = \frac{1}{l^2} + \sum_{m=l+1}^{\infty}\frac{1}{m^2} \le \frac{1}{l^2} + \int_l^{\infty}\frac{1}{u^2}\,du = \frac{1}{l^2} + \frac{1}{l} \le \frac{2}{l}.$$

Thus we arrive at

$$\begin{aligned}\sum_{n=1}^{\infty} P\Big(\frac{1}{k_n}\big|T_{k_n}\big| > \varepsilon\Big) &\leq 2K_\alpha \,\frac{1}{\varepsilon^2} \sum_{l=1}^{\infty} \frac{1}{l} \int_{[l-1,l)} x^2 \, dP_{X_1}(x) \\ &\leq 2K_\alpha \,\frac{1}{\varepsilon^2} \sum_{l=1}^{\infty} \int_{[l-1,l)} x \, dP_{X_1}(x) \\ &= 2K_\alpha \,\frac{1}{\varepsilon^2} \int_0^{\infty} x \, dP_{X_1}(x) \\ &= 2K_\alpha \,\frac{1}{\varepsilon^2}\, E(X_1) \\ &< +\infty,\end{aligned}$$

and our claim (4.8) is shown, and therefore the lemma is proved. □

For the proof of the next lemma, we first recall an exercise from analysis (with solution):

4.9 Exercise Suppose that $(a_n,\, n \in \mathbb{N})$ is a sequence of real or complex numbers converging to a. Then its sequence of arithmetic means

$$\frac{1}{n} \sum_{k=1}^{n} a_k, \qquad n \in \mathbb{N},$$

converges to a, too.

Solution Since $(a_n,\, n \in \mathbb{N})$ is a convergent sequence, it is bounded. Thus $\overline{a} = \sup_n |a_n| < +\infty$. Given $\varepsilon > 0$ choose $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$ we have $|a_n - a| < \varepsilon$. For $n \geq n_0$ we then also have

$$\begin{aligned}\Big|\frac{1}{n} \sum_{k=1}^{n} a_k - a\Big| &\leq \frac{1}{n} \sum_{k=1}^{n_0-1} |a_k - a| + \frac{1}{n} \sum_{k=n_0}^{n} |a_k - a| \\ &\leq \frac{n_0 - 1}{n}\,(\overline{a} + |a|) + \frac{n - n_0 + 1}{n}\,\varepsilon \\ &\leq \frac{n_0 - 1}{n}\,(\overline{a} + |a|) + \varepsilon.\end{aligned}$$

Therefore

$$\limsup_n \Big|\frac{1}{n} \sum_{k=1}^{n} a_k - a\Big| \leq \varepsilon,$$

and the exercise is done.

4.10 Lemma *For the sequence* $(Y_n,\, n \in \mathbb{N})$ *defined by* (4.6)

$$\frac{1}{k_n} \sum_{m=1}^{k_n} Y_m \xrightarrow{a.s.} E(X_1) \tag{4.9}$$

holds true.

Proof By construction we have

$$\begin{aligned}\lim_{m\to\infty} E(Y_m) &= \lim_{m\to\infty}\int_{[0,m)} x\,dP_{X_1}(x)\\ &= \int_{\mathbb{R}_+} x\,dP_{X_1}(x)\\ &= E(X_1),\end{aligned}$$

where we used the monotone convergence theorem, theorem A.54, for the second equality. With exercise 4.9 we conclude that

$$\frac{1}{n}\sum_{m=1}^{n} E(Y_m) \to E(X_1).$$

But then this convergence also holds true for the subsequence indexed by $(k_n,\, n \in \mathbb{N})$. Hence we get from lemma 4.8

$$\frac{1}{k_n}\sum_{m=1}^{k_n} Y_m \xrightarrow{\text{a.s.}} E(X_1),$$

as claimed. □

4.11 Lemma *For every* $\alpha > 1$ *and the sequence* $(k_n,\, n \in \mathbb{N})$ *defined in* (4.3)

$$\frac{1}{k_n}\sum_{m=1}^{k_n} X_m \xrightarrow{a.s.} E(X_1)$$

holds.

Proof For $n \in \mathbb{N}$ we get

$$P(X_n \neq Y_n) = P(X_n \geq n) = \sum_{l=n}^{\infty} P_{X_1}\big([l, l+1)\big).$$

Therefore

$$\begin{aligned}\sum_{n=1}^{\infty} P(X_n \neq Y_n) &= \sum_{n=1}^{\infty}\sum_{l=n}^{\infty} P_{X_1}\big([l, l+1)\big)\\ &= \sum_{l=1}^{\infty}\sum_{n=1}^{l} P_{X_1}\big([l, l+1)\big),\end{aligned}$$

with the same trick we used already twice in the proof of lemma 4.8. Thus

$$\begin{aligned}\sum_{n=1}^{\infty} P(X_n \neq Y_n) &= \sum_{l=1}^{\infty} l P_{X_1}\big([l, l+1)\big) \\ &= \sum_{l=1}^{\infty} \int_{[l,l+1)} l \, dP_{X_1}(x) \\ &\le \sum_{l=1}^{\infty} \int_{[l,l+1)} x \, dP_{X_1}(x) \\ &\le E(X_1).\end{aligned}$$

Therefore we have shown that $\sum_n P(X_n \neq Y_n) < +\infty$. The Borel–Cantelli–lemma, lemma 3.18, implies that $P(\limsup_n \{X_n \neq Y_n\}) = 0$. Let $\omega \in \complement \limsup_n \{X_n \neq Y_n\}$. Then there exists $n_0(\omega) \in \mathbb{N}$ so that for all $n \ge n_0(\omega)$, $X_n(\omega) = Y_n(\omega)$. Let N be the null set for the convergence in lemma 4.10. Then for ω in the set $\complement N \cap \complement \limsup_n \{X_n \neq Y_n\}$, which is a set of probability one, we get

$$\lim_{n\to\infty} \frac{1}{k_n} \sum_{m=1}^{k_n} X_m(\omega) = E(X_1),$$

as claimed. □

4.12 Theorem (Etemadi) *Suppose that $(X_n, n \in \mathbb{N})$ is a sequence of real valued, integrable, identically distributed random variables which are pairwise independent. Then $(X_n, n \in \mathbb{N})$ is subject to the strong law of large numbers.*

Proof As argued above, we may and will assume that $X_n \ge 0$, $n \in \mathbb{N}$. Set $S_n = \sum_{k=1}^{n} X_k$, and note that $(S_n, n \in \mathbb{N})$ is monotone increasing. Let $\alpha > 1$, and consider again the sequence $(k_n, n \in \mathbb{N})$ defined in (4.3). If $m > k_1 = \lfloor \alpha \rfloor$, then there exists precisely one $n \in \mathbb{N}$ so that $k_n < m \le k_{n+1}$, and we have

$$S_{k_n} \le S_m \le S_{k_{n+1}}. \tag{4.10}$$

By construction of k_n we get

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

Hence one finds

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

and

$$\frac{k_n}{m} > \frac{\alpha^n - 1}{\alpha^{n+1}} = \frac{1}{\alpha}\Big(1 - \frac{1}{\alpha^n}\Big).$$

The sequence $(1-\alpha^{-n},\ n \in \mathbb{N})$ increases to 1, so that for any $\varepsilon > 0$ there exists $n_\varepsilon \in \mathbb{N}$ so that for all $n \geq n_\varepsilon$,

$$1 - \frac{1}{\alpha^n} \geq \frac{1}{1+\varepsilon}.$$

Choose $\varepsilon = \alpha - 1$, and write n_α instead of n_ε. Then we have for all $n \geq n_\alpha$, and every $m \in \mathbb{N}$ with $k_n < m \leq k_{n+1}$ the inequality $m \leq \alpha^2 k_n$. Thus for such n and m inequality (4.10) gives

$$\frac{1}{\alpha^2}\frac{S_{k_n}}{k_n} \leq \frac{S_{k_n}}{m} \leq \frac{S_m}{m} \leq \frac{S_{k_{n+1}}}{m} \leq \alpha \frac{S_{k_{n+1}}}{k_{n+1}}.$$

From lemma 4.11 we know that there exists a null set N_α so that for all $\omega \in \complement N_\alpha$,

$$\lim_{n\to\infty} \frac{S_{k_n}(\omega)}{k_n} = E(X_1).$$

For $\omega \in \complement N_\alpha$ choose $n_\alpha(\omega) \geq n_\alpha \in \mathbb{N}$ such that for all $n \geq n_\alpha(\omega)$

$$\frac{1}{\alpha}\, E(X_1) \leq \frac{S_{k_n}(\omega)}{k_n} \leq \alpha E(X_1).$$

Thus for all $m > k_{n_\alpha(\omega)}$ we can find $n \in \mathbb{N}$ with $k_n < m \leq k_{n+1}$, and such that

$$\frac{1}{\alpha^3}\, E(X_1) \leq \frac{1}{\alpha^2}\,\frac{S_{k_n}(\omega)}{k_n} \leq \frac{S_m(\omega)}{m} \leq \alpha \frac{S_{k_{n+1}}(\omega)}{k_{n+1}} \leq \alpha^2 E(X_1).$$

Therefore, for all $m \in \mathbb{N}$ with $m \geq m(\alpha,\omega) \equiv k_{n_\alpha(\omega)} + 1$ we obtain the following inequalities

$$\Big(\frac{1}{\alpha^3} - 1\Big)E(X_1) \leq \frac{S_m(\omega)}{m} - E(X_1) \leq \big(\alpha^2 - 1\big)E(X_1). \tag{4.11}$$

Now we let α decrease to 1: Set $\alpha = 1 + 1/p$, $p \in \mathbb{N}$, and write $m_p(\omega)$ for $m(1 + 1/p, \omega)$. Define the null set

$$N = \bigcup_{p\in\mathbb{N}} N_{1+\frac{1}{p}},$$

and let $\omega \in \complement N$. Then $\omega \in \complement N_{1+1/p}$ for every $p \in \mathbb{N}$. Given $\varepsilon > 0$ choose $p \in \mathbb{N}$ such that

$$\max\Big(\big(1 - (1+p^{-1})^{-3}\big)E(X_1), \big((1+p^{-1})^2 - 1\big)E(X_1)\Big) \leq \varepsilon.$$

Since $\omega \in \complement N_{1+1/p}$, there exists $m_p(\omega)$ so that for all $m \geq m_p(\omega)$,

$$\Big|\frac{S_m(\omega)}{m} - E(X_1)\Big| \leq \varepsilon.$$

Thus $(S_m/m,\ m \in \mathbb{N})$ converges a.s. to $E(X_1)$, and therefore the sequence $(X_n,\ n \in \mathbb{N})$ is subject to the strong law of large numbers. $\square$

Every independent sequence $(X_n,\, n \in \mathbb{N})$ or random variables trivially is pairwise independent, so that we obtain the following classical theorem as a corollary:

4.13 Theorem (Kolmogorov) *Every independent sequence of integrable, identically distributed random variables is subject to the strong law of large numbers.*

4.3 The Theorem of Cramér–Chernoff

We end this chapter with a theorem due to Cramér and Chernoff which describes the speed of convergence of $n^{-1}S_n$ to $E(X_1)$ in terms of the probability that $n^{-1}S_n$ differs from $E(X_1)$. We assume throughout that $(X_n,\, n \in \mathbb{N})$ is an iid sequence of integrable random variables.

Set $\mu = E(X_1)$, and let ψ denote the moment generating function of X_1:

$$\psi(t) = E\big(e^{tX_1}\big), \qquad t \in \mathbb{R}.$$

Since the exponential function if convex, Jensen's inequality (A.45) gives the bound

$$\psi(t) \geq e^{t\mu}, \qquad t \in \mathbb{R},$$

and therefore we get for all $t \in \mathbb{R}$,

$$t\mu - \log \psi(t) \leq 0. \tag{4.12}$$

We define

$$I(x) = \sup_{t \in \mathbb{R}} \big(tx - \log \psi(t)\big), \qquad x \in \mathbb{R}. \tag{4.13}$$

For $t = 0$, $tx - \log \psi(t)$ is equal to zero, and therefore we have $I(x) \geq 0$. Thus we get from (4.12) that $I(\mu) = 0$. Let us observe that

$$I(x) = \sup_{t \geq 0} \big(tx - \log \psi(t)\big), \qquad x \geq \mu. \tag{4.14}$$

Indeed, if $t \leq 0$ and $x \geq \mu$, then $tx \leq t\mu$, and therefore $tx - \log \psi(t) \leq t\mu - \log \psi(t) \leq 0$, because of (4.12). Thus for $x \geq \mu$, the right hand side of (4.14) is equal to $I(x)$.

Before we begin the estimation which will lead to the Cramér–Chernoff–theorem, let us make another simple observation: If Z is a positive random variable, then $E(Z) \geq P(Z \geq 1)$. This is almost trivial:

$$E(Z) \geq \int_{\{Z \geq 1\}} Z \, dP \geq P(Z \geq 1).$$

As above let $S_n = \sum_{k=1}^n X_k$, and for $x \geq \mu$ consider the following estimation, valid for arbitrary $t \geq 0$:

$$\begin{aligned} P\Big(\frac{1}{n} S_n \geq x\Big) &= P\big(\exp(tS_n - ntx) \geq 1\big) \\ &\leq E\big(\exp(tS_n - ntx)\big) \\ &= e^{-ntx}\, \psi(t)^n \\ &= \exp\big(-n(tx - \log \psi(t))\big). \end{aligned}$$

Therefore

$$P\Big(\frac{1}{n} S_n \geq x\Big) \leq \inf_{t\geq 0} \exp\big(-n\big(tx - \log \psi(t)\big)\big) = \exp\Big(-n \sup_{t\geq 0}\big(tx - \log \psi(t)\big)\Big).$$

We have already remarked above, that for $x \geq \mu$ the last supremum actually is equal to $I(x)$. Thus we have proved the first of the two inequalities of following theorem, the proof of the second being completely analogous.

4.14 Theorem (Cramér–Chernoff) *Suppose that $(X_n,\, n \in \mathbb{N})$ is an iid sequence of integrable random variables. Then*

$$\begin{aligned} P\Big(\frac{1}{n} S_n \geq x\Big) &\leq e^{-nI(x)}, \qquad x \geq \mu, \\ P\Big(\frac{1}{n} S_n \leq x\Big) &\leq e^{-nI(x)}, \qquad x \leq \mu, \end{aligned} \tag{4.15}$$

hold true.

Chapter 5

Convergence in Law

In this and the next chapter we continue our study of the important notions of convergence used in probability. Here we shall investigate convergence in law, that is, the weak convergence of probability measures, and convergence of distribution functions. In the next chapter we shall prove that both are equivalent to the pointwise convergence of the associated characteristic function. The latter result will be the key to the proof of the central limit theorem for real valued random variables in its most general form.

For clarity of the presentation, throughout this chapter we only consider sequences of real valued random variables, and correspondingly sequences of probability measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, even though most of the material can discussed in a considerably more general context.

5.1 Weak Convergence, Convergence in Law

Notation If $(S, \mathcal{S})$ is a topological space, then $C_b(S)$ denotes the space of bounded, continuous (real or complex valued) functions on S, while $C_c(S)$ denotes the space of continuous (real or complex valued) functions on S with compact support.

5.1 Definition Let μ_n, $n \in \mathbb{N}$, μ be probability measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. The sequence $(\mu_n, n \in \mathbb{N})$ is said to *converge weakly* to μ, if for every $f \in C_b(\mathbb{R})$,

$$\lim_{n\to\infty} \int_{\mathbb{R}} f \, d\mu_n = \int_{\mathbb{R}} f \, d\mu \tag{5.1}$$

holds.

Notation We also denote the weak convergence of $(\mu_n, n \in \mathbb{N})$ to μ by $\mu_n \xrightarrow{\text{w}} \mu$.

5.2 Exercise

(a) Show that $(\varepsilon_{1/n}, n \in \mathbb{N})$ converges weakly to ε_0, where ε_x is the Dirac measure in $x \in \mathbb{R}$.

(b) Find B_1, $B_2 \in \mathcal{B}(\mathbb{R})$ such that $(\varepsilon_{1/n}(B_1),\ n \in \mathbb{N})$ converges to $\varepsilon_0(B_1)$, while $(\varepsilon_{1/n}(B_2),\ n \in \mathbb{N})$ does not converge to $\varepsilon_0(B_2)$.

(c) Let $\mu_n = \mathcal{N}(x_0, 1/n)$, $x_0 \in \mathbb{R}$, and show that $\mu_n \xrightarrow{\mathrm{w}} \varepsilon_{x_0}$. (*Hint:* Make a transformation of variables, and then use the dominated convergence theorem, theorem A.56.)

5.3 Definition A sequence $(X_n,\ n \in \mathbb{N})$ of real valued random variables is said to *converge in law* to a real valued random variable X, if the sequence $(P_{X_n},\ n \in \mathbb{N})$ converges weakly to P_X.

Notation If $(X_n,\ n \in \mathbb{N})$ converges in law to X we also write $X_n \xrightarrow{\mathrm{L}} X$.

Written out, $X_n \xrightarrow{\mathrm{L}} X$ means

$$\forall f \in C_b(\mathbb{R}) : \lim_{n\to\infty} \int_{\mathbb{R}} f \, dP_{X_n} = \int_{\mathbb{R}} f \, dP_X,$$

or equivalently

$$\forall f \in C_b(\mathbb{R}) : \lim_{n\to\infty} E(f \circ X_n) = E(f \circ X).$$

We turn to the question of uniqueness of weak limits. To this end, we introduce for $a, b \in \mathbb{R}$, $a \le b$, and $\varepsilon > 0$ the following function:

$$1^{\varepsilon}_{[a,b]}(x) = \begin{cases} 1, & \text{if } x \in [a,b], \\ 1 + \varepsilon^{-1}(x-a), & \text{if } x \in [a-\varepsilon, a), \\ 1 - \varepsilon^{-1}(x-b), & \text{if } x \in (b, b+\varepsilon], \\ 0, & \text{otherwise}. \end{cases} \tag{5.2}$$

Obviously, $1^{\varepsilon}_{[a,b]}$ is nothing but the indicator of the interval $[a,b]$, extended to the left and to the right by little “tents” so as to make it continuous. In particular, we have $1^{\varepsilon}_{[a,b]} \in C_b(\mathbb{R})$ for every choice of $a, b \in \mathbb{R}$, $a \le b$, and $\varepsilon > 0$.

5.4 Lemma *Let μ_1, μ_2 be two probability measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, such that for all $a, b \in \mathbb{R}$, $a \le b$, $\varepsilon > 0$,*

$$\int_{\mathbb{R}} 1^{\varepsilon}_{[a,b]} \, d\mu_1 = \int_{\mathbb{R}} 1^{\varepsilon}_{[a,b]} \, d\mu_2. \tag{5.3}$$

Then $\mu_1 = \mu_2$. In particular, if

$$\int_{\mathbb{R}} f \, d\mu_1 = \int_{\mathbb{R}} f \, d\mu_2 \tag{5.4}$$

holds true for all $f \in C_c(\mathbb{R})$, or for all $f \in C_b(\mathbb{R})$, then $\mu_1 = \mu_2$.

Proof In equation (5.3) let ε tend to zero through a sequence (e.g., $\varepsilon = 1/n$, $n \in \mathbb{N}$), then $1^{\varepsilon}_{[a,b]}$ converges pointwise to $1_{[a,b]}$, and all functions are majorized by the integrable function identically equal to one. Thus the dominated convergence theorem A.56 implies that $\mu_1([a,b]) = \mu_2([a,b])$ for all $a, b \in \mathbb{R}$, $a \leq b$. Hence μ_1 and μ_2 coincide on a $\cap$–stable generator of $\mathcal{B}(\mathbb{R})$, and therefore theorem A.21 implies that $\mu_1 = \mu_2$. □

5.5 Theorem *Weak limits of sequences of probability measures on* $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ *are unique.*

Proof If $\mu_n \xrightarrow{\text{w}} \mu_1$, and $\mu_n \xrightarrow{\text{w}} \mu_2$, then for every $f \in C_b(\mathbb{R})$, the integral of f with respect to μ_1 is equal to the integral of f with respect to μ_2. Thus by lemma 5.4 we find that $\mu_1 = \mu_2$. □

5.2 Portmanteau's Theorem

Recall that for any subset B of a topological space S, ∂B denotes the *boundary* of B, that is, $\partial B = \overline{B} \setminus B^o$, where $\overline{B}$ is the closure of B (smallest closed set containing B), and B^o is the open interior of B (largest open set contained in B).

5.6 Theorem *Suppose that* μ_n, $n \in \mathbb{N}$, μ *are probability measures on* $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. *The following statements are equivalent:*

(a) $(\mu_n, n \in \mathbb{N})$ *converges weakly to* μ.

(b) *For every bounded, uniformly continuous function* f, $\int f \, d\mu_n$ *converges to* $\int f \, d\mu$ *as* $n \to \infty$.

(c) *For every closed set* $A \subset \mathbb{R}$,

$$\limsup_{n\to\infty} \mu_n(A) \leq \mu(A).$$

(d) *For every open set* $O \subset \mathbb{R}$,

$$\liminf_{n\to\infty} \mu_n(O) \geq \mu(O).$$

(e) *For every* $B \in \mathcal{B}(\mathbb{R})$ *with* $\mu(\partial B) = 0$,

$$\lim_{n\to\infty} \mu_n(B) = \mu(B).$$

In order to keep the flow of the arguments of the proof of Portmanteau's theorem steady, we shall first derive some preparatory results. The first two actually are exercises in analysis.

Assume that (M, d) is a metric space. For $x \in M$ and $A \subset M$ define

$$d(x, A) = \inf\{d(x, z),\ z \in A\}. \tag{5.5}$$

Let us show that for every $A \subset M$, the mapping $x \mapsto d(x, A)$ is uniformly continuous: For any $x, y \in M$, $z \in A$ we get from the triangle inequality the estimate

$$d(x, A) \leq d(x, z) \leq d(x, y) + d(y, z),$$

so that

$$d(x, A) - d(y, z) \leq d(x, y).$$

Taking the supremum over $z \in A$ on the left hand side, we find that

$$d(x, A) - d(y, A) \leq d(x, y).$$

Interchanging above the roles of x and y (or — equivalently — using the symmetry of d), we obtain $|d(x, A) - d(y, A)| \leq d(x, y)$ and the uniform continuity of $x \mapsto d(x, A)$ is shown.

Next consider the following function φ on $\mathbb{R}+$:

$$\varphi(t) = \begin{cases} 1 - t, & \text{if } t \in [0, 1], \\ 0, & \text{if } t > 1. \end{cases}$$

Clearly, φ is uniformly continuous on $\mathbb{R}_+$. In what follows, we define for closed $A \subset \mathbb{R}$ and $\varepsilon > 0$ the function

$$1^\varepsilon_A(x) = \varphi\big(\varepsilon^{-1} d(x, A)\big), \tag{5.6}$$

where d is the usual metric on $\mathbb{R}$. In an *exercise* the reader will check that for the case $A = [a, b]$, this function coincides with the one in (5.2). Since the composition of two uniformly continuous functions is uniformly continuous, we have shown the first statement of

5.7 Lemma *For every $\varepsilon > 0$, 1^ε_A is uniformly continuous. Moreover, $1^\varepsilon_A(x) = 1$ for all $x \in A$, and as $\varepsilon \to 0$, 1^ε_A converges pointwise to 1_A.*

Proof $1^\varepsilon_A(x) = 1$ for $x \in A$ follows directly from the construction of 1^ε_A. For $x \notin A$ there exists $\delta > 0$ so that $d(x, A) > \delta$, because A is closed. Choose $\varepsilon < \delta$, then $\varepsilon^{-1} d(x, A) > 1$, and therefore $\varphi\big(\varepsilon^{-1} d(x, A)\big) = 0$. □

5.8 Lemma *Suppose that $(a_j,\ j \in \mathbb{N}_0)$, $(b_j,\ j \in \mathbb{N}_0)$ are two real or complex sequences. Then for every $k \in \mathbb{N}$, the following "summation by parts formula"*

$$\sum_{j=1}^{k} a_{j-1}(b_{j-1} - b_j) = a_0 b_0 - \sum_{j=1}^{k-1} (a_{j-1} - a_j) b_j - a_{k-1} b_k \tag{5.7}$$

holds true.

Proof

$$\begin{aligned}\sum_{j=1}^{k} a_{j-1}(b_{j-1}-b_j) &= \sum_{j=1}^{k} a_{j-1}b_{j-1} - \sum_{j=1}^{k} a_{j-1}b_j \\ &= a_0b_0 + \sum_{j=1}^{k-1} a_jb_j - \sum_{j=1}^{k-1} a_{j-1}b_j - a_{k-1}b_k \\ &= a_0b_0 + \sum_{j=1}^{k-1}(a_j - a_{j-1})b_j - a_{k-1}b_k. \qquad \square\end{aligned}$$

5.9 Lemma *Suppose that* μ *is a probability measure on* $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$*, and that* $A \subset \mathbb{R}$ *is closed. Define a family* $(A_\varepsilon,\ \varepsilon > 0)$ *of closed subsets of* $\mathbb{R}$ *by*

$$A_\varepsilon = \{x \in \mathbb{R},\ d(x, A) \leq \varepsilon\}. \tag{5.8}$$

Then there are at most countably many $\varepsilon > 0$ *so that* $\mu(\partial A_\varepsilon) > 0$.

Proof Observe that

$$\partial A_\varepsilon = \{x \in \mathbb{R},\ d(x, A) = \varepsilon\}.$$

Therefore, if $\varepsilon, \varepsilon > 0$, $\varepsilon \neq \varepsilon'$, then $\partial A_\varepsilon \cap \partial A_{\varepsilon'} = \emptyset$. For $m \in \mathbb{N}$ let

$$N_m = \Big\{\varepsilon > 0,\ \mu(\partial A_\varepsilon) \geq \frac{1}{m}\Big\}.$$

Then N_m has at most m elements: Suppose not so, i.e., $\varepsilon_1, \ldots, \varepsilon_r \in N_m$ with $r > m$. Then

$$1 = \mu(\mathbb{R}) \geq \mu\Big(\biguplus_{j=1}^{r} \partial A_{\varepsilon_j}\Big) = \sum_{j=1}^{r} \mu(\partial A_{\varepsilon_j}) \geq \frac{r}{m} > 1,$$

which is a contradiction. Now

$$\big\{\varepsilon > 0,\ \mu(\partial A_\varepsilon) > 0\big\} = \biguplus_{m \in \mathbb{N}} N_m,$$

and hence this set can at most have countably many elements. $\square$

Proof (of theorem 5.6)
"(a) $\Rightarrow$ (b)" This is trivial.

"(b) $\Rightarrow$ (c)" Let $A \subset \mathbb{R}$ be closed, and let $\varepsilon > 0$ be given. For every $m \in \mathbb{N}$ we have by hypothesis

$$\lim_{n\to\infty} \int 1_A^{1/m}\, d\mu_n = \int 1_A^{1/m}\, d\mu,$$

because $1_A^{1/m}$ is uniformly continuous. Set

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

Clearly, $A_m \downarrow A$ as $m \to \infty$. Thus the continuity of μ as a probability measure (cf. theorem A.23) entails $\lim_m \mu(A_m) = \mu(A)$. Consequently there exists $m_0 \in \mathbb{N}$ so that for all $m \geq m_0$ we get

$$\mu(A) \leq \mu(A_m) \leq \mu(A) + \varepsilon.$$

By construction we have $1_A \leq 1_A^{1/m} \leq 1_{A_m}$, so that for $m \geq m_0$,

$$\int 1_A^{1/m}\, d\mu \leq \int 1_{A_m}\, d\mu = \mu(A_m) \leq \mu(A) + \varepsilon.$$

On the other hand,

$$\mu_n(A) = \int 1_A\, d\mu_n \leq \int 1_A^{1/m}\, d\mu_n.$$

Hence we can estimate as follows

$$\begin{aligned}
\limsup_{n\to\infty} \mu_n(A) &\leq \limsup_{n\to\infty} \int 1_A^{1/m}\, d\mu_n \\
&= \lim_{n\to\infty} \int 1_A^{1/m}\, d\mu_n \\
&= \int 1_A^{1/m}\, d\mu \\
&\leq \mu(A) + \varepsilon.
\end{aligned}$$

$\varepsilon > 0$ was arbitrary, and therefore we have shown that (c) holds true.

"(c) $\Rightarrow$ (a)" First we prove that under the assumption of (c), we get for every $f \in C_b(\mathbb{R})$ the following inequality:

$$\limsup_{n\to\infty} \int f\, d\mu_n \leq \int f\, d\mu. \tag{5.9}$$

Temporarily we assume in addition that $0 < f < 1$. For $k \in \mathbb{N}$ set

$$F_j = f^{-1}\Big(\Big[\frac{j}{k}, 1\Big]\Big) = \Big\{x \in \mathbb{R},\ f(x) \geq \frac{j}{k}\Big\}, \qquad j = 0, 1, \dots, k.$$

Note that for every $j = 0, 1, \dots, k$, F_j is closed, and that $F_0 = \mathbb{R}$, $F_k = \emptyset$. Moreover, $(F_j,\ j = 0, \dots, k)$ is monotone decreasing. We have

$$\mathbb{R} = f^{-1}\big([0,1)\big) = \biguplus_{j=1}^{k} f^{-1}\Big(\Big[\frac{j-1}{k}, \frac{j}{k}\Big)\Big) = \biguplus_{j=1}^{k} \big(F_{j-1} \setminus F_j\big).$$

Thus we get for every probability measure ν on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, and every $k \in \mathbb{N}$ the equality

$$\int f\,d\nu = \sum_{j=1}^{k} \int_{F_{j-1}\setminus F_j} f\,d\nu.$$

On the set $F_{j-1} \setminus F_j$ the values of f are greater or equal to $(j-1)/k$ and strictly less than j/k. Hence we obtain the inequalities

$$\sum_{j=1}^{k} \frac{j-1}{k}\big(\nu(F_{j-1}) - \nu(F_j)\big) \le \int f\,d\nu \le \sum_{j=1}^{k} \frac{j}{k}\big(\nu(F_{j-1}) - \nu(F_j)\big).$$

We apply the "summation by parts formula" (5.7) to the sums on the left and on the right hand side, and get

$$\frac{1}{k}\sum_{j=1}^{k} \nu(F_j) \le \int f\,d\nu \le \frac{1}{k} + \frac{1}{k}\sum_{j=1}^{k} \nu(F_j), \tag{5.10}$$

where we used that $\nu(F_0) = 1$, $\nu(F_k) = 0$.

The validity of (c) together with the fact that F_j is closed entails for every $k \in \mathbb{N}$,

$$\limsup_{n\to\infty} \mu_n(F_j) \le \mu(F_j), \qquad j = 0, 1, \ldots, k.$$

Now we use the right hand side of (5.10) for the choice $\nu = \mu_n$ for the following estimation

$$\begin{aligned}
\limsup_{n\to\infty} \int f\,d\mu_n &\le \limsup_{n\to\infty}\Big(\frac{1}{k} + \frac{1}{k}\sum_{j=1}^{k} \mu_n(F_j)\Big) \\
&\le \frac{1}{k} + \frac{1}{k}\sum_{j=1}^{k} \limsup_{n\to\infty} \mu_n(F_j) \\
&\le \frac{1}{k} + \frac{1}{k}\sum_{j=1}^{k} \mu(F_j) \\
&\le \frac{1}{k} + \int f\,d\mu,
\end{aligned}$$

and in the last step we made use of the left hand inequality in (5.10) with the choice $\nu = \mu$. Next let k tend to infinity and obtain the claim (5.9) for the case that $0 < f < 1$. In order to remove the latter restriction, in a first step we only assume that $f > 0$. Then set $f_1 = (1 + \|f\|_\infty)^{-1} f$, where $\|f\|_\infty = \sup_{x\in\mathbb{R}} |f(x)|$. Then $0 < f_1 < 1$, so that for f_1 inequality (5.9) holds true. Multiplication of this inequality by $(1 + \|f\|_\infty)$ then gives the desired inequality for f. Secondly for $f \in C_b(\mathbb{R})$, there exists a constant $C \ge 0$ so that $f_2 = f + C > 0$. Since inequality (5.9) holds true for f_2, and $\int C\,d\nu = C$, $\nu = \mu$ or $\nu = \mu_n$, we just have to subtract on both sides of the inequality for f_2 the value C, in order to get (5.9) in general.

By replacing $f \in C_b(\mathbb{R})$ by $-f$ in inequality (5.9), we find

$$\liminf_{n\to\infty} \int f\,d\mu_n \geq \int f\,d\mu, \tag{5.11}$$

and both inequalities, (5.9) and (5.11), together yield

$$\lim_{n\to\infty} \int f\,d\mu_n = \int f\,d\mu,$$

for every $f \in C_b(\mathbb{R})$. Therefore $\mu_n \xrightarrow{\text{w}} \mu$, and hence (a) holds true.

"(c) $\Leftrightarrow$ (d)" This is obvious by taking complements.

"(c) $\Rightarrow$ (e)" Assume that (c) is true, and that $A \in \mathcal{B}(\mathbb{R})$. Then

$$\begin{aligned}
\mu(\overline{A}) &\geq \limsup_{n\to\infty} \mu_n(\overline{A}) \\
&\geq \limsup_{n\to\infty} \mu_n(A) \\
&\geq \liminf_{n\to\infty} \mu_n(A) \\
&\geq \liminf_{n\to\infty} \mu_n(A^o) \\
&\geq \mu(A^o),
\end{aligned}$$

and we used (d) in the last step. Since $\partial A = \overline{A} \setminus A^o$, the assumption $\mu(\partial A) = 0$ implies $\mu(\overline{A}) = \mu(A^o) = \mu(A)$, and therefore also equality everywhere above. In particular we get

$$\limsup_{n\to\infty} \mu_n(A) = \liminf_{n\to\infty} \mu_n(A) = \lim_{n\to\infty} \mu_n(A) = \mu(A).$$

Hence (e) holds.

"(e) $\Rightarrow$ (c)" Suppose that (e) holds and that $A \subset \mathbb{R}$ is closed. Define a family $(A_\varepsilon,\ \varepsilon \geq 0)$ of closed subsets of $\mathbb{R}$ as in (5.8). Obviously, as $\varepsilon \downarrow 0$, A_ε decreases to A. From lemma 5.9 we know, that there are at most countably many $\varepsilon > 0$ so that $\mu(\partial A_\varepsilon) > 0$. Therefore there exists a monotone decreasing sequence $(\varepsilon_k,\ k \in \mathbb{N})$ of strictly positive numbers ε_k such that $\mu(\partial A_{\varepsilon_k}) = 0$ for all $k \in \mathbb{N}$. Then (e) implies that for every $k \in \mathbb{N}$,

$$\lim_{n\to\infty} \mu_n(A_{\varepsilon_k}) = \mu(A_{\varepsilon_k})$$

holds true. Hence we find

$$\limsup_{n\to\infty} \mu_n(A) \leq \limsup_{n\to\infty} \mu_n(A_{\varepsilon_k}) = \lim_{n\to\infty} \mu_n(A_{\varepsilon_k}) = \mu(A_{\varepsilon_k})$$

for every $k \in \mathbb{N}$. Now let k tend to infinity, so that the continuity of μ (cf. theorem A.23) gives

$$\limsup_{n\to\infty} \mu_n(A) \le \mu(A),$$

and statement (c) is shown, and the proof of theorem 5.6 is complete. □

5.10 Exercise Use Portmanteau's theorem to prove that if $(X_n,\, n \in \mathbb{N})$ converges in law to a constant K, then it also converges in probability to K.

5.11 Theorem *Suppose that* $(X_n,\, n \in \mathbb{N})$ *is a sequence of real valued random variables which converges to a real valued random variable* X *in probability. Then it converges to* X *in law.*

Proof By statement (b) in Portmanteau's theorem, theorem 5.6, it is sufficient to show that

$$\lim_{n\to\infty} \int f\, dP_{X_n} = \int f\, dP_X$$

for every $f \in C_b(\mathbb{R})$ which is uniformly continuous. Given such an f, let $\varepsilon > 0$ be given, and choose $\delta > 0$ such that for all $x,\, y \in \mathbb{R}$ with $|x - y| < \delta$ we have $|f(x) - f(y)| < \varepsilon/2$. For $n \in \mathbb{N}$ set $A_n = \{|X_n - X| \ge \delta\}$. Then

$$\begin{aligned}
\Bigl|\int f\, dP_{X_n} - \int f\, dP_X\Bigr| &= \bigl|E\bigl(f\circ X_n - f\circ X\bigr)\bigr| \\
&\le E\bigl(\bigl|f\circ X_n - f\circ X\bigr|\bigr) \\
&= \int_{A_n} |f\circ X_n - f\circ X|\, dP + \int_{\complement A_n} |f\circ X_n - f\circ X|\, dP \\
&\le 2\|f\|_\infty P(A_n) + \frac{\varepsilon}{2} P\bigl(\complement A_n\bigr) \\
&\le 2\|f\|_\infty P(A_n) + \frac{\varepsilon}{2}.
\end{aligned}$$

By hypothesis, $P(A_n) \to 0$ as $n \to \infty$, and therefore we only have to choose $n_0 \in \mathbb{N}$ large enough so that for all $n \ge n_0$, $2\|f\|_\infty P(A_n) < \varepsilon/2$, in order to get

$$\Bigl|\int f\, dP_{X_n} - \int f\, dP_X\Bigr| < \varepsilon$$

for all $n \ge n_0$, and the proof is done. □

5.3 Convergence of Distribution Functions

Recall definition 2.7 of the distribution function F_X of a real valued random variable X. Just as well we can define the *distribution function* F of a probability measure μ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ by

$$F(x) = \mu\bigl((-\infty, x]\bigr), \qquad x \in \mathbb{R}, \tag{5.12}$$

and it is clear that the properties stated in theorem 2.9 hold true for such an F.

5.12 Lemma *Let F be a distribution function associated with a probability measure μ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ as in* (5.12)*. A point $x \in \mathbb{R}$ is a point of continuity of F, if and only if $\mu(\{x\}) = 0$.*

Proof F is continuous from the right (see theorem 2.9), so that F is continuous in x, if and only if for every sequence $(x_n, n \in \mathbb{N})$ in $\mathbb{R}$, which strictly increases to x, we have

$$\lim_{n\to\infty}\big(F(x) - F(x_n)\big) = 0.$$

By definition (5.12) this means

$$\begin{aligned} 0 &= \lim_{n\to\infty}\Big(\mu\big((-\infty, x]\big) - \mu\big((-\infty, x_n]\big)\Big) \\ &= \lim_{n\to\infty}\mu\big((x_n, x]\big) \\ &= \mu\Big(\bigcap_{n\in\mathbb{N}}(x_n, x]\Big) \\ &= \mu(\{x\}), \end{aligned}$$

where we used the continuity (cf. theorem A.23) of μ. □

5.13 Exercise Prove that every distribution function can have at most countably many discontinuities. (*Hint:* Show first that any discontinuity has to be a jump. Then consider sets of the form $U_m = \{x \in \mathbb{R},\ F(x) - F(x-) \geq 1/m\}$, $m \in \mathbb{N}$, where $F(x-)$ is the limit of F at x from the left, and make an argument similar to the one we used in the proof of lemma 5.9.)

5.14 Definition A sequence $(F_n, n \in \mathbb{N})$ of distribution functions is said to *converge to a distribution function F*, if for every point of continuity $x \in \mathbb{R}$ of F, $(F_n(x), n \in \mathbb{N})$ converges to $F(x)$.

5.15 Theorem *A sequence $(\mu_n, n \in \mathbb{N})$ of probability measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ converges weakly to a probability measure μ, if and only if the associated sequence $(F_n, n \in \mathbb{N})$ of distribution functions converges to the distribution function F of μ (in the sense of definition 5.14).*

For the proof we first recall a result from analysis: Any open subset B of $\mathbb{R}$ can be written as a union of at most countably many pairwise disjoint open intervals. Indeed, set $x \sim y$ for $x, y \in B$, if there exists an open interval $I \subset B$ so that $x, y \in I$. It is not hard to check that $\sim$ defines an equivalence relation on B, and that every equivalence class is an open interval. Therefore B is decomposed by $\sim$ into open, pairwise disjoint intervals. Since every open interval contains a rational point, there can be at most countably many of them.

Proof (of theorem 5.15) Suppose that $\mu_n \xrightarrow{\mathrm{w}} \mu$, and let $x \in \mathbb{R}$ be a point of continuity of F, so that by lemma 5.12 $\mu(\{x\}) = 0$. Then statement (e) of Portmanteau's theorem yields that

$$\lim_{n\to\infty} \mu_n\big((-\infty, x]\big) = \mu\big((-\infty, x]\big),$$

because $\{x\}$ is the boundary of $(-\infty, x]$. But this means that at every continuity point of F, $(F_n,\ n \in \mathbb{N})$ converges to F.

For the converse suppose that F_n converges to F in the sense of definition 5.14. Suppose that $B \subset \mathbb{R}$ is open. As recalled above, there exists at most countably many open intervals I_m which are pairwise disjoint, and whose union is B. For the following we assume that $B = \biguplus_{m\in\mathbb{N}} I_m$ — the case where B is decomposed only into finitely many intervals is easier. For every $m \in \mathbb{N}$ we can write $I_m = (a_m, b_m)$. Fatou's lemma, theorem A.55, gives

$$\liminf_{n\to\infty} \mu_n(B) = \liminf_{n\to\infty} \sum_{m=1}^{\infty} \mu_n(I_m) \geq \sum_{m=1}^{\infty} \liminf_{n\to\infty} \mu_n(I_m).$$

Let $\varepsilon > 0$. μ being continuous, we can find for every $m \in \mathbb{N}$ $a'_m > a_m$ and $b'_m < b_m$ such that

$$\mu(I'_m) \leq \mu(I_m) \leq \mu(I'_m) + \varepsilon 2^{-m},$$

and $I'_m = (a'_m, b'_m]$. Moreover, since F has only countably many points of discontinuity, we may choose the points a'_m, b'_m, $m \in \mathbb{N}$ in such a way that they are continuity points of F. Then we can estimate in the following way:

$$\begin{aligned}
\liminf_{n\to\infty} \mu_n(B) &\geq \sum_{m=1}^{\infty} \liminf_{n\to\infty} \mu_n(I'_m) \\
&= \sum_{m=1}^{\infty} \liminf_{n\to\infty}\big(F_n(b'_m) - F_n(a'_m)\big) \\
&= \sum_{m=1}^{\infty} \big(F(b'_m) - F(a'_m)\big) \\
&= \sum_{m=1}^{\infty} \mu(I'_m) \\
&\geq \sum_{m=1}^{\infty} \big(\mu(I_m) - \varepsilon 2^{-m}\big) \\
&= \mu(B) - \varepsilon.
\end{aligned}$$

$\varepsilon > 0$ was arbitrary, so that we obtain

$$\liminf_{n\to\infty} \mu_n(B) \geq \mu(B).$$

Statement (d) of Portmanteau's theorem, theorem 5.6, implies that $(\mu_n,\ n \in \mathbb{N})$ converges weakly to μ. □

We summarize the results about the various notions of convergence of random variables proved so far in the following diagram:

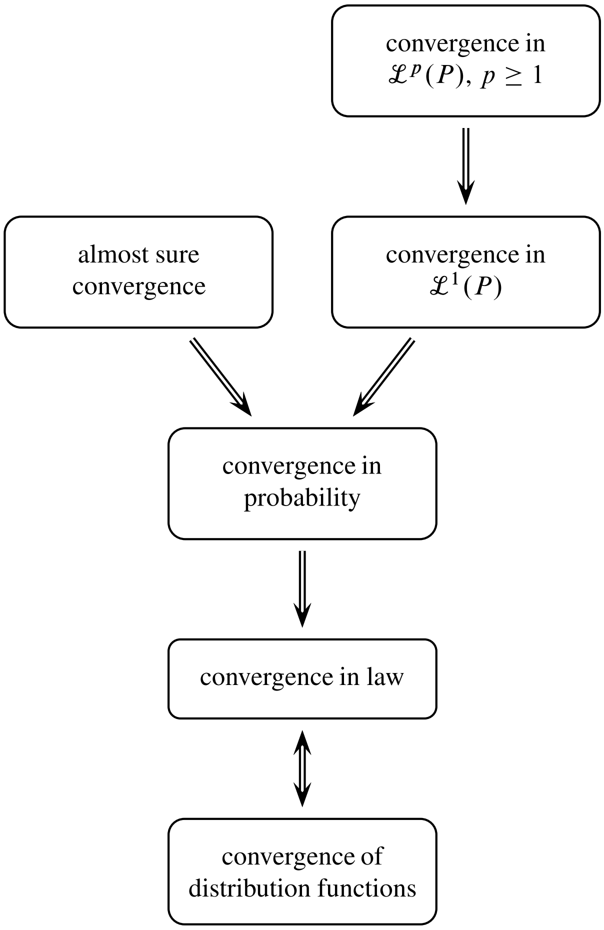

Figure 5.1: Convergence of random variables

We recall that additional to the implications in the diagram 5.1 there are the following two: (i) a sequence converging in probability has a subsequence which a.s. converges (a.s. with the same limit), and (ii) a sequence which converges in law to a constant also converges in probability to that constant.

Chapter 6

Prohorov's Theorem, Lévy's Continuity Lemma

In the present chapter we shall enter a deeper study of the weak convergence which we began in chapter 5. Our main aim is to establish the characterization of the weak convergence of a sequence of probability measures in terms of the associated sequence of characteristic functions, that is the proof of the continuity lemma by P. Lévy, theorem 6.21. It is this result which will allow for a nice proof of the central limit theorem in its most general form in the next chapter.

Throughout this chapter we shall treat families $\mathcal{M}$ of probability measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, the associated family $\mathcal{F}$ of distribution functions, and a family $\mathcal{X}$ of real valued random variables (defined on some probability space) whose laws coincide with those in $\mathcal{M}$ on the same footing. In fact, $\mathcal{M}$ and $\mathcal{F}$ are in one-to-one correspondence, $\mathcal{X}$ uniquely defines $\mathcal{M}$ and $\mathcal{F}$, and $\mathcal{M}$ or $\mathcal{F}$ define at least one family $\mathcal{X}$.

6.1 Definition A set $\mathcal{M}$ of probability measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is called *relatively compact*, if and only if every sequence in $\mathcal{M}$ has a subsequence which is weakly convergent to a probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. A set $\mathcal{F}$ of distribution functions is called *relatively compact*, if the associated set of probability measures is relatively compact, and a set $\mathcal{X}$ of real valued random variables is called *relatively compact*, if its associated set of laws is relatively compact.

6.2 Remarks Thus, by theorem 5.15 a family $\mathcal{F}$ of distribution functions is relatively compact, if and only if every sequence in $\mathcal{F}$ has a subsequence which converges in the sense of definition 5.14 to a distribution function. By the definition of convergence in law, definition 5.3, a family $\mathcal{X}$ of real valued random variables is relatively compact, if and only if every sequence in $\mathcal{X}$ has a subsequence which converges in law to a real valued random variable.

In topology one would rather say *sequentially (relatively) compact*. However, one can show that the set of all probability measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, equipped with the notion of weak convergence of sequences is actually a metric space (cf., e.g., [7, Appendix III] or [4, § 31]). Since in metric spaces sequential compactness is equivalent to compactness, the above naming is justified.

For a while we shall mainly concentrate on the investigation of relative compactness of sets $\mathscr{F}$ of distribution functions. We emphasize that the notion of convergence will always be understood as the one in definition 5.14.

Here is a rather trivial, but nevertheless important

6.3 Example Choose for $\mathscr{F}$ the set of *all* distribution functions on $\mathbb{R}$. Consider the sequence $(F_n,\, n \in \mathbb{N})$ given by

$$F_n(x) = 1_{[n,+\infty)}(x), \qquad x \in \mathbb{R}. \tag{6.1}$$

Thus, the associated sequence of probability measures is $(\varepsilon_n,\, n \in \mathbb{N})$. Obviously, the sequence $(F_n,\, n \in \mathbb{N})$ converges pointwise to the zero function. The latter not being a distribution function, we see that the set of all distribution functions is *not* relatively compact. Clearly, the "defect" is that it is allowed that "mass can vanish to infinity". Later we shall see that it is precisely this effect that one has to avoid in order to get a very powerful characterization of relative compactness, viz. Prohorov's theorem, theorem 6.9. $\diamond$

6.1 Helly's Theorem

Helly's theorem below states that the set of all distribution functions on $\mathbb{R}$ is "almost compact" (and hence any set of distribution functions is "almost relatively compact"): It is only property (c) in theorem 2.9 which characterizes distribution functions that in general fails to hold true for the limit function, that is, the effect that we had already observed in example 6.3.

6.4 Theorem (Helly) *Suppose that $(F_n,\, n \in \mathbb{N})$ is a sequence of distribution functions. Then there exists a monotone increasing function G with values in $[0,1]$, which is continuous from the right, and there exists a subsequence $(F_{n_k},\, k \in \mathbb{N})$ of $(F_n,\, n \in \mathbb{N})$, such that for every $x \in \mathbb{R}$ which is a point of continuity of G, $(F_{n_k}(x),\, k \in \mathbb{N})$ converges to $G(x)$.*

Proof Let $(q_k,\, k \in \mathbb{N})$ be an enumeration of $\mathbb{Q}$. For $n \in \mathbb{N}$ define the sequence

$$a^n = \big(F_n(q_k),\, k \in \mathbb{N}\big) \in \mathbb{R}^{\mathbb{N}}.$$

For every $k \in \mathbb{N}$ we then have the sequence $(a^n_k,\, n \in \mathbb{N})$ in $[0,1]$, and in particular, the set defined by this sequence is bounded. Statement (d) of lemma B.4, see appendix B, entails that the set in $\mathbb{R}^{\mathbb{N}}$ defined by $(a^n,\, n \in \mathbb{N})$ is relatively compact. Thus there exist a subsequence $(a^{n_l},\, l \in \mathbb{N})$, and a sequence $a = (a_k,\, k \in \mathbb{N}) \in \mathbb{R}^{\mathbb{N}}$ such that

$$\lim_{l \to \infty} \rho(a^{n_l}, a) = 0,$$

where ρ is the metric on $\mathbb{R}^{\mathbb{N}}$ defined in (B.1). Then statement (a) of lemma B.4 gives that for every $k \in \mathbb{N}$,

$$\lim_{l\to\infty} a_k^{n_l} = a_k.$$

In order to simplify the notation we shall assume from now on that the original sequence is identical with the extracted subsequence, i.e., we have $a_k^n \to a_k$, $n \to \infty$, for every $k \in \mathbb{N}$. Since for all $n, k \in \mathbb{N}$, $0 \le a_k^n \le 1$, we also find for the limit values $0 \le a_k \le 1$, $k \in \mathbb{N}$. For $k \in \mathbb{N}$ set $G_0(q_k) = a_k \in [0, 1]$, so that we obtain

$$\forall\, q \in \mathbb{Q} : F_n(q) \to G_0(q), \text{ as } n \to \infty.$$

We show that G_0 is increasing on $\mathbb{Q}$: Let $q_1, q_2 \in \mathbb{Q}$, $q_1 < q_2$, and let $\varepsilon > 0$ be given. Choose $n \in \mathbb{N}$ such that $|F_n(q_i) - G_0(q_i)| < \varepsilon/2$, $i = 1, 2$. Then

$$\begin{aligned} G_0(q_2) - G_0(q_1) &= G_0(q_2) - F_n(q_2) + F_n(q_2) - F_n(q_1) + F_n(q_1) - G_0(q_1) \\ &\ge -\varepsilon + F_n(q_2) - F_n(q_1) \\ &\ge -\varepsilon, \end{aligned}$$

because as a distribution function F_n is increasing. Since $\varepsilon > 0$ is arbitrary, we have shown our claim.

Define

$$G(x) = \inf_{q\in\mathbb{Q},\, q>x} G_0(q), \qquad x \in \mathbb{R}. \tag{6.2}$$

Since $0 \le G_0(q) \le 1$ for all $q \in \mathbb{Q}$, we get that $0 \le G(x) \le 1$ for all $x \in \mathbb{R}$. Note that in general we do not necessarily have equality of $G(q)$ and $G_0(q)$, $q \in \mathbb{Q}$.

First we show that G is monotone increasing. For $x \le y$ we have the inclusion $\{G_0(q), q > x\} \supset \{G_0(q), q > y\}$, so that we obtain

$$G(x) = \inf\{G_0(q), q > x\} \le \inf\{G_0(q), q > y\} = G(y).$$

Next we prove that G is continuous from the right. Let $x \in \mathbb{R}$ and $\varepsilon > 0$ be given. By construction of G there exists $q \in \mathbb{Q}$, $q > x$, such that

$$G_0(q) - \varepsilon \le G(x) \le G_0(q).$$

Set $\delta = q - x > 0$, and suppose that $y \in (x, x + \delta)$, viz. $x < y < q$. By construction of $G(y)$ we have $G(y) \le G_0(q)$, and since G is increasing we find

$$0 \le G(y) - G(x) \le \varepsilon,$$

proving the continuity from the right of G.

Finally we show that $(F_n, n \in \mathbb{N})$ converges to G at every point of continuity of G. Suppose that $x \in \mathbb{R}$ is such a point, and let $\varepsilon > 0$ be given. We claim that there exist $q_1, q_2 \in \mathbb{Q}$ with $q_1 < x < q_2$, and such that

$$G_0(q_2) - \varepsilon \le G(x) \le G_0(q_1) + \varepsilon \tag{6.3}$$

holds true. Indeed, the existence of $q_2 > x$, $q_2 \in \mathbb{Q}$, so that the left inequality holds follows directly from the construction of G. The continuity of G from the left x entails the existence of $x' \in \mathbb{R}$, $x' < x$, so that $G(x) \leq G(x') + \varepsilon$. Now choose a rational point $q_1 \in (x', x)$, then we get from the definition of $G(x')$ the inequality

$$G(x) \leq G(x') + \varepsilon \leq G_0(q_1) + \varepsilon,$$

and therefore (6.3) is proved. Now we can estimate as follows:

$$\begin{aligned} G(x) - \varepsilon &\leq G_0(q_1) \\ &= \lim_n F_n(q_1) \\ &\leq \liminf_n F_n(x) \\ &\leq \limsup_n F_n(x) \\ &\leq \limsup_n F_n(q_2) \\ &= G_0(q_2) \\ &\leq G(x) + \varepsilon. \end{aligned}$$

$\varepsilon > 0$ was arbitrary so that everywhere above we obtain equality, and hence

$$\liminf_n F_n(x) = \limsup_n F_n(x) = G(x),$$

which concludes the proof of theorem 6.4 □

6.2 Prohorov's Theorem

According to Helly's theorem, theorem 6.4, any set of distribution functions fails to be relatively compact by the fact that a limit function G might not have the properties $\lim_{x\to-\infty} G(x) = 0$, $\lim_{x\to\infty} G(x) = 1$, that is the effect we have already seen in example 6.3. On the level of the Lebesgue-Stieltjes measure μ_G defined by G (see example A.26), this means that μ_G can fail to be a probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, namely $\mu_G(\mathbb{R}) < 1$. In order to exclude such an effect, one makes the following definition.

6.5 Definition A set $\mathcal{M}$ of probability measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is called *tight*, if and only if for every $\varepsilon > 0$ there exists a compact set $K_\varepsilon \subset \mathbb{R}$ so that $\mu(K_\varepsilon) > 1 - \varepsilon$ for every $\mu \in \mathcal{M}$. A set $\mathcal{F}$ of distribution functions is called *tight*, if the associated set of probability measures is tight, and a set $\mathcal{X}$ of real valued random variables is called *tight*, if its associated set of laws is tight.

6.6 Remark Obviously, we can equivalently replace the inequality $\mu(K_\varepsilon) > 1 - \varepsilon$ in definition 6.5 by $\mu(K_\varepsilon) \geq 1 - \varepsilon$.

6.7 Exercise Show that any *finite* family of probability measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is tight. (*Hint:* Recall that every measure is continuous from below, theorem A.23.)

Obviously, the sequence of distribution functions in example 6.3 is not tight. [1]

6.8 Lemma *A set $\mathcal{F}$ of distribution functions on $\mathbb{R}$ is tight, if and only if for every $\varepsilon > 0$ there exists $R > 0$, so that $F(R) \geq 1 - \varepsilon$ and $F(-R) \leq \varepsilon$ hold true for every $F \in \mathcal{F}$.*

Proof This is an easy *exercise*. □

6.9 Theorem (Prohorov) *A family $\mathcal{M}$ of probability measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is relatively compact, if and only if it is tight.*

Proof Assume that $\mathcal{M}$ is relatively compact, and that $\mathcal{M}$ is not tight. Then there exists $\varepsilon > 0$ so that for every compact set $K \subset \mathbb{R}$ there exists $\mu_K \in \mathcal{M}$ with $\mu_K(K) < 1 - \varepsilon$. We choose the compact sets as $[-n, n]$, $n \in \mathbb{N}$, and denote the corresponding measure by μ_n. Thus we have a sequence $(\mu_n,\, n \in \mathbb{N})$ in $\mathcal{M}$ so that for every $n \in \mathbb{N}$, $\mu_n([-n, n]) < 1 - \varepsilon$. By hypothesis, the sequence $(\mu_n,\, n \in \mathbb{N})$ has a weakly convergent subsequence: $\mu_{n_k} \xrightarrow{\text{w}} \mu$ as $k \to \infty$, where μ is a probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. From Portmanteau's theorem 5.6, we get for every $x \in \mathbb{R}$,

$$\begin{aligned} \mu\big((-x, x)\big) &\leq \liminf_k \mu_{n_k}\big((-x, x)\big) \\ &\leq \liminf_n \mu_{n_k}([-n_k, n_k]) \\ &\leq 1 - \varepsilon, \end{aligned}$$

because for k large enough we have that $x \leq n_k$. But the statement that $\mu((-x, x)) < 1 - \varepsilon$ for all $x \in \mathbb{R}$ is in contradiction with the fact that μ is a probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$: We only have to let x increase through a sequence to infinity and use the continuity of μ to get $\mu(\mathbb{R}) \leq 1 - \varepsilon$. Thus relative compactness of $\mathcal{M}$ implies its tightness.

Now suppose that $\mathcal{M}$ is tight, and let $(\mu_n,\, n \in \mathbb{N})$ be any sequence in $\mathcal{M}$. Denote by $(F_n,\, n \in \mathbb{N})$ the sequence of distribution functions of $(\mu_n,\, n \in \mathbb{N})$, so that our assumption implies that $(F_n,\, n \in \mathbb{N})$ is tight. By theorem 5.15 it is sufficient to show that $(F_n,\, n \in \mathbb{N})$ has a subsequence which converges (in the sense of convergence of distribution functions, definition 5.14) to a distribution function. Helly's theorem 6.4 states that there exists a subsequence $(F_{n_k},\, k \in \mathbb{N})$ which converges to some function G at every continuity point of G, and G takes values in $[0, 1]$, is monotone increasing, and continuous from the right. To finish the proof we only have to show that $\lim_{x \to -\infty} G(x) = 0$, and $\lim_{x \to \infty} G(x) = 1$. Given $\varepsilon > 0$, by lemma 6.8 we can find $R > 0$ so that for all $k \in \mathbb{N}$, $F_{n_k}(-R) \leq \varepsilon$ and $F_{n_k}(R) \geq 1 - \varepsilon$.

[1] Here — and in the sequel — we follow the common practice to identify a sequence with the set defined by that sequence.

Moreover, we may assume that $-R$, R are points of continuity of G: Similarly as in exercise 5.13, G can have at most countably many discontinuities, so that if there is a discontinuity at R or $-R$ one can find a larger R so that $-R$, R are point of continuity, and at the same time the above inequalities for the F_{n_k} remain true, because F_{n_k} is monotone increasing. Since $(F_{n_k}, k \in \mathbb{N})$ converges to G at $\pm R$, we get $G(R) \geq 1 - \varepsilon$ and $G(-R) \leq \varepsilon$. Since G is monotone increasing we have shown: Given $\varepsilon > 0$ there exists $R > 0$ so that for all $x \geq R$, $1 \geq G(x) \geq 1 - \varepsilon$, $0 \leq G(-x) \leq \varepsilon$. Thus $\lim_{x\to-\infty} G(x) = 0$, and $\lim_{x\to\infty} G(x) = 1$, and the proof is complete. □

6.10 Remark Prohorov's theorem in the form given above is valid in the much more general situation of probability measures on a polish space. That is, instead of $\mathbb{R}$ (with its usual open sets) one considers a separable, metrizable space, which is complete in one of the metrics defining the topology, and equips it with its Borel σ–algebra, cf., e.g., [7].

6.3 Fourier Transform, Characteristic Functions

6.11 Definition For a probability measure μ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, the complex valued function

$$\hat{\mu}(t) = \int_{\mathbb{R}} e^{itx}\, d\mu(x), \qquad t \in \mathbb{R}, \tag{6.4}$$

on $\mathbb{R}$ is called the *characteristic function* of μ.

Thus for a real valued random variable X the characteristic function $\hat{P}_X$ of its law P_X coincides with the characteristic function φ_X of X, which we have already defined in definition 2.12. The mapping

$$\hat{\cdot} : \mu \mapsto \hat{\mu}$$

is also called the *Fourier transform* of μ. [2]

6.12 Examples The details of all examples below are left as *exercises*.

(a) For $\mu = \varepsilon_x$, $x \in \mathbb{R}$, $\hat{\mu}(t) = \exp(itx)$, $t \in \mathbb{R}$.

(b) If μ is the binomial law $\mathcal{B}(p, n)$, $p \in [0, 1]$, $n \in \mathbb{N}$, (see (2.5)), then

$$\hat{\mu}(t) = \big(1 + p(e^{it} - 1)\big)^n, \qquad t \in \mathbb{R}.$$

(c) If μ is the Poisson law $\mathcal{P}(\lambda)$, $\lambda > 0$, (see (2.6)), then

$$\hat{\mu}(t) = \exp\big(\lambda(e^{it} - 1)\big), \qquad t \in \mathbb{R}.$$

[2] We warn the reader that there are many different conventions for the Fourier transform in the literature.

(d) If μ is the normal law $\mathcal{N}(m,\sigma^2)$, $m \in \mathbb{R}$, $\sigma^2 > 0$, (see (2.10)), then

$$\hat{\mu}(t) = \exp\Big(imt - \frac{1}{2}\sigma^2 t^2\Big), \qquad t \in \mathbb{R}. \tag{6.5}$$

(e) If μ is the *Cauchy law*, that is, it has the density given by

$$\varphi(x) = \frac{1}{\pi}\,\frac{1}{1+x^2}, \qquad x \in \mathbb{R},$$

then

$$\hat{\mu}(t) = \exp(-|t|), \qquad t \in \mathbb{R}.$$

(*Hint:* Use the residue theorem of complex analysis.) ◇

6.13 Definition A complex valued function φ on $\mathbb{R}$ is *positive definite*, if for all $n \in \mathbb{N}$, $t_1, t_2, \ldots, t_n \in \mathbb{R}$, $z_1, z_2, \ldots, z_n \in \mathbb{C}$,

$$\sum_{k,l=1}^{n} z_k \overline{z_l}\, \varphi(t_k - t_l) \geq 0 \tag{6.6}$$

holds true.

6.14 Lemma *Suppose that φ is positive definite. Then $\varphi(0) \geq 0$, and $|\varphi(t)| \leq \varphi(0)$ for all $t \in \mathbb{R}$. Furthermore, $\overline{\varphi(t)} = \varphi(-t)$, $t \in \mathbb{R}$.*

Proof Choose $n = 2$, $t_1 = 0$, $t_2 = t \in \mathbb{R}$, $z_1 = 1$, $z_2 = z \in \mathbb{C}$, then equation (6.6) reads

$$\varphi(0) + \varphi(0)\,|z|^2 + \varphi(t)\,z + \varphi(-t)\,\overline{z} \geq 0. \tag{6.7}$$

If choose now $z = 1$ we get

$$2\varphi(0) + \varphi(t) + \varphi(-t) \geq 0.$$

In particular, the left hand side is real, and the choice $t = 0$ shows that $\varphi(0) \geq 0$. This implies that $\varphi(t) + \varphi(-t)$ is real for all $t \in \mathbb{R}$. Then we find for $z = i$ in equation (6.7)

$$2\varphi(0) + i\big(\varphi(t) - \varphi(-t)\big) \geq 0.$$

Therefore $\varphi(t) - \varphi(-t)$ is purely imaginary for all $t \in \mathbb{R}$, and we get

$$\begin{aligned} \varphi(t) + \varphi(-t) &= \overline{\varphi(t)} + \overline{\varphi(-t)}, \\ \varphi(t) - \varphi(-t) &= \overline{\varphi(-t)} - \overline{\varphi(t)}. \end{aligned}$$

Subtraction of these equalities implies that $\varphi(-t) = \overline{\varphi(t)}$. Now we prove the inequality $|\varphi(t)| \leq \varphi(0)$, $t \in \mathbb{R}$. If $\varphi(t) = 0$ it is obviously true. Suppose that $\varphi(t) \neq 0$. Choose $z = -\varphi(t)^{-1}|\varphi(t)|$, then inequality (6.7) gives

$$2\varphi(0) - 2|\varphi(t)| \geq 0,$$

and the proof is complete. □

6.15 Theorem *Let μ be a probability measure on* $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$*, Then its characteristic function $\hat{\mu}$ has the following properties:*

(a) *$\hat{\mu}$ is uniformly continuous on* $\mathbb{R}$.

(b) $\hat{\mu}(0) = 1$,

(c) *$\hat{\mu}$ is positive definite.*

Proof
Statement (a): Let $\varepsilon > 0$ be given. From exercise 6.7 we know that there exists $n_0 \in \mathbb{N}$ so that $\mu\big(\complement[-n_0, n_0]\big) \leq \varepsilon/4$. Consider $s, t \in \mathbb{R}$, $s < t$, then for $x \geq 0$,

$$\begin{aligned}
\left|e^{itx} - e^{isx}\right| &= \left|e^{isx/2}\right|\left|e^{itx/2}\right|\left|e^{i(t-s)x/2} - e^{-i(t-s)x/2}\right| \\
&= 2\left|\sin((t-s)x/2)\right| \\
&= 2\left|\int_0^{(t-s)x/2} \cos(u)\, du\right| \\
&\leq (t-s)x.
\end{aligned}$$

The cases $t < s$ and $x \leq 0$ can be done in a similar way, so that we get for all s, $t \in \mathbb{R}$, $x \in \mathbb{R}$ the inequality

$$\left|e^{itx} - e^{isx}\right| \leq |t-s||x|. \tag{6.8}$$

In particular, for $x \in \mathbb{R}$, with $|x| \leq n_0$,

$$\left|e^{itx} - e^{isx}\right| \leq n_0|t-s|.$$

The triangle inequality for Lebesgue integrals (A.22), which is proved in appendix A also for complex valued functions, gives

$$\begin{aligned}
|\hat{\mu}(t) - \hat{\mu}(s)| &= \left|\int_{\mathbb{R}} \left(e^{itx} - e^{isx}\right) d\mu(x)\right| \\
&\leq \int_{\mathbb{R}} \left|e^{itx} - e^{isx}\right| d\mu(x) \\
&= \int_{[-n_0,n_0]} \left|e^{itx} - e^{isx}\right| d\mu(x) + \int_{\complement[-n_0,n_0]} \left|e^{itx} - e^{isx}\right| d\mu(x) \\
&\leq n_0|t-s| + 2\,\frac{\varepsilon}{4}.
\end{aligned}$$

Now choose $\delta = (2n_0)^{-1}\varepsilon$ then $|t-s| < \delta$ implies $|\hat{\mu}(t) - \hat{\mu}(s)| < \varepsilon$.

Statement (b): is obvious.

Statement (c): This follows from

$$\begin{aligned}\sum_{k,l=1}^{n} z_k \overline{z_l}\, \hat{\mu}(t_k - t_l) &= \int_{\mathbb{R}} \sum_{k,l=1}^{n} z_k\, e^{it_k x}\, \overline{z_l}\, e^{-it_l x}\, d\mu(x) \\ &= \int_{\mathbb{R}} \Big| \sum_{k=1}^{n} z_k\, e^{it_k x} \Big|^2 d\mu(x) \\ &\geq 0.\end{aligned}$$

□

Statement (c) and lemma 6.14 provide us with the

6.16 Corollary *For every probability measure μ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, $|\hat{\mu}(t)| \leq 1$ and $\overline{\hat{\mu}(t)} = \hat{\mu}(-t)$ hold true for every $t \in \mathbb{R}$.*

(The statement of this corollary is also easily seen directly: $|\hat{\mu}(t)| \leq 1$, $t \in \mathbb{R}$, follows from the triangle inequality (A.22), and equation (A.25) entails the hermiticity property $\overline{\hat{\mu}(t)} = \hat{\mu}(-t)$, $t \in \mathbb{R}$.)

Notation From now on we denote the set of all probability measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ by $\mathcal{W}(\mathbb{R})$, and the set of all functions with the properties given in (a), (b), (c) of theorem 6.15 by $\mathcal{K}(\mathbb{R})$. Functions in $\mathcal{K}(\mathbb{R})$ will simply be called *characteristic functions* — this naming will be justified below (see theorem 6.18).

With this terminology we have proved in theorem 6.15 that the Fourier transform $\hat{\cdot}$ maps $\mathcal{W}(\mathbb{R})$ into $\mathcal{K}(\mathbb{R})$. Next we prove the injectivity of this mapping, which justifies the naming "characteristic" function:

6.17 Theorem *The Fourier transform $\hat{\cdot}$ is an injective mapping from $\mathcal{W}(\mathbb{R})$ into $\mathcal{K}(\mathbb{R})$.*

Proof Suppose that $\mu_1, \mu_2 \in \mathcal{W}(\mathbb{R})$ have the same characteristic function: $\hat{\mu}_1 = \hat{\mu}_2$. We have to show that $\mu_1 = \mu_2$. By lemma 5.4 it is sufficient to prove for all $f \in C_c(\mathbb{R})$ that

$$\int f\, d\mu_1 = \int f\, d\mu_2.$$

Without loss of generality we may assume that f is not the function which is identically zero, and that $\|f\|_\infty \leq 1$. Let $\varepsilon > 0$ be given, and we may suppose that $\varepsilon \leq 1$. Since the family consisting of μ_1 and μ_2 is tight (see exercise 6.7), there exists $R > 0$ so that $\mu_i(\complement[-R, R]) \leq \varepsilon/8$, $i = 1, 2$. Making R large enough, if necessary, we can also assume that $\operatorname{supp} f \subset [-R, R]$. f is continuous on $[-R, R]$, and we have $f(-R) = f(R) = 0$. Hence we can apply the theorem of Weierstraß C.2: There exists a trigonometric polynomial

$$f_n(x) = \sum_{k=-n}^{n} a_{n,k}\, e^{i\pi kx/R}, \qquad n \in \mathbb{N},\ a_{n,k} \in \mathbb{C},\ x \in \mathbb{R},$$

such that

$$\sup_{x\in[-R,R]} |f(x) - f_n(x)| \le \frac{\varepsilon}{4} \le \frac{1}{4}. \tag{6.9}$$

Now we estimate as follows:

$$\begin{aligned}
\Big|\int_{\mathbb{R}} f\, d\mu_1 - \int_{\mathbb{R}} f\, d\mu_2\Big| &\le \int_{\mathbb{R}} |f - f_n|\, d\mu_1 + \int_{\mathbb{R}} |f - f_n|\, d\mu_2 \\
&\qquad + \Big|\int_{\mathbb{R}} f_n\, d\mu_1 - \int_{\mathbb{R}} f_n\, d\mu_2\Big| \\
&= \int_{\mathbb{R}} |f - f_n|\, d\mu_1 + \int_{\mathbb{R}} |f - f_n|\, d\mu_2.
\end{aligned}$$

The last step comes from the fact that μ_1 and μ_2 have the same characteristic function, and this implies — by linearity of the integrals — that the integrals of f_n with respect to μ_1 and μ_2 coincide. Write μ for μ_1 or μ_2, then the last two integrals can be bounded in the following way:

$$\begin{aligned}
\int_{\mathbb{R}} |f - f_n|\, d\mu &= \int_{[-R,R]} |f - f_n|\, d\mu + \int_{\complement[-R,R]} |f - f_n|\, d\mu \\
&= \int_{[-R,R]} |f - f_n|\, d\mu + \int_{\complement[-R,R]} |f_n|\, d\mu \\
&\le \frac{\varepsilon}{4}\, \mu\big([-R,R]\big) + \|f_n\|_\infty\, \mu\big(\complement[-R,R]\big) \\
&\le \frac{\varepsilon}{4} + \frac{\varepsilon}{8}\, \|f_n\|_\infty.
\end{aligned}$$

f_n is periodic with period $2R$, so that $\|f_n\|_\infty$ is equal to the supremum of f_n on $[-R, R]$. On the other hand, by assumption $\|f\|_\infty \le 1$, and we have inequality 6.9, which entails that $\|f_n\|_\infty \le 2$. Thus both integrals in question contribute in absolute value at most $\varepsilon/2$, and therefore

$$\Big|\int_{\mathbb{R}} f\, d\mu_1 - \int_{\mathbb{R}} f\, d\mu_2\Big| \le \varepsilon.$$

The remark that $\varepsilon > 0$ was arbitrary concludes the proof. □

We conclude this section with

6.18 Theorem (Bochner) *The Fourier transform* $\hat{\cdot}$ *is surjective from* $\mathcal{W}(\mathbb{R})$ *to* $\mathcal{K}(\mathbb{R})$*, that is, every function in* $\mathcal{K}(\mathbb{R})$ *is the characteristic function of a probability measure* $\mu \in \mathcal{W}(\mathbb{R})$.

The proof of this theorem is deferred to appendix D. Taken together with the uniqueness theorem 6.17, we thus have

6.19 Corollary *The Fourier transform $\hat{\cdot}$ is a bijection from $\mathcal{W}(\mathbb{R})$ onto $\mathcal{K}(\mathbb{R})$.*

6.4 Lévy's Continuity Lemma

6.20 Lemma *Suppose that $(\mu_n,\, n \in \mathbb{N})$ is a sequence in $\mathcal{W}(\mathbb{R})$, and that the associated sequence $(\hat{\mu}_n,\, n \in \mathbb{N})$ of characteristic functions in $\mathcal{K}(\mathbb{R})$ converges pointwise to $\hat{\mu} \in \mathcal{K}(\mathbb{R})$. Then $(\mu_n,\, n \in \mathbb{N})$ is tight.*

Proof For $R_1 > 0$, $n \in \mathbb{N}$ we use the estimate (3.14) of corollary 3.27 as follows (recall that γ is the standard Gaussian measure on the real line, cf. (3.13)):

$$\begin{aligned}
\mu_n\big(\complement[-R_1, R_1]\big) &\le \mu_n\big(\complement(-R_1, R_1)\big) \\
&\le \frac{\sqrt{e}}{\sqrt{e}-1} \int_{\mathbb{R}} \Big(1 - \hat{\mu}_n\Big(\frac{t}{R_1}\Big)\Big)\, d\gamma(t) \\
&\le \frac{\sqrt{e}}{\sqrt{e}-1} \left(\int_{\mathbb{R}} \Big|1 - \hat{\mu}\Big(\frac{t}{R_1}\Big)\Big|\, d\gamma(t) \right. \\
&\qquad\qquad \left. + \int_{\mathbb{R}} \Big|\hat{\mu}\Big(\frac{t}{R_1}\Big) - \hat{\mu}_n\Big(\frac{t}{R_1}\Big)\Big|\, d\gamma(t) \right).
\end{aligned}$$

Both integrands are majorized by the γ–integrable constant 2, and the first converges for every $t \in \mathbb{R}$ to zero as R_1 tends to infinity, because $\hat{\mu}$ is continuous with value 1 at 0, while the second integrand tends for every $t \in \mathbb{R}$ to zero as $n \to \infty$ due to our hypothesis. Therefore the dominated convergence theorem A.56 justifies the following argument: Given $\varepsilon > 0$ we can find $R_1 > 0$ so that the first term above (including the constant in front) is less than $\varepsilon/2$. This R_1 being fixed, we can choose $n_0 \in \mathbb{N}$ so that for all $n \ge n_0$ the second term (again including the constant in front) is less than $\varepsilon/2$. Thus for all $n \ge n_0$ we get

$$\mu_n\big(\complement[-R_1, R_1]\big) \le \varepsilon.$$

On the other hand, by exercise 6.7 we can find $R_2 > 0$ so that for all $n \in \{1, 2, \ldots, n_0 - 1\}$ we have

$$\mu_n\big(\complement[-R_2, R_2]\big) \le \varepsilon.$$

Set $R = \max(R_1, R_2)$, then

$$\mu_n\big(\complement[-R, R]\big) \le \varepsilon$$

for all $n \in \mathbb{N}$. Thus the sequence $(\mu_n,\, n \in \mathbb{N})$ is tight. $\square$

6.21 Theorem (Lévy's Continuity Lemma) *Assume that* $(\mu_n, n \in \mathbb{N})$ *is a sequence in* $\mathcal{W}(\mathbb{R})$*, and that* $\mu \in \mathcal{W}(\mathbb{R})$*. Let* $(\hat{\mu}_n, n \in \mathbb{N})$*,* $\hat{\mu}$ *be the associated characteristic functions.* $(\mu_n, n \in \mathbb{N})$ *converges weakly to* μ*, if and only if* $(\hat{\mu}_n, n \in \mathbb{N})$ *converges pointwise to* $\hat{\mu}$*.*

Proof "$\Rightarrow$" For every $t \in \mathbb{R}$, the mapping $x \mapsto \exp(itx)$ belongs to $C_b(\mathbb{R})$, so that $\mu_n \xrightarrow{\text{w}} \mu$ directly by the definition of weak convergence implies $\hat{\mu}_n(t) \to \hat{\mu}(t)$, as $n \to \infty$.

"$\Leftarrow$" Lemma 6.20 implies that $(\mu_n, n \in \mathbb{N})$ is tight. From Prohorov's theorem 6.9 we then get that $(\mu_n, n \in \mathbb{N})$ is relatively compact. Thus every subsequence $(\mu_{n'})$ of $(\mu_n, n \in \mathbb{N})$ contains a further subsequence $(\mu_{n''})$ which converges to some probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, say ν. From the first part of the proof we find that $(\hat{\mu}_{n''})$ converges pointwise to $\hat{\nu}$. On the other hand, by hypothesis $(\hat{\mu}_{n''})$ converges pointwise to $\hat{\mu}$. Thus we have that $\hat{\nu} = \hat{\mu}$, and the uniqueness theorem 6.17 implies that $\nu = \mu$. Thus so far we have proved: Every subsequence $(\mu_{n'})$ of $(\mu_n, n \in \mathbb{N})$ contains a subsequence $(\mu_{n''})$ which converges weakly to μ. But then $(\mu_n, n \in \mathbb{N})$ must converge weakly to μ: Suppose not so, then there exist $\varepsilon > 0$ and $f \in C_b(\mathbb{R})$ such that for every $n \in \mathbb{N}$ there exists $n' \in \mathbb{N}$, $n' \geq n$, with

$$\left| \int f \, d\mu_{n'} - \int f \, d\mu \right| > \varepsilon.$$

But this subsequence $(\mu_{n'})$ cannot contain a subsequence which converges weakly to μ, which is a contradiction to the fact proved above. □

We consider $\mathcal{W}(\mathbb{R})$ as equipped with the topology of weak convergence, and $\mathcal{K}(\mathbb{R})$ as equipped with the topology of pointwise convergence. Then Lévy's continuity lemma in combination with theorem 6.17 and theorem 6.18 gives the following result

6.22 Corollary *The Fourier transform* $\hat{\cdot}$ *is a* homeomorphism*, that is, a bicontinuous bijection, from* $\mathcal{W}(\mathbb{R})$ *onto* $\mathcal{K}(\mathbb{R})$*.*

6.5 Characteristic Functions and Moments

Notation For $n \in \mathbb{N}$, $\mathcal{W}_n(\mathbb{R})$ denotes the set of all probability measures μ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ so that

$$\int_{\mathbb{R}} |x|^n \, d\mu(x) < +\infty.$$

6.23 Definition If $\mu \in \mathcal{W}_n(\mathbb{R})$, then

$$M_n(\mu) = \int_{\mathbb{R}} x^n \, d\mu(x), \qquad n \in \mathbb{N}, \tag{6.10}$$

is called the *n–th moment* of μ.

6.24 Theorem *Assume that $\hat{\mu}$ is the characteristic function of a probability measure $\mu \in \mathcal{W}_n(\mathbb{R})$, $n \in \mathbb{N}$. Then the derivatives $\hat{\mu}^{(k)}$, $k = 1, 2, \dots, n$, exist on all of $\mathbb{R}$ and are bounded, uniformly continuous functions. Moreover,*

$$\hat{\mu}^{(k)}(t) = i^k \int e^{itx} x^k \, d\mu(x), \qquad t \in \mathbb{R},\ k \in \{1, \dots, n\}, \tag{6.11}$$

and in particular

$$\hat{\mu}^{(k)}(0) = i^k M_k(\mu), \qquad k \in \{1, \dots, n\}, \tag{6.12}$$

hold true.

Proof If $\mu \in \mathcal{W}_n(\mathbb{R})$, $n \in \mathbb{N}$, then Hölder's inequality (A.37) shows that $\mu \in \mathcal{W}_k(\mathbb{R})$ for all $k \in \{1, \dots, n\}$. Suppose that $k \in \{0, 1, \dots, n-1\}$, and consider the function

$$f_k(t) = i^k \int e^{itx} x^k \, d\mu(x), \qquad t \in \mathbb{R}.$$

We show that f_k is differentiable, and that the derivative can be computed by differentiation under the integral. Let $t \in \mathbb{R}$, and let $(t_n,\ n \in \mathbb{N})$ be a sequence in $\mathbb{R}$ with $t_n \neq t$ for all $n \in \mathbb{N}$, and $t_n \to t$, $n \to \infty$. Then

$$\frac{f_k(t) - f_k(t_n)}{t - t_n} = i^k \int \frac{e^{itx} - e^{it_n x}}{t - t_n} x^k \, d\mu(x).$$

From inequality (6.8) we get

$$\left| \frac{e^{itx} - e^{it_n x}}{t - t_n} x^k \right| \leq |x|^{k+1},$$

which — by hypothesis and Hölder's inequality — is an integrable majorant of the integrand above, which is uniform in $n \in \mathbb{N}$. Thus we may apply the dominated convergence theorem A.56 to interchange the limit $n \to \infty$ with the integral, and obtain

$$f_k'(t) = i^{k+1} \int e^{itx} x^{k+1} \, d\mu(x), \qquad t \in \mathbb{R}.$$

Using this formula successively for $\hat{\mu}$ we get (6.11), and (6.12). The boundedness of $\hat{\mu}^{(k)}$, $k \in \{1, \dots, n\}$, is trivial. That it is uniformly continuous is shown in the same way as for $\hat{\mu}$, see the proof of (a) of theorem 6.15. □

6.25 Corollary *Suppose that $\mu \in \mathcal{W}_n(\mathbb{R})$, $n \in \mathbb{N}$. Then*

$$\hat{\mu}(t) = \sum_{k=0}^{n} M_k(\mu) \frac{(it)^k}{k!} + \frac{t^n}{n!} \theta_n(t), \qquad t \in \mathbb{R}, \tag{6.13a}$$

with a remainder θ_n satisfying

$$|\theta_n(t)| \leq \sup_{0 \leq \tau \leq 1} \left| \hat{\mu}^{(n)}(\tau t) - \hat{\mu}^{(n)}(0) \right|, \qquad t \in \mathbb{R}. \tag{6.13b}$$

In particular, θ_n is continuous in $t = 0$ with $\theta_n(0) = 0$.

Proof By theorem 6.24, $\hat{\mu}$ is n times continuously differentiable on $\mathbb{R}$, with $\hat{\mu}^{(k)}(0) = i^k M_k(\mu)$, $k = 1, \dots, n$. Therefore the formulae (6.13) follow directly from Taylor's theorem E.1. □

The case $n = 2$ will be of special importance for us in the next section, and for this case we make the last result somewhat more precise:

6.26 Corollary *Suppose that $\mu \in \mathcal{W}_2(\mathbb{R})$. Then*

$$\hat{\mu}(t) = 1 + itM_1(\mu) - \frac{1}{2}t^2 M_2(\mu) + \frac{1}{2}t^2\theta_\mu(t), \qquad t \in \mathbb{R}, \tag{6.14a}$$

with a function θ_μ which is continuous in $t = 0$ with $\theta_\mu(0) = 0$. Moreover, for all $\varepsilon > 0$, $t \in \mathbb{R}$, there exists $\delta > 0$ so that for all $\alpha > 0$, and all $\mu \in \mathcal{W}_2(\mathbb{R})$ the following inequality holds true:

$$\left|\theta_\mu\Big(\frac{t}{\alpha}\Big)\right| \le \varepsilon M_2(\mu) + 2\int_{\{|x|\ge\delta\alpha\}} x^2\, d\mu(x), \qquad t \in \mathbb{R}. \tag{6.14b}$$

Proof We only have to prove the second statement. Let $\varepsilon > 0$, and $t \in \mathbb{R}$ be given. If $t = 0$ we can choose any $\delta > 0$ and have nothing to prove. Thus let $t \neq 0$ and choose $\delta = \varepsilon/|t| > 0$. Choose $\mu \in \mathcal{W}_2(\mathbb{R})$. For $n = 2$ the right hand side of inequality (6.13b) is equal to

$$\sup_{0\le\tau\le1}\left|\int \big(e^{i\tau tx} - 1\big)\, x^2\, d\mu(x)\right| \le \sup_{0\le\tau\le1}\int \big|e^{i\tau tx} - 1\big|\, x^2\, d\mu(x).$$

Hence for all $\alpha > 0$ we find

$$\left|\theta_\mu\Big(\frac{t}{\alpha}\Big)\right| \le \sup_{0\le\tau\le1}\int \big|e^{i\tau tx/\alpha} - 1\big|\, x^2\, d\mu(x).$$

For $x \in \mathbb{R}$ with $|x| < \delta\alpha$ we obtain with inequality (6.8) for all $\tau \in [0, 1]$ the estimate

$$\big|e^{i\tau tx/\alpha} - 1\big| \le \tau|t|\,\frac{|x|}{\alpha} \le \varepsilon.$$

Then

$$\begin{aligned}
\left|\theta_\mu\Big(\frac{t}{\alpha}\Big)\right| &\le \sup_{0\le\tau\le1}\Big(\int_{\{|x|<\delta\alpha\}} \big|e^{i\tau tx/\alpha} - 1\big|\, x^2\, d\mu(x) \\
&\qquad\qquad + \int_{\{|x|\ge\delta\alpha\}} \big|e^{i\tau tx/\alpha} - 1\big|\, x^2\, d\mu(x)\Big) \\
&\le \sup_{0\le\tau\le1}\Big(\varepsilon\int_{\{|x|<\delta\alpha\}} x^2\, d\mu(x) + 2\int_{\{|x|\ge\delta\alpha\}} x^2\, d\mu(x)\Big) \\
&\le \varepsilon\int_{\mathbb{R}} x^2\, d\mu(x) + 2\int_{\{|x|\ge\delta\alpha\}} x^2\, d\mu(x) \\
&= \varepsilon M_2(\mu) + 2\int_{\{|x|\ge\delta\alpha\}} x^2\, d\mu(x),
\end{aligned}$$

and the proof is complete. □

6.27 Remark It is important to observe that in the second statement of corollary 6.26, the choice of $\delta > 0$ is *independent* of the choice of $\mu \in \mathcal{W}_2(\mathbb{R})$.

Chapter 7

Central Limit Theorem

As a motivation let us consider an *iid* sequence $(X_n,\ n \in \mathbb{N})$ in $\mathcal{L}^2(P)$ with $V(X_1) > 0$. Then it is also an *iid* sequence in $\mathcal{L}^1(P)$, and the theorem of Kolmogorov 4.13 states that

$$\frac{1}{n} S_n \xrightarrow{\text{a.s.}} E(X_1), \quad n \to \infty,$$

where

$$S_n = \sum_{k=1}^{n} X_k.$$

The question arises how the random variable S_n/n varies around its mean and limit $E(X_1)$. Since S_n/n converges a.s. to a constant, in the limit its variance vanishes. Therefore, we have to normalize this random variable appropriately in order to get a nontrivial answer. The right way to do this is to divide by its standard deviation, in other words, we consider the standardization of S_n/n:

$$Z_n = \frac{S_n/n - E(X_1)}{\sigma(S_n/n)} = \frac{S_n - nE(X_1)}{\sqrt{nV(X_1)}},$$

where we used Bienaymé's theorem 2.16 to compute $\sigma(S_n/n)$. Set $m = E(X_1)$, $\sigma^2 = V(X_1)$, and

$$Y_{nk} = \frac{X_k - m}{\sqrt{n\sigma^2}}, \qquad n \in \mathbb{N},\ k = 1, 2, \ldots, n.$$

Then we are interested in the asymptotic behavior of $\sum_{k=1}^{n} Y_{nk}$ for $n \to \infty$. The relevant notion of convergence for this problem is convergence in law.

7.1 Preparations

First we prove a simple lemma from analysis:

7.1 Lemma (Compound Interest Formula) *Assume that $n \in \mathbb{N}$, $a_1, \ldots, a_n$, $b_1, \ldots, b_n \in \mathbb{C}$. Then*

$$\prod_{k=1}^{n} a_k - \prod_{k=1}^{n} b_k = \sum_{k=1}^{n} a_1 \cdots a_{k-1} (a_k - b_k) b_{k+1} \cdots b_n, \tag{7.1}$$

and

$$\Big|\prod_{k=1}^{n} a_k - \prod_{k=1}^{n} b_k\Big| \le M_{a,b}^{n-1} \sum_{k=1}^{n} |a_k - b_k|, \tag{7.2}$$

holds, where $M_{a,b} = \max\{|a_k|, |b_k|, k = 1, \ldots, n\}$. In particular, if $|a_k|$, $|b_k| \le 1$ for all $k \in \{1, \ldots, n\}$, then

$$\Big|\prod_{k=1}^{n} a_k - \prod_{k=1}^{n} b_k\Big| \le \sum_{k=1}^{n} |a_k - b_k|, \tag{7.3}$$

If $a_j = a$, $j \in \{1, \ldots, n\}$, $b_k = b$, $k \in \{1, \ldots, n\}$, then

$$|a^n - b^n| \le n\,|a - b|\,(\max(|a|, |b|))^{n-1}. \tag{7.4}$$

Moreover, if $(\lambda_n, n \in \mathbb{N})$ is a sequence in $\mathbb{C}$ which converges to $\lambda \in \mathbb{C}$, then

$$\Big(1 + \frac{\lambda_n}{n}\Big)^n \tag{7.5}$$

converges to $\exp(\lambda)$ *as n tends to infinity.*

Proof We compute as follows:

$$\begin{aligned}
&\sum_{k=1}^{n} a_1 \cdots a_{k-1} (a_k - b_k) b_{k+1} \cdots b_n \\
&\qquad = \sum_{k=1}^{n} a_1 \cdots a_k b_{k+1} \cdots b_n - \sum_{k=1}^{n} a_1 \cdots a_{k-1} b_k \cdots b_n \\
&\qquad = a_1 \cdots a_n - b_1 \cdots b_n \\
&\qquad\qquad + \sum_{k=1}^{n-1} a_1 \cdots a_k b_{k+1} \cdots b_n - \sum_{k=2}^{n} a_1 \cdots a_{k-1} b_k \cdots b_n.
\end{aligned}$$

Clearly the last two sums cancel each other, so that we obtain (7.1). The inequalities (7.2)–(7.4) are direct consequences. For the proof of (7.5), we first observe the following form of the compound interest formula, whose proof we leave as an *exercise* to the reader.

$$\Big(1 + \frac{\lambda}{n}\Big)^n \xrightarrow[n]{} e^{\lambda}, \qquad \lambda \in \mathbb{C}. \tag{7.6}$$

(*Hint:* Prove the bound $|e^z - (1+z)| \le |z|^2 e^{|z|}$, $z \in \mathbb{C}$, and use inequality (7.4).)
With (7.6) it is sufficient to show that

$$\left|\Big(1+\frac{\lambda_n}{n}\Big)^n - \Big(1+\frac{\lambda}{n}\Big)^n\right| \to 0, \quad n \to \infty.$$

We use (7.4) to estimate this expression from above by

$$\begin{aligned} n\left|\frac{\lambda_n}{n} - \frac{\lambda}{n}\right| \Big(\max\big(|1+\lambda_n/n|,|1+\lambda/n|\big)\Big)^{n-1} & \\ \le |\lambda_n - \lambda| \Big(1 + \frac{|\lambda_n| + |\lambda|}{n}\Big)^{n-1}. \end{aligned}$$

Let $n_0 \in \mathbb{N}$ be large enough so that for all $n \ge n_0$ we have $|\lambda_n - \lambda| \le 1$. Then we find for such $n \in \mathbb{N}$,

$$\begin{aligned} \left|\Big(1+\frac{\lambda_n}{n}\Big)^n - \Big(1+\frac{\lambda}{n}\Big)^n\right| &\le |\lambda_n - \lambda| \left(1 + \frac{1+2|\lambda|}{n}\right)^{n-1} \\ &\le |\lambda_n - \lambda| \left(1 + \frac{1+2|\lambda|}{n}\right)^{n}. \end{aligned}$$

Now $|\lambda_n - \lambda|$ converges to zero with $n \to \infty$, while

$$\left(1 + \frac{1+2|\lambda|}{n}\right)^{n}$$

converges to $\exp(1 + 2|\lambda|)$, so that the whole expression vanishes in the limit as n tends to infinity. □

7.2 Lemma *Assume that $X_1, X_2, \dots, X_n$ are independent real valued random variables with characteristic functions φ_{X_k}, $k = 1, \dots, n$. Then*

$$\varphi_{X_1+\cdots+X_n}(t) = \varphi_{X_1}(t)\cdots\varphi_{X_n}(t), \qquad t \in \mathbb{R}, \tag{7.7}$$

where $\varphi_{X_1+\cdots+X_n}$ is the characteristic function of $X_1 + \cdots + X_n$.

Proof For $t \in \mathbb{R}$ we get

$$\varphi_{X_1+\cdots+X_n}(t) = E\Big(\prod_{k=1}^n e^{itX_k}\Big) = \prod_{k=1}^n E\Big(e^{itX_k}\Big) = \prod_{k=1}^n \varphi_{X_k}(t),$$

where we used theorems 2.25, 2.30. □

7.3 Example Consider an *iid* sequence of integrable random variables. Kolmogorov's strong law of large numbers 4.13 implies that almost surely

$$\frac{1}{n}(X_1 + \cdots + X_n) \to E(X_1) = m, \quad n \to \infty,$$

and lemma 3.16 entails that this also holds true in probability, that is, the weak law of large numbers holds true for $(X_n, n \in \mathbb{N})$. We show this now with the help of characteristic functions and Lévy's continuity lemma, theorem 6.21. From exercise 5.10 we know that it is sufficient to prove convergence in law, and theorem 6.21 shows that it is sufficient to prove pointwise convergence of the characteristic functions. The characteristic function of the constant random variable $E(X_1) = m$ is given by $\exp(imt)$, $t \in \mathbb{R}$. Thus we have to show that for every $t \in \mathbb{R}$,

$$\varphi_{(X_1+\cdots+X_n)/n}(t) \to e^{imt}, \quad n \to \infty.$$

We have

$$\varphi_{(X_1+\cdots+X_n)/n}(t) = \varphi_{X_1+\cdots+X_n}\Big(\frac{t}{n}\Big) = \varphi_{X_1}\Big(\frac{t}{n}\Big)^n,$$

with lemma 7.2. From corollary 6.25 we get that

$$\varphi_{X_1}\Big(\frac{t}{n}\Big) = 1 + i\,\frac{t}{n}\,m + \frac{t}{n}\,\theta_1\Big(\frac{t}{n}\Big),$$

and $\theta_1(t/n)$ converges to zero with $n \to \infty$, so that

$$i\,t\,m + t\,\theta_1\Big(\frac{t}{n}\Big) \to i\,tm, \quad n \to \infty.$$

Thus the last statement of lemma 7.1 gives that

$$\Big(1 + i\,\frac{t}{n}\,m + \frac{t}{n}\,\theta_1\Big(\frac{t}{n}\Big)\Big)^n \to e^{itm}, \quad n \to \infty,$$

and we are done. ◇

7.2 Poisson's Theorem

7.4 Theorem *Suppose that for every $n \in \mathbb{N}$, Y_{nk}, $k = 1, \dots, n$, are independent, identically distributed Bernoulli random variables with parameter $p_n \in [0, 1]$. Assume furthermore that $np_n \to \alpha > 0$ as $n \to \infty$. Then the sequence $(X_n, n \in \mathbb{N})$, $X_n = Y_{n1} + \cdots + Y_{nn}$, $n \in \mathbb{N}$, converges in law to a random variable with Poisson law of parameter α.*

Proof For $n \in \mathbb{N}$, X_n is binomially distributed with parameters n, p_n, and therefore its characteristic function is given by

$$\varphi_{X_n}(t) = \big(1 + p_n(e^{it} - 1)\big)^n, \qquad t \in \mathbb{R}.$$

Put $\lambda_n = np_n(\exp(it) - 1)$, which converges to $\alpha(\exp(it) - 1)$ as $n \to \infty$. Thus with lemma 7.1 we get that

$$\varphi_{X_n}(t) \to e^{\alpha(e^{it}-1)}, \qquad t \in \mathbb{R},$$

as n tends to infinity. The proof is finished by theorem 6.21, and the remark that the last expression is the characteristic function of a Poisson random variable with parameter α (see example 6.12.(c)). □

7.3 Central Limit Theorem

The next theorem is the classical central limit theorem of *Laplace–DeMoivre*. We shall get it as a special case from theorem 7.10, but its proof is also an easy *exercise* based on theorem 6.21, lemma 7.1 and lemma 7.2.

7.5 Theorem (Laplace–DeMoivre) *Assume that* $(X_n,\ n \in \mathbb{N})$ *is an* iid *sequence of real valued random variables in* $\mathcal{L}^2(P)$*, such that* $V(X_1) > 0$*. For* $n \in \mathbb{N}$ *set*

$$Z_n = \frac{X_1 + \cdots + X_n - nE(X_1)}{\sqrt{nV(X_1)}}.$$

Then $(Z_n,\ n \in \mathbb{N})$ *converges in law to a standard normal random variable.*

For the remainder of this section we suppose that $X = (X_n,\ n \in \mathbb{N})$ is an independent sequence of random variables in $\mathcal{L}^2(P)$, such that for every $n \in \mathbb{N}$, $V(X_n) > 0$. Throughout we shall use the following

Notation For $n \in \mathbb{N}$, $\varepsilon > 0$, set

$$\begin{aligned}
\xi_n &= E(X_n), \\
\sigma_n^2 &= V(X_n), \\
s_n^2 &= \sum_{k=1}^{n} \sigma_k^2, \\
Y_n &= \frac{\sum_{k=1}^{n}(X_k - \xi_k)}{s_n}, \\
L_n^X(\varepsilon) &= \frac{1}{s_n^2} \sum_{k=1}^{n} \int_{\{|x-\xi_k| \geq \varepsilon s_n\}} (x - \xi_k)^2 \, dP_{X_k}(x).
\end{aligned}$$

7.6 Definition The sequence $X = (X_n,\ n \in \mathbb{N})$ satisfies the

(a) *Lindeberg condition*, if $\lim_n L_n^X(\varepsilon) = 0$ for every $\varepsilon > 0$,

(b) *Feller condition*, if
$$\lim_{n\to\infty} \left(\max_{1 \leq k \leq n} \frac{\sigma_k}{s_n} \right) = 0.$$

7.7 Exercise Show that if $(X_n,\ n \in \mathbb{N})$ is an *iid* sequence in $\mathcal{L}^2(P)$ with $V(X_1) > 0$, then it satisfies the Lindeberg condition.

7.8 Lemma *If the sequence* $X = (X_n,\ n \in \mathbb{N})$ *satisfies the Lindeberg condition, then it also satisfies the Feller condition.*

Proof Let $\varepsilon > 0$ be given. For $n \in \mathbb{N}$, $k = 1, \dots, n$ consider

$$\begin{aligned}
\sigma_k^2 &= \int_{\mathbb{R}} (x-\xi_k)^2\, dP_{X_k}(x) \\
&= \int_{\{|x-\xi_k|<\varepsilon s_n\}} (x-\xi_k)^2\, dP_{X_k}(x) + \int_{\{|x-\xi_k|\ge\varepsilon s_n\}} (x-\xi_k)^2\, dP_{X_k}(x) \\
&\le \varepsilon^2 s_n^2 + \int_{\{|x-\xi_k|\ge\varepsilon s_n\}} (x-\xi_k)^2\, dP_{X_k}(x) \\
&\le \varepsilon^2 s_n^2 + \sum_{k=1}^n \int_{\{|x-\xi_k|\ge\varepsilon s_n\}} (x-\xi_k)^2\, dP_{X_k}(x) \\
&= \big(\varepsilon^2 + L_n^X(\varepsilon)\big) s_n^2.
\end{aligned}$$

Hence we get

$$\Big(\max_{1\le k\le n} \frac{\sigma_k^2}{s_n^2}\Big) \le \varepsilon^2 + L_n^X(\varepsilon),$$

and consequently

$$\limsup_n \Big(\max_{1\le k\le n} \frac{\sigma_k^2}{s_n^2}\Big) \le \varepsilon^2.$$

Since $\varepsilon > 0$ was arbitrary we find that

$$\lim_{n\to\infty} \Big(\max_{1\le k\le n} \frac{\sigma_k^2}{s_n^2}\Big) = 0.$$

Observe that

$$\max_{1\le k\le n} \frac{\sigma_k^2}{s_n^2} = \Big(\max_{1\le k\le n} \frac{\sigma_k}{s_n}\Big)^2,$$

and it follows that X satisfies the Feller condition. □

7.9 Lemma *Let $X = (X_n, n \in \mathbb{N})$ be as above, and let $Z = (Z_n, n \in \mathbb{N})$ be an independent sequence of normally distributed random variables such that $E(Z_n) = E(X_n) = \xi_n$, $V(Z_n) = V(X_n) = \sigma_n^2$, $n \in \mathbb{N}$. If X satisfies the Lindeberg condition then also Z.*

Proof Since we have everywhere centered expressions of the form $X_n - \xi_n$, we may assume without loss of generality that $\xi_n = 0$ for all $n \in \mathbb{N}$. Then

$$\begin{aligned}
L_n^Z(\varepsilon) &= \frac{1}{s_n^2} \sum_{k=1}^n \int_{\{|x|\ge\varepsilon s_n\}} x^2\, dP_{Z_k}(x) \\
&= \frac{1}{s_n^2} \sum_{k=1}^n \frac{1}{\sqrt{2\pi\sigma_k^2}} \int_{\{|x|\ge\varepsilon s_n\}} x^2\, e^{-x^2/(2\sigma_k^2)}\, dx \\
&= \frac{1}{s_n^2} \sum_{k=1}^n \sigma_k^2 \frac{1}{\sqrt{2\pi}} \int_{\{|x|\ge\varepsilon s_n/\sigma_k\}} x^2\, e^{-x^2/2}\, dx.
\end{aligned}$$

Set
$$\alpha_n = \max_{1\le k\le n} \frac{\sigma_k}{s_n} > 0,$$
then we find
$$\begin{aligned} L_n^Z(\varepsilon) &\le \frac{1}{s_n^2}\sum_{k=1}^n \sigma_k^2 \frac{1}{\sqrt{2\pi}} \int_{\{|x|\ge\varepsilon/\alpha_n\}} x^2\, e^{-x^2/2}\, dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{\{|x|\ge\varepsilon/\alpha_n\}} x^2\, e^{-x^2/2}\, dx. \end{aligned}$$
Since X satisfies the Lindeberg condition, by lemma 7.8 it also satisfies the Feller condition, and therefore $\alpha_n \to 0$ with $n \to \infty$. Since the function under the last integral is integrable, it follows that the integral converges to zero as n tends to infinity.□

7.10 Theorem (Lindeberg–Lévy) *Let $X = (X_n,\ n \in \mathbb{N})$ be as before, and assume that it satisfies the Lindeberg condition. Then $Y = (Y_n,\ n \in \mathbb{N})$ as defined above converges in law to a standard normal random variable.*

Proof Without loss of generality, we may assume that the sequence X is centered: $\xi_n = 0$ for all $n \in \mathbb{N}$. Let $Z = (Z_n,\ n \in \mathbb{N})$ be an independent sequence of normally distributed random variables as in lemma 7.9: $Z_n \sim \mathcal{N}(0, \sigma_n^2)$, $n \in \mathbb{N}$. Because of Lévy's continuity lemma, theorem 6.21, it is sufficient to prove that the sequence of characteristic funcction $(\varphi_{Y_n},\ n \in \mathbb{N})$ converges pointwise to the characteristic function $\varphi_0(t) = \exp(-t^2/2)$, $t \in \mathbb{R}$, of the standard normal law. From lemma 7.1 we get for $t \in \mathbb{R}$,
$$\varphi_{Y_n}(t) = \prod_{k=1}^n \varphi_{X_k/s_n}(t) = \prod_{k=1}^n \varphi_{X_k}\Big(\frac{t}{s_n}\Big).$$
Thus
$$\begin{aligned} |\varphi_{Y_n}(t) - \varphi_0(t)| &= \Big|\prod_{k=1}^n \varphi_{X_k}\Big(\frac{t}{s_n}\Big) - \prod_{k=1}^n \exp\Big(-\frac{\sigma_k^2 t^2}{2s_n^2}\Big)\Big| \\ &= \Big|\prod_{k=1}^n \varphi_{X_k}\Big(\frac{t}{s_n}\Big) - \prod_{k=1}^n \varphi_{Z_k}\Big(\frac{t}{s_n}\Big)\Big| \\ &\le \sum_{k=1}^n \Big|\varphi_{X_k}\Big(\frac{t}{s_n}\Big) - \varphi_{Z_k}\Big(\frac{t}{s_n}\Big)\Big|, \end{aligned}$$
where made use of inequality (7.3) of lemma 7.1. Now we apply corollary 6.26 (note that $\xi_k = E(X_k) = 0$, $\sigma_k^2 = E(X_k^2)$):
$$\begin{aligned} \varphi_{X_k}\Big(\frac{t}{s_n}\Big) &= 1 - \frac{1}{2}\frac{t^2}{s_n^2}\sigma_k^2 + \frac{1}{2}\frac{t^2}{s_n^2}\theta_{P_{X_k}}\Big(\frac{t}{s_n}\Big), \\ \varphi_{Z_k}\Big(\frac{t}{s_n}\Big) &= 1 - \frac{1}{2}\frac{t^2}{s_n^2}\sigma_k^2 + \frac{1}{2}\frac{t^2}{s_n^2}\theta_{P_{Z_k}}\Big(\frac{t}{s_n}\Big). \end{aligned}$$

so that

$$\varphi_{X_k}\Big(\frac{t}{s_n}\Big) - \varphi_{Z_k}\Big(\frac{t}{s_n}\Big) = \frac{1}{2}\,\frac{t^2}{s_n^2}\Big(\theta_{P_{X_k}}\Big(\frac{t}{s_n}\Big) - \theta_{P_{Z_k}}\Big(\frac{t}{s_n}\Big)\Big),$$

and we find

$$|\varphi_{Y_n}(t) - \varphi_0(t)| \le \frac{1}{2}\,\frac{t^2}{s_n^2}\sum_{k=1}^{n}\Big(\Big|\theta_{P_{X_k}}\Big(\frac{t}{s_n}\Big)\Big| + \Big|\theta_{P_{Z_k}}\Big(\frac{t}{s_n}\Big)\Big|\Big).$$

Let $\varepsilon > 0$ be given. By corollary 6.26 there exists $\delta > 0$, so that for all $k \in \mathbb{N}$ we have the following estimates

$$\Big|\theta_{P_{X_k}}\Big(\frac{t}{s_n}\Big)\Big| \le \varepsilon\sigma_k^2 + 2\int_{\{|x|\ge\delta s_n\}} x^2\, dP_{X_k}(x),$$
$$\Big|\theta_{P_{Z_k}}\Big(\frac{t}{s_n}\Big)\Big| \le \varepsilon\sigma_k^2 + 2\int_{\{|x|\ge\delta s_n\}} x^2\, dP_{Z_k}(x).$$

We insert this above and get

$$\begin{aligned}|\varphi_{Y_n}(t) - \varphi_0(t)| &\le \frac{t^2}{s_n^2}\Big(\varepsilon\sum_{k=1}^{n}\sigma_k^2 + s_n^2 L_n^X(\delta) + s_n^2 L_n^Z(\delta)\Big)\\ &= t^2\big(\varepsilon + L_n^X(\delta) + L_n^Z(\delta)\big).\end{aligned}$$

By hypothesis we have that $L_n^X(\delta) \to 0$ as $n \to \infty$, and by lemma 7.9 we also have that $L_n^Z(\delta) \to 0$ as $n \to \infty$. Thus we obtain

$$\limsup_n |\varphi_{Y_n}(t) - \varphi_0(t)| \le t^2\varepsilon,$$

and since $\varepsilon > 0$ was arbitrary, the proof is concluded. □

7.11 Remark A family $(X_{nj},\ j = 1,\dots,n,\ n \in \mathbb{N})$ of real valued random variables is called *asymptotically negligible*, if for every $\varepsilon > 0$

$$\lim_{n\to\infty}\max_{j=1,\dots,n} P(|X_{nj}| > \varepsilon) = 0.$$

Feller has proved that under the (mild) assumption, that the sequence X in theorem 7.10 is such that $(X_{nj},\ j = 1,\dots,n,\ n \in \mathbb{N})$ with

$$X_{nj} = \frac{1}{s_n}(X_j - \xi_j)$$

is asymptotically negligible (ξ_j and s_n as above), the Lindeberg condition is also *necessary* in order that the central limit theorem holds true (cf., e.g., [5, § 28]).

Appendix A

Lebesgue Integration

This appendix is devoted to an account of the essentials of the measure and integration theory developed by H. Lebegue and E. Borel. Often, proofs are not given, instead the interested reader is asked to fill in details as exercises or is referred to the pertinent literature, such as [2–4, 6, 9, 13, 17, 18]. However, here and there some arguments are sketched.

The basic idea of measure theory is very simple: Approximate “complicated” sets by “simple” sets, of which one can determine in some elementary, e.g., “geometric”, way their “size” (length, area, volume, duration etc.), and “take the limit”. For example, one can approximate a circle in the plane — e.g., from the outside or the inside — by figures made up from rectangles or triangles, for which we have simple formulae to compute their areas, and in the limit we obtain the area of the circle. It turns out, however, that this simple idea leads to rather abstract and complicated looking constructions and arguments. The reason is that at the outset it is not so clear what the “good class of complicated sets” is, and what in general it means to “approximate” them by “simple sets”. Next one has to give a meaning to the idea to approximate not only the complicated sets but also their “size”. The abstraction of the idea of simple sets will be semi-rings of subsets of a given set, “complicated sets” will be the elements in a σ–algebra of subsets of that given set. The notion of “approximation” will then be casted as the σ–algebra generated by a semi-ring (or other types of systems of subsets of a given set). These ideas will be formulated in the next section. In the following section, section A.3, we shall study the idea of associating a size with a set. In particular, the idea of approximating the size of a “complicated” set by those of “simple” sets will lead to the central question of the extension of a σ–additive, non-negative set function from a semi-ring to the σ–algebra generated by it. This problem is solved via *Carathédory’s theorem*. Being equipped with these tools, we construct the Lebesgue integral of functions in section A.7, and investigate its properties in the subsequent sections of this appendix.

Throughout this appendix we assume that we are given some basic, non-empty set Ω. All appearing sets will be subsets of Ω.

A.1 Systems of Sets

We shall consider various subsets $\mathcal{A}$ of the power set $\mathcal{P}(\Omega)$ of Ω, that is, $\mathcal{A}$ is a *system* or a *family* of subsets of Ω.

A.1 Definition Consider $\mathcal{A} \subset \mathcal{P}(\Omega)$.

(a) $\mathcal{A}$ is called a *semiring*, if $\mathcal{A}$ is stable under intersections ("$\cap$–stable"), and if for $A, B \in \mathcal{A}$ there exist $n \in \mathbb{N}$, and $C_1, C_2, \ldots, C_n \in \mathcal{A}$, which are pairwise disjoint, and such that
$$A \setminus B = \biguplus_{k=1}^{n} C_k$$
holds true.

(b) $\mathcal{A}$ is a *ring*, if $\mathcal{A}$ is stable with respect unions ("$\cup$–stable") and differences ("$\setminus$–stable").

(c) $\mathcal{A}$ is an *algebra*, if it is a ring and $\Omega \in \mathcal{A}$.

(d) $\mathcal{A}$ is a *σ–algebra*, if $\mathcal{A}$ is an algebra, and $\mathcal{A}$ is stable with respect to countable unions, that is, if $(A_n,\ n \in \mathbb{N})$ is a sequence in $\mathcal{A}$, then $\cup_{n\in\mathbb{N}} A_n$ belongs to $\mathcal{A}$.

A.2 Remark In the literature a σ–algebra is often also called a *σ–field*.

A.3 Exercise

(a) Show that $\mathcal{A}$ is a σ–algebra over Ω, if and only if the following hold true:

  (i) $\Omega \in \mathcal{A}$;

  (ii) $\mathcal{A}$ is stable under complements ("$\complement$–stable");

  (iii) $\mathcal{A}$ is stable under countable unions.

(b) Show that every σ–algebra is stable with respect to countable intersections.

(c) Suppose that $\mathcal{A}$ is a σ–algebra over Ω, and that $B \subset \Omega$. Then the family $\mathcal{A} \cap B = \{C \subset \Omega,\ C = A \cap B,\ A \in \mathcal{A}\}$ forms a σ–algebra over B, called the *trace σ–algebra* of $\mathcal{A}$ on B.

A.4 Definition Suppose that $\mathcal{A}$ is a σ–algebra over Ω, then $(\Omega, \mathcal{A})$ is called a *measurable space*.

A.5 Example Suppose that $\Omega = \mathbb{R}^n$, $n \in \mathbb{N}$. For $a, b \in \mathbb{R}^n$ so that $a_i \leq b_i$, $i = 1, \ldots, n$, denote
$$(a, b] = \{x \in \mathbb{R}^n,\ a_i < x_i \leq b_i,\ i = 1, \ldots, n\},$$

called a *right semiclosed interval* in $\mathbb{R}^n$. (Analogously on defines *left semiclosed intervals*.) The family $\mathcal{J}^n$ of right semiclosed intervals forms a semiring over $\mathbb{R}^n$, but not a ring. The family of all subsets of $\mathbb{R}^n$ which can be written as pairwise disjoint unions of finitely many intervals in $\mathcal{J}^n$ forms a ring, but not an algebra. One obtains an algebra if one includes in the process of forming these unions also sets where some or all of the a_i (b_i, resp.) take the value $-\infty$ ($+\infty$, resp., but then always with strict inequalities above). But this will *not* yield a σ–algebra — in fact, one cannot construct a σ–algebra this way through any process involving only countably many steps, that is, by induction over $\mathbb{N}$. (However, it is possible to do so using a process of *transfinite induction*.) $\diamond$

As indicated in the last example, except in trivial cases it is almost never possible to construct a σ–algebra in an explicit manner. Instead one often defines σ–algebras as the minimal σ–algebra containing a certain given system of sets. First one shows in an *exercise* that if $(\mathcal{A}_i,\ i \in I)$ is a family of σ–algebras (rings, algebras resp.), I any non-empty index set, then also $\mathcal{A} = \cap_{i\in I}\mathcal{A}_i$ is a σ–algebra (ring, algebra, resp.). With this in mind, one makes the following

A.6 Definition Suppose that $\mathcal{M} \subset \mathcal{P}(\Omega)$. Then $\mathcal{R}(\mathcal{M})$, $\mathcal{A}(\mathcal{M})$, $\sigma(\mathcal{M})$ denote the smallest ring, algebra, σ–algebra resp., containing the family $\mathcal{M}$ of subsets of Ω. They are called the *ring, algebra, σ–algebra, resp., generated by $\mathcal{M}$*.

It is not hard to show that if $\mathcal{S}$ is a semiring over Ω, then the ring $\mathcal{R}(\mathcal{S})$ generated by $\mathcal{S}$ is given by the family of all (finite) pairwise disjoint unions of sets in $\mathcal{S}$.

A.7 Example Consider again $\Omega = \mathbb{R}^n$, $n \in \mathbb{N}$, and the semiring $\mathcal{J}^n$ of right semiclosed intervals. The σ–algebra $\sigma(\mathcal{J}^n)$ generated by $\mathcal{J}^n$ is called the *Borel σ–algebra over* $\mathbb{R}^n$, and it is denoted by $\mathcal{B}(\mathbb{R}^n)$.

For notational simplicity we continue our discussion for the case that $n = 1$, and abbreviate $\mathcal{J} = \mathcal{J}^1$. Consider any open interval (a, b), and the associated closed interval $[a, b]$. Then we can write

$$(a,b) = \bigcup_{n\in\mathbb{N}} \Big(a, b - \frac{1}{n}\Big], \qquad [a,b] = \bigcap_{n\in\mathbb{N}} \Big(a - \frac{1}{n}, b\Big],$$
$$(a,b] = \bigcap_{n\in\mathbb{N}} \Big(a, b + \frac{1}{n}\Big) = \bigcup_{n\in\mathbb{N}} \Big[a + \frac{1}{n}, b\Big].$$

Therefore, $\mathcal{B}(\mathbb{R})$ contains all left semiclosed, all open, and all closed intervals (and in particular all singletons $\{x\}$ with $x \in \mathbb{R}$). Moreover, if we consider the minimal σ–algebra over $\mathbb{R}$ which contains all open (closed, resp.) intervals, then the third relation above shows that it also contains all right semiclosed intervals. Minimality then implies that all these σ–algebras are equal, and a trivial modification of our arguments shows that the same is true if we had defined $\mathcal{B}(\mathbb{R})$ as the σ–algebra generated by the left semiclosed intervals. Thus we have proved that $\mathcal{B}(\mathbb{R})$ is generated by all intervals, or by the open, the closed, the right or the left semiclosed intervals.

Consider now the family $\mathcal{J}_0$ of intervals of the form $(-\infty, a]$, $a \in \mathbb{R}$. Since $\sigma(\mathcal{J}_0)$ is stable with respect to differences, for any a, $b \in \mathbb{R}$ with $a \leq b$ we have that $(a, b] = (-\infty, b] \setminus (-\infty, a]$ belongs to $\sigma(\mathcal{J}_0)$. Thus we get $\mathcal{B}(\mathbb{R}) \subset \sigma(\mathcal{J}_0)$. Conversely, since for every $a \in \mathbb{R}$, the interval $(-\infty, a] = \cup_n(-n, a]$ belongs to $\sigma(\mathcal{J}_0)$, we find that $\sigma(\mathcal{J}_0) \subset \mathcal{B}(\mathbb{R})$. That is, $\mathcal{B}(\mathbb{R}) = \sigma(\mathcal{J}_0)$, or in other words, $\mathcal{J}_0$ is a $\cap$–stable generator of $\mathcal{B}(\mathbb{R})$.

We go one step further. Observe that any open set $O \subset \mathbb{R}$ can be written as a countable union of open intervals: To see this, let $\mathcal{J}_{\mathbb{Q}}$ denote the family of all open intervals in $\mathbb{R}$ with rational endpoints, so that $\mathcal{J}_{\mathbb{Q}}$ is countable. Let $(I_n,\, n \in \mathbb{N})$ be an enumeration of $\mathcal{J}_{\mathbb{Q}}$. Then

$$O = \bigcup_{n \in \mathbb{N},\, I_n \subset O} I_n.$$

Indeed, that the set on the right hand side is a subset of O is trivial, while if $x \in O$, then because O is open there is an open interval containing x which is a subset O, and by shrinking this interval a bit, we find that $x \in I_n \subset O$ for some $n \in \mathbb{N}$. Thus O is a subset of the right hand side above, and the equality is proved. But now it follows, that $\mathcal{B}(\mathbb{R})$ contains the family $\mathcal{O}$ all open subsets of $\mathbb{R}$. Since by definition $\sigma(\mathcal{O})$ is the smallest σ–algebra containing $\mathcal{O}$, we must have $\mathcal{B}(\mathbb{R}) \supset \sigma(\mathcal{O})$. On the other hand, since $\mathcal{O}$ — and therefore $\sigma(\mathcal{O})$ — also contains all open intervals, which generate $\mathcal{B}(\mathbb{R})$, we get that $\mathcal{B}(\mathbb{R}) \subset \sigma(\mathcal{O})$, and so we find altogether that $\mathcal{B}(\mathbb{R}) = \sigma(\mathcal{O})$: $\mathcal{B}(\mathbb{R})$ is generated by the family of all open subsets of $\mathbb{R}$. Since the complements of the sets in $\mathcal{O}$ precisely give the family of all closed subsets of $\mathbb{R}$, we see that $\mathcal{B}(\mathbb{R})$ is equal to the σ–algebra generated by the family of closed subsets of $\mathbb{R}$.

All this holds in the same way for the general case of $\Omega = \mathbb{R}^n$, $n \in \mathbb{N}$.

Let $A \subset \mathbb{R}^n$. Then the trace σ–algebra $A \cap \mathcal{B}(\mathbb{R}^n)$ is denoted by $\mathcal{B}(A)$, and it is called the *Borel σ–algebra over A*. As an exercise (not completely trivial!), the reader may want to prove that $\mathcal{B}(A)$ coincides with the σ–algebra generated by the family $A \cap \mathcal{O}_n$, i.e., the open sets of A, where $\mathcal{O}_n$ denotes the open sets in $\mathbb{R}^n$. $\diamond$

Motivated by the last example we make the

A.8 Definition Suppose that $(\Omega, \mathcal{T})$ is a topological space. Then $\sigma(\mathcal{T})$ is called the *Borel σ–algebra* of $(\Omega, \mathcal{T})$.

A.9 Exercise Show that for all $B \in \mathcal{B}(\mathbb{R})$, $h \in \mathbb{R}$, $B + h \in \mathcal{B}(\mathbb{R})$. (*Hint:* Prove that the family of sets in $\mathcal{B}(\mathbb{R})$ for which the statement holds true is a σ–algebra containing any of the generators of $\mathcal{B}(\mathbb{R})$.)

A.10 Exercise Show that for $n \in \mathbb{N}$, the family $\mathcal{B}(\mathbb{R}) \times \cdots \times \mathcal{B}(\mathbb{R})$ (n factors) is a generator of $\mathcal{B}(\mathbb{R}^n)$. (*Hint:* Show first that for each $i \in \{1, \ldots, n\}$, sets of the form $\mathbb{R} \times \cdots \times \mathbb{R} \times B_i \times \mathbb{R} \times \cdots \times \mathbb{R}$, belong to $\mathcal{B}(\mathbb{R}^n)$. To this end, show that for each i the system of such sets forms a σ–algebra over $\mathbb{R}^n$.)

A.2 Dynkin Systems, Monotone Class Theorem

This section is a shortened version of chapter 2 in [27]. The interested reader can find more results, more details, and proofs there (more exercises as well!).

We have already seen that interesting σ–algebras are often constructed in a somewhat indirect, abstract fashion. As a consequence, it is often not quite straightforward to check whether a certain statement holds true for all sets in a given σ–algebra. In this section we introduce some powerful tools for such purposes.

A.11 Definition A non-empty family $\mathcal{S}$ of subsets of Ω is called a

(a) *π–system*, if $\mathcal{S}$ is $\cap$–stable;

(b) *d-system*, if the following hold true

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

  (ii) $A, B \in \mathcal{S}, A \subset B \Rightarrow B \setminus A \in \mathcal{S}$,

  (iii) if $(A_n, n \in \mathbb{N})$ is a sequence in $\mathcal{S}$ which increases to A, then $A \in \mathcal{S}$.

A.12 Remark Sometimes d-systems are also called *Dynkin systems* or *monotone systems*.

A.13 Definition If $\mathcal{M} \subset \mathcal{P}(\Omega)$, $d(\mathcal{M})$ denotes the smallest d-system containing $\mathcal{M}$, and it is called the *d-system generated by $\mathcal{M}$*.

A.14 Lemma *$\mathcal{S}$ is a d-system over Ω, if and only if the following hold true:*

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

(ii) $A \in \mathcal{S} \Rightarrow \complement A \in \mathcal{S}$,

(iii) *for every pairwise disjoint sequence $(A_n, n \in \mathbb{N})$ in $\mathcal{S}$, $\biguplus_n A_n \in \mathcal{S}$.*

A.15 Theorem *$\mathcal{S}$ is a σ–algebra over Ω, if and only if $\mathcal{S}$ is a π–system and a d-system.*

A.16 Theorem (Monotone Class Theorem) *Suppose that $\mathcal{C}$ is a π–system, and that $\mathcal{S}$ is a d-system such that $\mathcal{C} \subset \mathcal{S}$. Then $\sigma(\mathcal{C}) \subset \mathcal{S}$ holds true.*

A.3 Set Functions, Measures

A.17 Definition Assume that $\mathcal{S} \subset \mathcal{P}(\Omega)$ with $\emptyset \in \mathcal{S}$.

(a) A mapping μ from $\mathcal{S}$ into $[0, +\infty]$ with $\mu(\emptyset) = 0$ is called a *set function* on $\mathcal{S}$.

(b) A set function μ on $\mathcal{S}$ is called *additive*, if the following holds true: Whenever $A_1, \ldots, A_n \in \mathcal{S}$ are pairwise disjoint, and such that their union belongs to $\mathcal{S}$, then

$$\mu\Big(\biguplus_{i=1}^{n} A_i\Big) = \sum_{i=1}^{n} \mu(A_i). \tag{A.1}$$

(c) A set function μ on $\mathcal{S}$ is called *countably* (or σ–) *additive*, if the following holds true: If $(A_n,\ n \in \mathbb{N})$ is a pairwise disjoint sequence in $\mathcal{S}$, and such that the union $\uplus_n A_n$ belongs to $\mathcal{S}$, then

$$\mu\Big(\biguplus_{n=1}^{\infty} A_n\Big) = \sum_{n=1}^{\infty} \mu(A_n). \tag{A.2}$$

(d) A set function μ on $\mathcal{S}$ is called *σ–finite*, if there exists a sequence $(A_n,\ n \in \mathbb{N})$ in $\mathcal{S}$ whose union is equal to Ω, and which is such that $\mu(A_n) < +\infty$ for every $n \in \mathbb{N}$. If $\mu(A) < +\infty$ for all $A \in \mathcal{S}$, then μ is called *finite* on $\mathcal{S}$.

(e) If μ is a countably additive set function on a σ–algebra $\mathcal{A}$ over Ω, then μ is called a *measure* on $(\Omega, \mathcal{A})$, and $(\Omega, \mathcal{A}, \mu)$ is called a *measure space*.

(f) If $(\Omega, \mathcal{A}, \mu)$ is a measure space such that $\mu(\Omega) = 1$, then μ is called a *probability measure*, and $(\Omega, \mathcal{A}, \mu)$ is called a *probability space*.

A.18 Remark Often what in (a) above is called simply a "set function" is named more precisely a *positive* (or — according to the convention used — *non-negative*) *set function*. Similarly, in (b) *finitely additive* instead of simply "additive" would be more precise. Since there will be no danger of confusion for the present purposes we shall stick with the simpler namings. Moreover, sometimes a countably additive set function which is defined on a ring is called a *pre-measure*.

Now we can sketch the basic procedure of measure theory. Usually, often based on heuristic and/or geometric considerations, one defines some notion of size for sets in a semiring, such as for rectangles or triangles. Then one tries to extend this basic set function to a measure on the σ–algebra generated by that semiring. Typically, one considers the generated ring as an intermediate step, proves that the given basic set function has an extension to a *countably* additive set function. Most of the time, the latter is the most difficult step. Once one has this, one can apply a result of general nature, such as Carathédory's theorem which we will discuss in the next section, to conclude that there exists an extension as a measure (which often will be unique) to the generated σ–algebra.

A.19 Examples

(a) Let Ω be as above, and let $\mathcal{A}$ be any σ–algebra over Ω. For any $\omega_0 \in \Omega$, and $A \in \mathcal{A}$ set

$$\varepsilon_{\omega_0}(A) = \begin{cases} 1, & \text{if } \omega_0 \in A, \\ 0, & \text{otherwise.} \end{cases}$$

It is easy to see that ε_{ω_0} is a measure on $(\Omega, \mathcal{A})$, called the *Dirac measure in the point* ω_0. Obviously ε_{ω_0} is a finite measure.

(b) Let $\Omega = \mathbb{N}$, $\mathcal{A} = \mathcal{P}(\mathbb{N})$, and for $A \subset \mathbb{N}$ set

$$\mu(A) = \text{number of elements in } A.$$

Again in this case it is not hard to show that μ defines a measure on $(\mathbb{N}, \mathcal{P}(\mathbb{N}))$, called the *counting measure*. It is not finite, but it is σ–finite.

(c) Let $\Omega = \mathbb{R}^n$, $n \in \mathbb{N}$, and choose $\mathcal{S}$ as the semiring $\mathcal{I}^n$ of all right semiclosed intervals. For $(a,b] \in \mathcal{I}^n$ define

$$\lambda\big((a,b]\big) = \prod_{i=1}^{n}(b_i - a_i), \tag{A.3}$$

that is, we associate with the hyperrectangle $(a,b]$ its usual volume. It is obvious, that λ defines a set function on $\mathcal{S}$ in the sense of definition A.17 (a). Now consider the ring $\mathcal{R}(\mathcal{I}^n)$ generated by $\mathcal{I}^n$. As mentioned above, this ring is given by the family of all subsets in $\mathbb{R}^n$ which can be written as a finite pairwise disjoint union of sets in $\mathcal{I}^n$. Therefore it is also clear how one can extend λ to $\mathcal{R}(\mathcal{I}^n)$ to become an additive set function thereon. However, it is not at all obvious whether this extension is *countably* additive (it is!) — we will come back to this in the next subsection.

Some elementary properties of additive set functions and measures are given in the following

A.20 Theorem

(a) *Suppose that μ is an additive set function on a ring $\mathcal{R}$. Let A, $B \in \mathcal{R}$ be such that $\mu(A \cap B) < +\infty$, then*

$$\mu(A \cup B) = \mu(A) + \mu(B) - \mu(A \cap B). \tag{A.4}$$

μ is increasing*: If $A \subset B$, then $\mu(A) \le \mu(B)$ holds true. μ is* subtractive *in the sense that $A \subset B$ and $\mu(A) < +\infty$ entail*

$$\mu(B \setminus A) = \mu(B) - \mu(A). \tag{A.5}$$

Moreover, for all $A_1, \dots, A_n$, $n \in \mathbb{N}$, the inequality

$$\mu\Big(\bigcup_{i=1}^{n} A_i\Big) \le \sum_{i=1}^{n} \mu(A_i) \tag{A.6}$$

holds.

(b) *Suppose that μ is a measure on the measurable space $(\Omega, \mathcal{A})$. Then for every sequence $(A_n, n \in \mathbb{N})$ in $\mathcal{A}$ the inequality*

$$\mu\Big(\bigcup_{n=1}^{\infty} A_n\Big) \leq \sum_{n=1}^{\infty} \mu(A_n) \tag{A.7}$$

holds true.

A very useful application of the monotone class theorem A.16 is the following result, whose proof can be found, e.g., in chapter 2 of [27]:

A.21 Theorem *Suppose that μ_1 and μ_2 are two probability measures on $(\Omega, \mathcal{A})$, and that $\mathcal{E}$ is a $\cap$–stable generator of $\mathcal{A}$, such that μ_1 and μ_2 coincide on $\mathcal{E}$. Then $\mu_1 = \mu_2$.*

A.22 Remark Suppose that μ is an additive set function on an algebra $\mathcal{A}$. Then the fact that μ is increasing (see (a) above) entails that μ is finite on $\mathcal{A}$, if and only if $\mu(\Omega) < +\infty$. In particular every probability measure μ on a measurable space $(\Omega, \mathcal{A})$ is finite on $\mathcal{A}$.

Assume that $(A_n, n \in \mathbb{N})$ is an increasing, respectively decreasing, sequence of subsets of Ω. We set

$$\lim_n A_n = \begin{cases} \bigcup_{n\in\mathbb{N}} A_n, & \text{if } (A_n, n \in \mathbb{N}) \text{ is increasing,} \\ \bigcap_{n\in\mathbb{N}} A_n, & \text{if } (A_n, n \in \mathbb{N}) \text{ is decreasing.} \end{cases}$$

If $A = \lim_n A_n$ in either of these cases, we also simply write $A_n \uparrow A$ for the case that $(A_n, n \in \mathbb{N})$ is increasing, and $A_n \downarrow A$ for the case that this sequence decreases. If μ is a set function defined on $\mathcal{S}$, then μ is called *continuous from below (above, resp.) in $A \in \mathcal{S}$*, if for every sequence $(A_n, n \in \mathbb{N})$ of sets in $\mathcal{S}$ which increases (decreases, resp.) to A we have

$$\mu(\lim_n A_n) = \lim_n \mu(A_n). \tag{A.8}$$

If μ is continuous from below (above, resp.) in every $A \in \mathcal{S}$, then μ is simply called *continuous from below (above, resp.).*

A.23 Theorem *Suppose that μ is an additive set function on a ring $\mathcal{R}$.*

(a) *μ is countably additive, if and only if μ is continuous from below.*

(b) *μ is continuous from above, if and only if μ is continuous from above in the empty set. If μ is countably additive, then μ is continuous in $\emptyset$ (from above).*

(c) *If μ is finite on $\mathcal{R}$, then countable additivity, continuity from above, from below, and continuity in $\emptyset$ are equivalent.*

A.4 Extension Problem, Carathédory's Theorem

Suppose that we are given a set function μ on a semiring $\mathcal{S}$ (cf. definition A.17 (a)). Since we know already that $\mathcal{R}(\mathcal{S})$ consists of all subsets $A \subset \Omega$ which can be written as

$$A = \biguplus_{i=1}^{n} A_i, \tag{A.9}$$

for some pairwise disjoint $A_1, \ldots, A_n \in \mathcal{S}$, $n \in \mathbb{N}$, it is easy to extend μ to $\mathcal{R}(\mathcal{S})$ via the recipe (for simplicity we denote the extension again by μ)

$$\mu(A) = \sum_{i=1}^{n} \mu(A_i). \tag{A.10}$$

It is not hard (though somewhat boring) to check, that this extension is well-defined (in general it is possible that $A \in \mathcal{R}(\mathcal{S})$ has different representations of the form (A.9)), and that μ is additive on $\mathcal{R}(\mathcal{S})$. We want to extend μ now further to $\sigma(\mathcal{S}) = \sigma(\mathcal{R}(\mathcal{S}))$ as a measure $\hat{\mu}$, i.e., $\hat{\mu}$ should be countably additive on $\sigma(\mathcal{R}(\mathcal{S}))$. But since the restriction of $\hat{\mu}$ to $\mathcal{R}(\mathcal{S})$ has to coincide with μ, we find that μ had to be countably additive on $\mathcal{R}(\mathcal{S})$ in order that this be possible. Therefore, the right way to put the extension problem is the following: Suppose that we are given a countably additive set function μ on a ring $\mathcal{R}$. Does there exist a countably additive set function $\hat{\mu}$ on $\sigma(\mathcal{R})$ whose restriction to $\mathcal{R}$ is equal to μ? And if it does, is it unique? The question of existence is answered by the following

A.24 Theorem (Carathéodory) *Every countably additive set function μ on a ring $\mathcal{R}$ has a countably additive extension to the σ–algebra $\sigma(\mathcal{R})$.*

The proof relies on a rather ingenious idea of Carathédory which we are going to sketch now.

First one defines the following set function μ^* on $\mathcal{P}(\Omega)$: For $A \subset \Omega$ which is such that there is no sequence in $\mathcal{R}$ whose union covers A we set $\mu^*(A) = +\infty$. Otherwise, we define

$$\mu^*(A) = \inf\Big\{\sum_{n=1}^{\infty} \mu(A_n),\ (A_n,\ n \in \mathbb{N}) \text{ is a cover of } A \text{ in } \mathcal{R}\Big\}. \tag{A.11}$$

One proves that μ^* is an *outer measure* on $\mathcal{P}(\Omega)$: μ^* is an increasing set function on $\mathcal{P}(\Omega)$ so that $\mu^*(\cup_n A_n) \leq \sum_n \mu^*(A_n)$ for every sequence $(A_n,\ n \in \mathbb{N})$ in $\mathcal{P}(\Omega)$.

Now comes the remarkable observation of Carathéodory: Define

$$\mathcal{A}(\mu^*) = \{A \in \mathcal{P}(\Omega),\ \mu^*(Q) = \mu^*(Q \cap A) + \mu^*(Q \cap \complement A),\ \forall Q \in \mathcal{P}(\Omega)\}. \tag{A.12}$$

Then one can show: a) $\mathcal{A}(\mu^*)$ is a σ–algebra, b) μ^* restricted to $\mathcal{A}(\mu^*)$ is a measure, c) $\mathcal{A}(\mu^*)$ contains $\sigma(\mathcal{R})$, d) μ^* restricted to $\mathcal{R}$ is equal to μ. Thus one has proved theorem A.24.

For the question of uniqueness one proves

A.25 Theorem *Suppose that μ is as in the hypothesis of theorem A.24. If μ is σ–finite on $\mathcal{R}$, then the extension of theorem A.24 is unique.*

A.26 Example We continue the discussion of example A.19 (a). For convenience of the notation we specialize to the case $n = 1$, but at the same time we consider the more general case where we are given a real valued, increasing (not necessarily *strictly* increasing) function α on $\mathbb{R}$ which is continuous from the right. The case of example A.19 (a) (for $n = 1$) is then given by the choice $\alpha(x) = x$, $x \in \mathbb{R}$. Thus we assign the value

$$\mu_\alpha((a, b]) = \alpha(b) - \alpha(a) \tag{A.13}$$

to the interval $(a, b]$. μ_α is extended to the ring $\mathcal{R}(\mathcal{I})$ consisting all finite pairwise disjoint unions of right semiclosed intervals in the same way as above, see (A.10). In view of theorem A.24 we only have to prove that μ_α is countably additive on $\mathcal{R}(\mathcal{I})$. Once we have shown that, it follows that μ_α has an extension to a measure on $\mathcal{B}(\mathbb{R})$. This extension is unique by theorem A.25, because μ_α is σ–finite on $\mathcal{R}(\mathcal{I})$. Moreover, since μ_a is finite on $\mathcal{R}(\mathcal{I})$, we can apply theorem A.23 (c), and therefore it is sufficient to show that μ_α is continuous in $\emptyset$. Continuity from below is trivial, so that we have reduced our problem to show that if $(A_n,\, n \in \mathbb{N})$ is a decreasing sequence in $\mathcal{R}(\mathcal{I})$ with $A_n \downarrow \emptyset$, then $\mu_\alpha(A_n) \to 0$, as $n \to +\infty$. We only sketch the argument. Suppose now that $A_n \downarrow \emptyset$, but $\lim_n \mu_\alpha(A_n) = \inf_n \mu_\alpha(A_n) = \varepsilon > 0$. We are heading for a contradiction. Every A_n is the union of finitely many pairwise disjoint intervals of the form $(a, b]$. By shrinking each of these intervals a bit we can use the continuity of α to find in each of them a *closed* interval without changing the size too much, i.e., for each $n \in \mathbb{N}$ we can find a compact subset C_n of A_n, so that $\mu_\alpha(C_n) > 0$. Moreover, one can arrange this in such a way that the sequence $(C_n,\, n \in \mathbb{N})$ of compact subsets is decreasing, too, and since $\mu_\alpha(C_n) > 0$, we have that they are all non-empty. This means that every finite intersection of sets in this sequence is non-empty. Now the finite intersection property of compact sets implies that their intersection is non-empty: $\cap_n A_n \supset \cap_n C_n \neq \emptyset$. Thus $(A_n,\, n \in \mathbb{N})$ cannot decrease to $\emptyset$: contradiction!

Thus we obtain the unique existence of a measure on $\mathcal{B}(\mathbb{R})$, denoted again by μ_α, which assigns to every interval $(a, b]$, $a \leq b$, the size (A.13). It is called the *Lebesgue–Stieljes–measure* defined by α. For the case $\alpha = \mathrm{id}$, it is called the *Lebesgue measure*, it is denoted by λ. $\diamond$

A.27 Exercise Show that for every $x \in \mathbb{R}$, $\lambda(\{x\}) = 0$. Conclude that every countable subset of $\mathbb{R}$ has Lebesgue measure zero, and in particular, $\lambda(\mathbb{Q}) = 0$. Moreover, for all $a, b \in \mathbb{R}$, $a \leq b$, $\lambda((a, b)) = \lambda([a, b)) = \lambda([a, b]) = b - a$.

A.5 Measurable Mappings, Image Measures

Assume that Ω_1 is a non-empty set, that $(\Omega_2, \mathcal{A}_2)$ is a measurable space, and that T is a mapping from Ω_1 to Ω_2. We define $\sigma(T) = T^{-1}(\mathcal{A}_2)$, called the *$\sigma$–algebra generated by T*.

A.28 Exercise Show that $\sigma(T)$ is a σ–algebra over Ω_1.

A.29 Definition Assume that $(\Omega_i, \mathcal{A}_i)$, $i = 1, 2$, are two measurable spaces, and that T is a mapping from Ω_1 into Ω_2. T is called $\mathcal{A}_1$–$\mathcal{A}_2$*–measurable*, if and only if $\sigma(T) \subset \mathcal{A}_1$.

A.30 Remark If the σ–algebras are understood from the context we shall just say "T is measurable". $\sigma(T)$ is the smallest σ–algebra over Ω_1 making the map T measurable: If $\mathcal{A}_1' \subset \sigma(T)$ and T is $\mathcal{A}_1'/\mathcal{A}_2$–measurable then $\mathcal{A}_1' = \sigma(T)$.

The proofs of the following two theorems are easy *exercises*:

A.31 Theorem *Suppose that $(\Omega_1, \mathcal{A}_1, \mu)$ is a measure space, and that $(\Omega_2, \mathcal{A}_2)$ is a measurable space. Assume furthermore that T is a measurable mapping from Ω_1 into Ω_2. For $A \in \mathcal{A}_2$ set*

$$T\mu(A) = \mu \circ T^{-1}(A) = \mu\bigl(T^{-1}(A)\bigr). \tag{A.14}$$

Then $T\mu$ is a measure on $(\Omega_2, \mathcal{A}_2)$, called the image measure of μ under T.

Notation Sometimes it will be convenient to denote the image measure $T\mu$ of μ under T by μ_T.

A.32 Theorem *Suppose that $(\Omega_i, \mathcal{A}_i)$, $i = 1, 2, 3$, are measurable spaces, and that T_i, $i = 1, 2$, is a measurable mapping from Ω_i into Ω_{i+1}. Then $T_2 \circ T_1$ is $\mathcal{A}_1$–$\mathcal{A}_3$–measurable.*

A.33 Lemma *Suppose that Ω_1 is a non-empty set, that $(\Omega_2, \mathcal{A}_2)$ is a measurable space, and that $\mathcal{C}$ a generator of $\mathcal{A}_2$: $\sigma(\mathcal{C}) = \mathcal{A}_2$. Let T be a mapping from Ω_1 into Ω_2. Then $\sigma(T) = \sigma\bigl(T^{-1}(\mathcal{C})\bigr)$, i.e., $T^{-1}(\mathcal{C})$ is a generator of $\sigma(T)$.*

Proof First we show the inclusion $\sigma(T) \supset \sigma(T^{-1}(\mathcal{C}))$: By definition, $\sigma(T) = T^{-1}(\mathcal{A}_2)$, so that $\sigma(T) \supset T^{-1}(\mathcal{C})$. Since $\sigma(T^{-1}(\mathcal{C}))$ is the smallest σ–algebra containing $T^{-1}(\mathcal{C})$, we find that $\sigma(T) \supset \sigma(T^{-1}(\mathcal{C}))$. For the converse inclusion set

$$\mathcal{S} = \bigl\{A \subset \Omega_2,\ T^{-1}(A) \in \sigma\bigl(T^{-1}(\mathcal{C})\bigr)\bigr\}.$$

By construction we have that $\mathcal{C} \subset \mathcal{S}$. On the other hand, an easy *exercise* shows that $\mathcal{S}$ is a σ–algebra over Ω_2. Therefore $\mathcal{S} \supset \sigma(\mathcal{C}) = \mathcal{A}_2$. As a consequence we find that $T^{-1}(\mathcal{A}_2) \subset \sigma(T^{-1}(\mathcal{C}))$, that is, $\sigma(T) \subset \sigma(T^{-1}(\mathcal{C}))$. □

Now suppose that $(\Omega_i, \mathcal{A}_i)$, $i = 1, 2$, are measurable spaces, that $\mathcal{C}$ is a generator of $\mathcal{A}_2$, and that T is a mapping from Ω_1 to Ω_2, such that $T^{-1}(\mathcal{C}) \subset \mathcal{A}_1$. It follows that $\sigma(T^{-1}(\mathcal{C})) \subset \mathcal{A}_1$, because $\sigma(T^{-1}(\mathcal{C}))$ is the smallest σ–algebra containing $T^{-1}(\mathcal{C})$. From lemma A.33 we get that $\sigma(T) \subset \mathcal{A}_1$, i.e., T is measurable. Conversely, if T is measurable, then obviously $T^{-1}(\mathcal{C}) \subset \mathcal{A}_1$. Thus we have proved the

A.34 Theorem *Suppose that $(\Omega_i, \mathcal{A}_i)$, $i = 1, 2$, are two measurable spaces, that $\mathcal{C}$ is a generator of $\mathcal{A}_2$, and that T is a mapping from Ω_1 into Ω_2. T is $\mathcal{A}_1$–$\mathcal{A}_2$–measurable, if and only if $T^{-1}(\mathcal{C}) \subset \mathcal{A}_1$.*

From theorem A.34 we immediately get

A.35 Corollary *Consider two topological spaces $(\Omega_i, \mathcal{T}_i)$, $i = 1, 2$, and let $\mathcal{A}_i$ be the Borel–σ–algebras: $\mathcal{A}_i = \sigma(\mathcal{T}_i)$, $i = 1, 2$. Then every continuous mapping from Ω_1 into Ω_2 is measurable.*

A.36 Exercise Choose $(\Omega_1, \mathcal{A}_1) = (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$, $(\Omega_2, \mathcal{A}_2) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$, and for $i \in \{1, \dots, n\}$ let π_i denote the canonical projection from $\mathbb{R}^n$ onto $\mathbb{R}$: $\pi_i(x) = x_i$, $x = (x_1, \dots, x_n) \in \mathbb{R}^n$. Show that π_i is measurable. (*Hint:* Show that π_i is continuous.) Redo exercise A.10 using the measurability of $\pi_1, \dots, \pi_n$.

For families of mappings we make the following

A.37 Definition Assume that I is a non-empty index set, and that $(T_i,\, i \in I)$ is a family of mappings, such that T_i, $i \in I$, maps Ω into Ω_i. Suppose furthermore that for every $i \in I$, $(\Omega_i, \mathcal{A}_i)$ is a measurable space. $\sigma(T_i,\, i \in I)$ is the smallest σ–algebra over Ω which makes all mappings T_i, $i \in I$, measurable.

A.38 Exercise Show that

$$\sigma(T_i,\, i \in I) = \sigma\Big(\bigcup_{i \in I} T_i^{-1}(\mathcal{A}_i)\Big)$$

We shall now specialize to real valued functions. From our discussion of the Borel σ–algebra $\mathcal{B}(\mathbb{R})$ in example A.7 we immediately get

A.39 Lemma *Suppose that f is a real valued function on $(\Omega, \mathcal{A})$. It is measurable, if and only if for all $x \in \mathbb{R}$, any and therefore all of the following inclusions hold true:*

$$\{f \le x\} \in \mathcal{A}, \quad \{f < x\} \in \mathcal{A}, \quad \{f \ge x\} \in \mathcal{A}, \quad \{f > x\} \in \mathcal{A}.$$

Above we made use of the shorthand

$$\{f \le x\} = \{\omega \in \Omega,\ f(\omega) \le x\},$$

and similarly for the other inclusions. We shall make use of this notation throughout.

Based on lemma A.39 one can now step by step prove the following results, the proofs are left as *exercises* to the interested reader.

A.40 Theorem *Suppose that f and g are measurable functions on $(\Omega, \mathcal{A})$.*

(a) *The following sets belong to $\mathcal{A}$: $\{f \le g\}$, $\{f < g\}$, $\{f = g\}$, $\{f \ne g\}$.*

(b) *The functions $f \pm g$, $f \cdot g$, $f \wedge g$, $f \vee g$ are measurable.*

(c) *The functions* $|f|$, $f^{\pm}$ *are measurable.*

A.41 Theorem *Assume that* $(f_n,\ n \in \mathbb{N})$ *is a sequence of measurable functions.*

(a) $\sup_n f_n$, $\inf_n f_n$, $\limsup_n f_n$, $\liminf_n f_n$ *are measurable.*

(b) *If it exists, the pointwise limit of* $(f_n,\ n \in \mathbb{N})$ *is measurable.*

A.6 Lebesgue Measure, Vitali Set

Similarly as in example A.26, one can construct the *Lebesgue measure* in the n–dimensional case, i.e., on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$, as the unique extension of the prescription (A.3) to $\mathcal{B}(\mathbb{R}^n)$ (with an obvious intermediate step passing $\mathcal{R}(\mathcal{I}^n)$). Also in this case the Lebesgue measure will be simply denoted by λ. The Lebesgue measure λ on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ inherits important invariance properties from its predecessor A.3 on the right semiclosed intervals. First we discuss this for $n = 1$: Let $h \in \mathbb{R}$, and consider the operation of translation τ_h by h: $x \mapsto x + h$, $x \in \mathbb{R}$. Clearly, this is a continuous and therefore measurable mapping (see corollary A.35) from $\mathbb{R}$ into itself. By theorem A.31 it defines another measure $\lambda_h = \tau_h \lambda$ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Since the length of an interval does not change under translation, we find that λ and λ_h coincide on $\mathcal{I}$. Since $\mathcal{R}(\mathcal{I})$ consists of pairwise disjoint finite unions of elements in $\mathcal{I}$, λ and λ_h also coincide on $\mathcal{R}(\mathcal{I})$. We know that λ is σ–finite on $\mathcal{R}(\mathcal{I})$ (even on $\mathcal{I}$), so that we can apply theorem A.25 to conlude that $\lambda = \lambda_h$. Thus λ is invariant under translations.

Now we can prove that $\mathcal{B}(\mathbb{R}) \subsetneq \mathcal{P}(\mathbb{R})$, that is, there exist subsets of $\mathbb{R}$ which do not belong to $\mathcal{B}(\mathbb{R})$.

A.42 Example (Vitali Set) Consider the set $[0, 1]$. Let us say that $x, y \in [0, 1]$ are equivalent, denoted by $x \sim y$, if $x - y \in \mathbb{Q}$. It is easy to see that $\sim$ is indeed an equivalence relation. Thus under $\sim$ the interval $[0, 1]$ is decomposed into its equivalence classes, denoted by $[x]$, $x \in [0, 1]$. Define a subset $V \subset [0, 1]$ by choosing exactly one representative from each of these equivalence classes.[1] Let assume that $V \in \mathcal{B}(\mathbb{R})$ and bring this assumption to a contradiction.

Define a set $\hat{V}$ by

$$\hat{V} = \biguplus_{q \in [-1,1]_{\mathbb{Q}}} (V + q),$$

where $[-1, 1]_{\mathbb{Q}}$ denotes the set of rational numbers in $[-1, 1]$. Our assumption entails that $\hat{V} \in \mathcal{B}(\mathbb{R})$ (cf. exercise A.9). Now for $q, q' \in [-1, 1]_{\mathbb{Q}}$, $q \neq q'$, we have that $V + q$ and $V + q'$ are disjoint: Suppose that $x \in (V + q) \cap (V + q')$, that is, $x - q \in V$, and $x - q' \in V$. Since $(x - q) - (x - q') = q' - q$ is rational, this entails that $x - q$ and $x - q'$ are equivalent, but by construction of V, there is only one representative of the class $[x - q]$ in V. Thus such an x cannot exist. Next we observe that $[0, 1] \subset \hat{V}$. Indeed, let $x \in [0, 1]$, consider its equivalence class $[x]$,

[1] Here we make use of the *axiom of choice*, which the author *always* assumes to hold true. In fact, the validity of the axiom of choice is crucial for this example.

and let y denote its representative in V. Then $x - y = q \in [-1,1]_{\mathbb{Q}}$, and therefore $x = y + q \in \hat{V}$. By construction we have $\hat{V} \subset [-1,2]$. The fact that λ is increasing (see theorem A.20.(a)) implies that $1 \leq \lambda(\hat{V}) \leq 3$. On the other hand, from the countable additivity of λ we get

$$\lambda(\hat{V}) = \lambda\Big(\biguplus_{q\in[-1,1]_{\mathbb{Q}}} (V+q)\Big) = \sum_{q\in[-1,1]_{\mathbb{Q}}} \lambda(V+q) = \sum_{q\in[-1,1]_{\mathbb{Q}}} \lambda(V),$$

where we used the invariance of λ under translations in the last equality. The assumption $\lambda(V) = 0$ is in contradiction with $0 = \lambda(\hat{V}) \geq 1$, while the assumption $\lambda(V) > 0$ entails the contradiction $3 \geq \lambda(\hat{V}) = +\infty$. $\diamond$

We return to the general case of the Lebesgue measure λ on $(\mathbb{R}^n, \mathcal{B}(R^n))$, $n \in \mathbb{N}$. A linear (therefore continuous and hence measurable) mapping ρ from $\mathbb{R}^n$ into itself, which leaves the Euclidean metric of $\mathbb{R}^n$ invariant, is called a *Euclidean motion*. The set of all Euclidean motions forms a group, called the *Euclidean group*, and every element ρ in this group can be represented as a rotation (i.e., an element in the orthogonal group $O(n)$), followed by a translation. As in the case $n = 1$, each Euclidean motion ρ defines a new measure $\lambda_\rho = \rho\lambda$. λ and λ_ρ coincide on $\mathcal{I}^n$. Similarly as above it follows that $\lambda = \lambda_\rho$. That is, λ is invariant under the Euclidean group, and in particular under translations. It is interesting and important to observe that — up to constant — the Lebesgue measure is *characterized* by the last property, namely one has the following

A.43 Theorem *The Lebesgue measure λ on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ is invariant under the Euclidean group. Every measure on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ which is translation invariant is a constant multiple of the Lebesgue measure λ. In particular, every translation invariant measure on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ which assigns 1 to the unit hypercube $[0,1]^n$ is equal to the Lebesgue measure.*

For the proof of this theorem, and more invariance and transformation properties of the Lebesque measure, the interested reader is referred to the pertinent literature, for example to [4, § 8].

We conclude this section with the following remark. Suppose that $A \in \mathcal{B}(\mathbb{R}^n)$. We have already defined the trace σ–algebra $\mathcal{B}(A)$ of $\mathcal{B}(\mathbb{R}^n)$ on A (see exercise A.3 (c)). All sets in $\mathcal{B}(A)$ belong to $\mathcal{B}(\mathbb{R}^n)$, so that it makes sense to speak of the restriction of λ to $\mathcal{B}(A)$. This restriction defines a measure on $\mathcal{B}(A)$, and we shall call it again *Lebesgue measure*, and denote it by λ.

A.7 Lebesgue Integral

In this section we construct the Lebesgue integral. Throughout we consider a given measure space $(\Omega, \mathcal{A}, \mu)$. It will be convenient and useful that we consider instead of only real valued functions the larger set of *numerical functions*. That is, functions which are also allowed to take the values $\pm\infty$. We write

$$\begin{aligned}\overline{\mathbb{R}} &= \mathbb{R} \cup \{-\infty, +\infty\} = [-\infty, +\infty],\\ \overline{\mathbb{R}}_+ &= \mathbb{R}_+ \cup \{+\infty\} = [0, +\infty],\\ \overline{\mathbb{R}}_- &= \mathbb{R}_- \cup \{-\infty\} = [-\infty, 0].\end{aligned}$$

The Borel σ–algebra $\mathcal{B}(\overline{\mathbb{R}})$ on $\overline{\mathbb{R}}$ is simply given by $\sigma(\mathcal{B}(\mathbb{R}), \{\pm\infty\})$, and similarly for $\mathcal{B}(\overline{\mathbb{R}}_\pm)$. Thus, for example $A \in \mathcal{B}(\overline{\mathbb{R}})$, if and only if either $A \in \mathcal{B}(\mathbb{R})$ or A is of one of the forms $A' \cup \{-\infty, +\infty\}$, $A' \cup \{-\infty\}$, $A' \cup \{+\infty\}$ with $A' \in \mathcal{B}(\mathbb{R})$. The discussion of measurability of numerical functions is practically the same as for real valued functions, the necessary modifications being most of the time obvious. [2] In order to avoid clumsy notation (and clumsy ways of speaking), we shall not distinguish between real valued and numerical functions, unless there is danger of confusion.

The construction of the Lebesgue integral has three steps, the first is to define the integral for elementary functions which we define now:

A.44 Definition An *elementary function* is a measurable function with values in $\overline{\mathbb{R}}_+$ which takes only finitely many values. $\mathcal{E}(\Omega, \mathcal{A})$ denotes the set of all elementary functions on $(\Omega, \mathcal{A})$.

Let $f \in \mathcal{E}(\Omega, \mathcal{A})$, and suppose that $\alpha_1, \dots, \alpha_n$, $n \in \mathbb{N}$, are its values. Set $A_i = f^{-1}(\{\alpha_i\})$, $i = 1, \dots, n$. Then $A_1, \dots, A_n$ form a decomposition of Ω in $\mathcal{A}$. Moreover,

$$f = \sum_{i=1}^{n} \alpha_i \, 1_{A_i}. \tag{A.15}$$

However, except in trivial cases, this representation of $f \in \mathcal{E}(\Omega, \mathcal{A})$ is not unique. For $f \in \mathcal{E}(\Omega, \mu)$ with a representation given by (A.15) set

$$\int_\Omega f \, d\mu = \sum_{i=1}^{n} \alpha_i \, \mu(A_i). \tag{A.16}$$

A routine exercise shows that the left hand side in A.16 does not depend on the representation of f, i.e., it is *well-defined.*

A.45 Definition The left hand side of equation (A.16) is called the *Lebesgue integral of the elementary function f (with respect to μ).*

It is straightforward to show that the Lebesgue integral, considered as a mapping from the positive cone $\mathcal{E}(\Omega, \mathcal{A})$ into $\mathbb{R}_+$, is linear and increasing.

[2] However, when forming sums and differences of numerical functions one has to take care not to subtract $+\infty$ from $+\infty$ etc. For the formation products we make throughout the convention $0 \cdot \pm\infty = \pm\infty \cdot 0 = 0$.

In the second step we take limits. We take advantage of the fact that a real increasing sequence which is bounded, converges to a limit. Thus, if one allows the "value" $+\infty$ — which we do — then every real increasing sequence converges. To make this work in our situation one first observes the following basic result:

A.46 Theorem *$\mathcal{M}_+(\Omega, \mathcal{A})$ denotes the positive cone of positive, measurable functions on Ω. Then for every $f \in \mathcal{M}_+(\Omega, \mathcal{A})$ there exists an increasing sequence $(f_n,\, n \in \mathbb{N})$ in $\mathcal{E}(\Omega, \mathcal{A})$ which converges pointwise to f.*

It is not hard to construct such a sequence $(f_n,\, n \in \mathbb{N})$ for given $f \in \mathcal{M}_+(\Omega, \mathcal{A})$: For $n \in \mathbb{N}$, decompose the interval $[0, n)$ into intervals of the form $[k/2^n, (k+1)/2^n)$, $k = 0, 1, \dots, n2^n - 1$. Then let $A_k^n \in \mathcal{A}$ denote the inverse image of these intervals under f, and adjoin the set $A_{n2^n}^n = \{f \geq n\} \in \mathcal{A}$. On A_k^n set $f = k/2^n$, while on $A_{n2^n}^n$ set $f = n$. The so constructed sequence fulfills (*exercise!*):

$$\lim_n f_n(\omega) = \sup f_n(\omega) = f(\omega), \qquad \omega \in \Omega. \tag{A.17}$$

For $f \in \mathcal{M}_+(\Omega, \mathcal{A})$ with increasing sequence $(f_n,\, n \in \mathbb{N})$ in $\mathcal{E}(\Omega, \mathcal{A})$ so that pointwise $f_n \uparrow f$ we set

$$\int_\Omega f \, d\mu = \lim \int_\Omega f_n \, d\mu = \sup \int_\Omega f_n \, d\mu. \tag{A.18}$$

Note that the right hand sides converge, because the facts that $(f_n,\, n \in \mathbb{N})$ is increasing, and that — as we remarked above — the Lebesgue integral is monotone increasing on $\mathcal{E}(\Omega, \mathcal{A})$ implies that the sequence $(\int_\Omega f_n \, d\mu,\, n \in \mathbb{N})$ is increasing. Again one has to check that the left hand side in (A.18) is well-defined.

A.47 Definition For every function $f \in \mathcal{M}_+(\Omega, \mathcal{A})$ the left hand side of formula (A.18) is called the *Lebesgue integral of f*.

It is not hard to see that the Lebesgue integral is linear and increasing as a mapping from the positive cone $\mathcal{M}_+(\Omega, \mathcal{A})$ into $\overline{\mathbb{R}}_+ = [0, +\infty]$.

Finally consider a general measurable function on $(\Omega, \mathcal{A})$. From theorem A.40 (c) we know that $f^\pm$ belong to $\mathcal{M}_+(\Omega, \mathcal{A})$. Therefore their Lebesgue integrals are already defined.

A.48 Definition A real valued or numerical function f on Ω is called *(μ–)integrable*, if f is measurable and $\int_\Omega f^\pm \, d\mu$ are both finite. The set of all μ–integrable functions is denoted by $\mathcal{L}^1(\Omega, \mathcal{A}, \mu)$.

In an *exercise* the reader is invited to check that a measurable function belongs to $\mathcal{L}^1(\Omega, \mathcal{A}, \mu)$, if and only if $|f|$ has a finite μ–integral.

A.49 Definition For $f \in \mathcal{L}^1(\Omega, \mathcal{A}, \mu)$ its *Lebesgue integral (with respect to μ)* is defined by

$$\int_\Omega f \, d\mu = \int_\Omega f^+ \, d\mu - \int_\Omega f^- \, d\mu. \tag{A.19}$$

A.50 Remarks The notion used here varies greatly in the literature, and for us it will be convenient, too, to change it here and there. First, it is quite common for the notation of $\mathcal{L}^1(\Omega, \mathcal{A}, \mu)$ to omit some of the "arguments", depending on what one wants to emphasize. Often we shall simply write $\mathcal{L}^1(\mu)$. For the integral in (A.19) one also finds the following notations:

$$\int_\Omega f(\omega)\, d\mu(\omega), \qquad \int_\Omega f(\omega)\, \mu(d\omega), \qquad (\mu, f), \qquad \langle \mu, f \rangle.$$

If f is positive, measurable or integrable, and A belongs to $\mathcal{A}$, then also the function $f\,1_A$ is positive, measurable or integrable, resp. One sets

$$\int_A f\, d\mu = \int_\Omega f\, 1_A\, d\mu. \tag{A.20}$$

If we write $\int f\, d\mu$, then we always mean $\int_\Omega f\, d\mu$. Suppose that $A = B \uplus C$, B, $C \in \mathcal{A}$. Then clearly $1_A = 1_B + 1_C$, so that we get the *additivity* of the Lebesgue integral: For positive, measurable or integrable f, and disjoint B, C in $\mathcal{A}$, we have

$$\int_{B \uplus C} f\, d\mu = \int_B f\, d\mu + \int_C f\, d\mu. \tag{A.21}$$

It is not difficult so see that the Lebesgue integral, considered as a mapping from $\mathcal{L}^1(\Omega, \mathcal{A}, \mu)$ to $\mathbb{R}$ is increasing, that is, if $f, g \in \mathcal{L}^1(\Omega, \mathcal{A}, \mu)$ with $f \leq g$, then

$$\int f\, d\mu \leq \int g\, d\mu.$$

The *triangle inequality*

$$\Big| \int f\, d\mu \Big| \leq \int |f|\, d\mu \tag{A.22}$$

follows for all $f \in \mathcal{L}^1(\Omega, \mathcal{A}, \mu)$. This again entails that $\mathcal{L}^1(\Omega, \mathcal{A}, \mu)$ is a vector space, and the Lebesgue integral acts linearly on this vector space.

We also need to discuss complex valued functions. To this end we remark that as a topological space $\mathbb{C}$ is nothing but $\mathbb{R}^2$ with its standard topology. Thus we can endow $\mathbb{C}$ with the Borel σ–algebra $\mathcal{B}(\mathbb{R}^2)$, for which we then write $\mathcal{B}(\mathbb{C})$. The set $\mathcal{L}^1_{\mathbb{C}}(\Omega, \mathcal{A}, \mu)$ of *complex valued μ–integrable functions f* is defined as the set of those functions f which are such that $\mathrm{Re}(f)$ and $\mathrm{Im}(f)$ belong to $\mathcal{L}^1(\Omega, \mathcal{A}, \mu)$. If there is no danger of confusion, we also simply write $\mathcal{L}^1(\mu)$ etc. for $\mathcal{L}^1_{\mathbb{C}}(\Omega, \mathcal{A}, \mu)$. The reader checks that $f \in \mathcal{L}^1_{\mathbb{C}}(\Omega, \mathcal{A}, \mu)$, if and only if $|f| \in \mathcal{L}^1(\Omega, \mathcal{A}, \mu)$. For $f \in \mathcal{L}^1_{\mathbb{C}}(\Omega, \mathcal{A}, \mu)$ we set

$$\int_\Omega f\, d\mu = \int_\Omega \mathrm{Re}(f)\, d\mu + i \int_\Omega \mathrm{Im}(f)\, d\mu, \tag{A.23}$$

and call it the *Lebesgue integral (with respect to μ)* of the integrable complex valued function f. Note that this definition implies

$$\mathrm{Re}\Big(\int f\, d\mu\Big) = \int \mathrm{Re}(f)\, d\mu, \qquad \mathrm{Im}\Big(\int f\, d\mu\Big) = \int \mathrm{Im}(f)\, d\mu, \tag{A.24}$$

and

$$\overline{\left(\int f\,d\mu\right)} = \int \overline{f}\,d\mu. \tag{A.25}$$

Similarly as above, we get that $\mathscr{L}^1_{\mathbb{C}}(\Omega, \mathcal{A}, \mu)$ is a (complex) vector space, and the Lebesgue integral acts (complex) linearly thereon. The triangle inequality in the form (A.22) holds also in the complex case, but its proof is a bit more involved: Suppose that $f \in \mathscr{L}^1_{\mathbb{C}}(\Omega, \mathcal{A}, \mu)$, and consider the estimation

$$\mathrm{Re}\Big(\big(\overline{\int f\,d\mu}\big)\,f\Big) \le \Big|\Big(\overline{\int f\,d\mu}\Big)\,f\Big| = \Big|\int f\,d\mu\Big|\,|f|.$$

With this and the linearity of the Lebesgue integral we can compute in the following way

$$\begin{aligned}
\Big|\int f\,d\mu\Big|^2 &= \mathrm{Re}\Big(\big(\overline{\int f\,d\mu}\big)\big(\int f\,d\mu\big)\Big)\\
&= \mathrm{Re}\Big(\int \big(\overline{\int f\,d\mu}\big)\,f\,d\mu\Big)\\
&= \int \mathrm{Re}\Big(\big(\overline{\int f\,d\mu}\big)\,f\Big)\,d\mu\\
&\le \int \Big|\int f\,d\mu\Big|\,|f|\,d\mu\\
&= \Big|\int f\,d\mu\Big| \int |f|\,d\mu,
\end{aligned}$$

from which (A.22) follows also in the complex case.

We end this section with a few special examples of computation of Lebesgue integrals.

A.51 Examples

(a) Let $\omega_0 \in \Omega$, consider the Dirac measure ε_{ω_0} in ω_0, and any (real or complex valued) measurable function on $(\Omega, \mathcal{A})$. Then we get

$$\int f\,d\varepsilon_{\omega_0} = f(\omega_0). \tag{A.26}$$

Indeed, if $f = 1_A$, $A \in \mathcal{A}$, then by definition of the Dirac measure and of the Lebesgue integral of the elementary function f, we find

$$\int f\,d\varepsilon_{\omega_0} = \varepsilon_{\omega_0}(A) = 1_A(\omega_0) = f(\omega_0),$$

and formula (A.26) is proved for such functions. Now we just have to go through the steps of the construction of the Lebesgue integral with this, in order to prove (A.26) in the general case.

(b) First we make the following remark: Suppose that $(\mu_n,\, n \in \mathbb{N})$ is a sequence of measures on a measurable space $(\Omega, \mathcal{A})$, and that $(\alpha_n,\, n \in \mathbb{N})$ is a sequence of positive numbers. Then

$$\mu(A) = \sum_{n=1}^{\infty} \alpha_n \mu_n(A), \qquad A \in \mathcal{A} \tag{A.27}$$

defines a measure on $(\Omega, \mathcal{A})$ which will be denoted by

$$\mu = \sum_{n=1}^{\infty} \alpha_n \mu_n. \tag{A.28}$$

Now consider $\Omega = \mathbb{R}$, $\mathcal{A} = \mathcal{B}(\mathbb{R})$, and Dirac measures ε_n in the natural numbers $n \in \mathbb{N}$. Let $(a_n,\, n \in \mathbb{N})$ be a sequence of numbers which are positive or such that the associated series converges absolutely. Furthermore, let f be any measurable function such that $f(n) = a_n$, $n \in \mathbb{N}$. (For example, we may choose $f = 0$ on $\complement\mathbb{N}$, and $f(n) = a_n$, $n \in \mathbb{N}$, or f can be constructed such that its graph is the polygon through the numbers (n, a_n), $n \in \mathbb{N}_0$, $a_0 = 0$, with $f(x) = 0$ for $x < 0$, etc.) Let $\mu = \sum_n \varepsilon_n$. Then we get with (a) (*exercise!*)

$$\int f\, d\mu = \sum_{n=1}^{\infty} a_n.$$

Thus we see that series are nothing but special Lebesgue integrals. In particular, all the results on Lebesgue integrals which will follow apply *mutatis mutandis* to series.

(c) Choose $\Omega = [0,1]$, $\mathcal{A} = \mathcal{B}([0,1])$, $\mu = \lambda$, and consider the *Dirichlet function* f, which is defined to be 0 on the set of rational numbers in $[0,1]$, and 1 on its complement, the irrational numbers in $[0,1]$. Since $[0,1]_{\mathbb{Q}} = \mathbb{Q} \cap [0,1]$ is a countable union of one-point sets, and since these belong to $\mathcal{B}([0,1])$ as we have already remarked (see example A.7), we know that $[0,1]_{\mathbb{Q}} \in \mathcal{B}([0,1])$, and so does its complement. Also, we know from exercise A.27 that $\lambda([0,1]_{\mathbb{Q}}) = 0$. Thus we get

$$\int_{[0,1]} f\, d\lambda = 1.$$

The Dirichlet function is the typical example of a function whose Lebesgue integral is well-defined, but which is not Riemann–integrable.

Next we convince ourselves that "null sets do not matter for integration". A set $N \subset \Omega$ is called a *(μ–) null set*, if $N \in \mathcal{A}$ and $\mu(N) = 0$. We say that a statement p on Ω is *true (μ–) a.e.* or *holds (μ–) a.e.*, if there exists a μ–null set N so that $p(\omega)$ is true for all $\omega \in \complement N$.

A.52 Theorem *Suppose that f is a real valued or numerical, measurable function on Ω.*

(a) *If $f \geq 0$, then*

$$\int f \, d\mu = 0 \quad \Leftrightarrow \quad f = 0, \ \mu\text{–a.e.}$$

(b) *If N is a μ–null set, then $f \, 1_N$ is integrable, and*

$$\int_N f \, d\mu = 0.$$

Proof For the proof of the implication "$\Rightarrow$" in (a) we first observe that for every $\varepsilon > 0$, the estimate

$$0 = \int f \, d\mu \geq \int_{\{f \geq \varepsilon\}} f \, d\mu \geq \varepsilon \, \mu(\{f \geq \varepsilon\})$$

shows that $\mu(\{f \geq \varepsilon\}) = 0$. Then

$$\mu(\{f > 0\}) = \mu\Big(\bigcup_n \Big\{f \geq \frac{1}{n}\Big\}\Big) = \lim_n \mu\Big(f \geq \frac{1}{n}\Big) = 0.$$

For the implication "$\Leftarrow$" let $N = \{f > 0\}$, so that N is a μ–null set. Denote by i_N the (numerical, measurable) function which is equal to $+\infty$ on N, and zero on its complement. Observe that the sequence $(n \, 1_N, \, n \in \mathbb{N})$ is a sequence of elementary functions which increases pointwise to i_N. Therefore the definition of the Lebesgue integral for elementary functions yields

$$\int i_N \, d\mu = \lim_n \int n \, 1_N \, d\mu = \lim_n n \cdot \mu(N) = 0.$$

Since $f \leq i_N$ on Ω, the fact that the Lebesgue integral is monotone increasing yields that the integral of f is zero, too.

The statement in (b) is proved by the observation that for any measurable function f the functions $f^{\pm} \, 1_N$ are equal to zero μ–a.e. Therefore (a) implies that their μ–integrals are zero, and so $f \, 1_N$ is integrable with integral zero. $\square$

We end this section with the following important theorem which describes the behavior of Lebesgue integrals under measurable transformations.

A.53 Theorem (Transformation Theorem) *Assume that $(\Omega_1, \mathcal{A}_1, \mu)$ is a measurable space, let $(\Omega_2, \mathcal{A}_2)$ be a a measure space, and assume that T is a measurable mapping from Ω_1 into Ω_2, while f is a measurable function on Ω_2. f belongs to $\mathcal{L}^1(\Omega_2, \mathcal{A}_2, T\mu)$ (for the definition of $T\mu$ recall theorem A.31), if and only if $f \circ T$ belongs to $\mathcal{L}^1(\Omega_1, \mathcal{A}_1, \mu)$. In this case or if f is positive, the following relation holds true*

$$\int_{\Omega_1} f \circ T \, d\mu = \int_{\Omega_2} f \, dT\mu. \tag{A.29}$$

The proof for the case $f = 1_A$, $A \in \mathcal{A}_2$, is completely straightforward. Then one just has to use linearity of the Lebesgue integral and to go through the steps of its construction.

A.8 Limit Theorems

The powerful limit theorems for the Lebesgue integral which we shall present here are one of the reasons, why Lebesgue integrals are so popular and useful in mathematics. The proofs cannot be given within the scope of this appendix. The interested reader is referred to the pertinent literature, where one can also find useful variations of the results collected here.

A.54 Theorem (Monotone Convergence Theorem) *Suppose that* $(f_n,\ n \in \mathbb{N})$ *is an increasing sequence of functions in* $\mathcal{M}_+(\Omega, \mathcal{A})$*, and let* $f = \lim_n f_n = \sup_n f_n$*. Then*

$$\int f\,d\mu = \lim_n \int f_n\,d\mu = \sup_n \int f_n\,d\mu \tag{A.30}$$

holds true.

A.55 Theorem (Fatou's Lemma) *Suppose that* $(f_n,\ n \in \mathbb{N})$ *is a sequence of functions in* $\mathcal{M}_+(\Omega, \mathcal{A})$*. Then*

$$\int \liminf_n f_n\,d\mu \leq \liminf_n \int f_n\,d\mu. \tag{A.31}$$

A.56 Theorem (Dominated Convergence Theorem) *Assume that* $(f_n, \in \mathbb{N})$ *is a sequence of measurable functions which converges pointwise to* f*, and that there exists an integrable function* g*, so that* $|f_n| \leq g$ *for all* $n \in \mathbb{N}$*. Then* f *is integrable, and*

$$\int f\,d\mu = \int \lim_n f_n\,d\mu = \lim_n \int f_n\,d\mu \tag{A.32}$$

holds true.

The following theorem of central importance for the calculation of Lebesgue integrals on the real line can be shown as a consequence of the dominated convergence theorem, theorem A.56:

A.57 Theorem *Consider a finite, closed interval* $[a, b]$ *on the real line, and suppose that* f *is a real valued,* $\mathcal{B}([a, b])$*–*$\mathcal{B}(\mathbb{R})$*–measurable function on* $[a, b]$*. If* f *is Riemann–integrable, then* f *is also Lebesgue–integrable with respect to* λ*, and both integrals coincide.*

A.58 Remark In general, the statement of the theorem is *not* true for intervals which are not compact.

So typically, in order to compute a Lebesgue integral of a function on the real line, one approximates the given function by Riemann–integrable (e.g., piecewise continuous) functions, uses the fundamental theorem of calculus to compute the Riemann integrals of the approximations, and then uses a convergence argument, such as theorem A.54 or A.56, to finish the calculation.

A.9 $\mathcal{L}^p$–Spaces

A.59 Definition For $p \geq 1$ the set of measurable functions f which are such that $|f|^p$ is integrable with respect to μ is denoted by $\mathcal{L}^p(\Omega, \mathcal{A}, \mu)$.

A.60 Remark As for $\mathcal{L}^1(\Omega, \mathcal{A}, \mu)$, we shall occasionally leave out some of the "arguments", i.e., we also write $\mathcal{L}^p(\Omega)$, $\mathcal{L}^p(\mu)$ or the like, unless there is danger of confusion.

A.61 Definition For $p \geq 1$, define a mapping $\|\cdot\|_p$ from $\mathcal{L}^p(\Omega, \mathcal{A}, \mu)$ into $\mathbb{R}_+$ by

$$\|f\|_p = \Big(\int |f|^p \, d\mu\Big)^{1/p}, \qquad f \in \mathcal{L}^p(\Omega, \mathcal{A}, \mu). \tag{A.33}$$

If $p = 1$, $f, g \in \mathcal{L}^1(\mu)$ then

$$\|f + g\|_1 = \int |f + g| \, d\mu \leq \int (|f| + |g|) \, d\mu = \|f\|_1 + \|g\|_1,$$

which shows that $\|\cdot\|_1$ defines a semi-norm on $\mathcal{L}^1(\mu)$. Next we show that this is true for general $p \geq 1$, and we shall consider $p > 1$. To this end, let $\alpha = 1/p$, $\beta = 1 - \alpha$. Elementary calculus shows that the function $t \mapsto t^\alpha$, $t \in \mathbb{R}_+$, is concave, therefore its tangent at $t = 1$, lies above the graph of this function. Since the tangent at $t = 1$ is given by the line $t \mapsto 1 - \alpha + \alpha t = \beta + \alpha t$, we get for all $t \in (0, +\infty)$,

$$t^\alpha \leq \alpha t + \beta.$$

Replacing t by t/s, $t, s \in (0, +\infty)$ we get the more symmetric inequality

$$t^\alpha s^\beta \leq \alpha t + \beta s, \qquad s, t > 0. \tag{A.34}$$

Consider $f \in \mathcal{L}^p(\mu)$, $g \in \mathcal{L}^q(\mu)$ which are positive, and such that the integrals of f^p and g^q are equal to 1, where $q > 1$ is such that $1/p + 1/q = 1$, i.e., q is *conjugate to* p. Since above we set $\alpha = 1/p$, we have $1/q = \beta$. From inequality (A.34) we get for every $\omega \in \Omega$,

$$f(\omega)g(\omega) = \big(f(\omega)^p\big)^\alpha \big(g(\omega)^q\big)^\beta \leq \alpha f(\omega)^p + \beta g(\omega)^q.$$

Integrating this inequality (and using the fact that the Lebesgue integral is an increasing mapping) we find

$$\int f \, g \, d\mu \leq 1. \tag{A.35}$$

If $f \in \mathcal{L}^p(\mu)$, $g \in \mathcal{L}^q(\mu)$ are positive, and such that the integrals of f^p, g^q resp., are non-zero, then we can set

$$f' = \frac{f}{\|f\|_p}, \quad g' = \frac{g}{\|g\|_q},$$

and we can apply inequality (A.35) to f', g' instead. The result is the inequality

$$\int fg\,d\mu \leq \|f\|_p\,\|g\|_q. \tag{A.36}$$

Next we drop the assumption that f, g be positive, then with the triangle inequality (A.22) we can estimate as follows:

$$\Big|\int fg\,d\mu\Big| \leq \int |f|\,|g|\,d\mu \leq \|f\|_p\,\|g\|_q. \tag{A.37}$$

Finally we drop the condition that the integrals of $|f|^p$, $|g|^q$ be non-zero: If one of them is (or both are) zero, it follows from theorem A.52 that the corresponding function is zero μ–a.e., so that — with another application of this theorem — inequality (A.37) is trivially true. We have proved the following

A.62 Theorem (Hölder's Inequality) *Suppose that p, $q > 1$ are conjugate to each other, and that $f \in \mathcal{L}^p(\Omega, \mathcal{A}, \mu)$, $g \in \mathcal{L}^q(\Omega, \mathcal{A}, \mu)$. Then* Hölder's inequality (A.37) *holds true.*

A.63 Remark For $p = q = 2$, inequality A.37 is also called *Schwarz' inequality*.

A.64 Theorem (Minkowski's Inequality) *Suppose that $p \geq 1$, and that f, $g \in \mathcal{L}^p(\Omega, \mathcal{A}, \mu)$. Then* Minkowski's inequality

$$\|f + g\|_p \leq \|f\|_p + \|g\|_p \tag{A.38}$$

holds. In particular, $\|\cdot\|_p$ is a semi-norm on $\mathcal{L}^p(\Omega, \mathcal{A}, \mu)$.

Proof We may assume that $p > 1$, and that f and g are positive. Then

$$\begin{aligned}
\|f + g\|_p^p &= \int (f + g)^p\,d\mu \\
&= \int (f + p)^{p-1}(f + g)\,d\mu \\
&= \int (f + g)^{p-1} f\,d\mu + \int (f + g)^{p-1} g\,d\mu.
\end{aligned}$$

Now we apply to both terms on the right hand side Hölder's inequality (A.37) (note that for given $p > 1$, its conjugate is $q = p/p - 1$):

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

and similarly for the other term:

$$\int (f + g)^{p-1} g\,d\mu \leq \|f + g\|_p^{p-1}\|g\|_p.$$

This proves inequality (A.38). $\square$

A.65 Corollary *For every $p \geq 1$, $\mathcal{L}^p(\Omega, \mathcal{A}, \mu)$ is a vector space.*

A.66 Remark Actually, above we have not stated explicitly whether the functions in theorems A.62, A.64, are real or complex valued. However, the attentive reader will have remarked that both are also true in the complex case, so that the previous corollary can be interpreted in both cases accordingly.

Next we consider convergence in $\mathcal{L}^p(\Omega, \mathcal{A}, \mu)$.

A.67 Definition Suppose that $(f_n, \in \mathbb{N})$ is a sequence in $\mathcal{L}^p(\Omega, \mathcal{A}, \mu)$. It is called a *Cauchy sequence*, if and only if $\|f_n - f_m\|_p$ converges to zero as $m, n \to +\infty$. It is said to *converge to* $f \in \mathcal{L}^p(\Omega, \mathcal{A}, \mu)$, if and only if $\|f_n - f\|_p$ converges to zero as $n \to +\infty$.

Assume that the sequence $(f_n, n \in \mathbb{N})$ in $\mathcal{L}^p(\mu)$ converges to $f \in \mathcal{L}^p(\mu)$, and suppose furthermore that $f = g$ μ–a.e., but $f \neq g$. Then $(f_n, n \in \mathbb{N})$ converges to g, too. Thus we see that in general limits of $\mathcal{L}^p(\mu)$–convergence are *not unique* — of course, this nothing but a reflection of the fact that $\|\cdot\|_p$ is only a *semi*–norm, and (in general) *not* a norm. On the other hand, this is not as grave as it seems, because the non-uniqueness can be completely described: Let $\mathcal{N}_\mu$ denote the vector space of measurable functions on $(\Omega, \mathcal{A})$ which are μ–a.e. equal to zero. Then it is clear that $\mathcal{N}_\mu$ is a vector subspace of $\mathcal{L}^p(\mu)$ for every $p \geq 1$ (cf. theorem A.52 (a)). Thus the following definition makes sense:

A.68 Definition For $p \geq 1$, $L^p(\Omega, \mathcal{A}, \mu)$ denotes the quotient of $\mathcal{L}^p(\Omega, \mathcal{A}, \mu)$ by $\mathcal{N}_\mu$.

Let $p \geq 1$, and denote by $[f]_\mu$ the $\mathcal{N}_\mu$–equivalence class of $f \in \mathcal{L}^p(\mu)$, i.e., $[f]_\mu \in L^p(\mu)$, and $f' \in [f]_\mu$ if and only if there exists $g \in \mathcal{N}_\mu$, so that $f - f' = g$. Thus $\|f - f'\|_p = 0$, and so $\|f\|_p = \|f'\|_p$. Hence the value of $\|f\|_p$ does not depend on the choice of the representant chosen in $[f]_\mu$. Therefore we can set $\|[f]_\mu\|_p = \|f\|_p$, and on $L^p(\mu)$, $\|\cdot\|_p$ is a *norm*.

Suppose that $(f_n, n \in \mathbb{N})$ is a sequence in $\mathcal{L}^p(\mu)$ which converges to $f \in \mathcal{L}^p(\mu)$. Then it follows from the triangle inequality for $\|\cdot\|_p$ that $(f_n, n \in \mathbb{N})$ is an $\mathcal{L}^p(\mu)$–Cauchy sequence. For the converse we have the following result:

A.69 Theorem (Riesz–Fischer Theorem) *For every $p \geq 1$, $\mathcal{L}^p(\Omega, \mathcal{A}, \mu)$ is complete.*

For the proof we refer the interested reader to the literature mentioned at the beginning of this appendix. In particular, also $L^p(\Omega, \mathcal{A}, \mu)$, $p \geq 1$, is complete, and therefore it is a *Banach space*.

A.10 The Theorem of Fubini–Tonelli

In this section we assume that $(\Omega_i, \mathcal{A}_i, \mu_i)$, $i = 1, 2$, are two σ–finite measure spaces. Set $\Omega = \Omega_1 \times \Omega_2$, and define a σ–algebra $\mathcal{A}$ as the σ–algebra generated by the Cartesian products $A_1 \times A_2$, $A_i \in \mathcal{A}_i$, $i = 1, 2$. $\mathcal{A}$ is called the *product σ–algebra* of $\mathcal{A}_1$ and $\mathcal{A}_2$, sometimes also denoted by $\mathcal{A}_1 \otimes \mathcal{A}_2$. Furthermore, for $A_i \in \mathcal{A}_i$, $i = 1, 2$, define

$$\mu(A_1 \times A_2) = \mu_1(A_1)\, \mu_2(A_2). \tag{A.39}$$

Then μ is extended in the usual way to a σ–finite, countably additive set function on the ring generated by $\mathcal{A}_1 \times \mathcal{A}_2$. Therefore, by theorems A.24 and A.25, μ has a unique extension to a measure on $(\Omega_1 \times \Omega_2, \mathcal{A}_1 \otimes \mathcal{A}_2)$, which is called the *product measure* of μ_1 and μ_2, also denoted by $\mu_1 \otimes \mu_2$. The measure space $(\Omega_1 \times \Omega_2, \mathcal{A}_1 \otimes \mathcal{A}_2, \mu_1 \otimes \mu_2)$ is called the *product (measure) space* of $(\Omega_1, \mathcal{A}_1, \mu_1)$ and $(\Omega_2, \mathcal{A}_2, \mu_2)$. The product measure space of σ–finite measure spaces $(\Omega_i, \mathcal{A}_i, \mu_i)$, $i = 1, \ldots, n$, $n \in \mathbb{N}$, can be constructed in the same way.

A.70 Exercise

(a) Denote by π_i, $i = 1, 2$, the canonical projection of $\Omega = \Omega_1 \times \Omega_2$ onto Ω_i: $\pi_i(\omega) = \omega_i$, $\omega = (\omega_1, \omega_2) \in \Omega$. Show that $\mathcal{A}_1 \otimes \mathcal{A}_2$ is the smallest σ–algebra on Ω, so that the projections π_1, π_2 become measurable.

(b) Suppose that $\mathcal{C}_i$, $i = 1, 2$, is a generator of $\mathcal{A}_i$. Show that $\mathcal{A}_1 \otimes \mathcal{A}_2 = \sigma(\mathcal{C}_1 \times \mathcal{C}_2)$. (*Hint:* This can be done in a similar way as in exercise A.10.)

(c) Show that for $n \in \mathbb{N}$, $\mathcal{B}(\mathbb{R}^n) = \mathcal{B}(\mathbb{R}) \otimes \cdots \otimes \mathcal{B}(\mathbb{R})$ (n factors). Conclude that the Lebesgue measure λ^n on $\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ is the n–fold product measure of the one-dimensional Lebesgue measure: $\lambda^n = \lambda \otimes \cdots \otimes \lambda$.

A.71 Theorem (Fubini–Tonelli Theorem) *Consider the product space*

$$(\Omega_1 \times \Omega_2, \mathcal{A}_1 \otimes \mathcal{A}_2, \mu_1 \otimes \mu_2)$$

of two σ–finite measure spaces $(\Omega_i, \mathcal{A}_i, \mu_i)$, $i = 1, 2$, and let f be an $\mathcal{A}_1 \otimes \mathcal{A}_2$–measurable function.

(a) *For every $\omega_1 \in \Omega_1$, the function $f(\omega_1, \cdot)$ on Ω_2 is $\mathcal{A}_2$–measurable.*

(b) *If f is positive or if $f(\omega_1, \cdot)$ is μ_2–integrable for every $\omega_1 \in \Omega_1$, the mapping*

$$\omega_1 \mapsto \int_{\Omega_2} f(\omega_1, \omega_2)\, d\mu_2(\omega_2)$$

is $\mathcal{A}_1$–measurable.

(c) *The statements analogous to (a) and (b), when the indices* 1 *and* 2 *are interchanged, hold true.*

(d) *f is $\mu_1 \otimes \mu_2$–integrable, if and only if*

$$\int_{\Omega_1} \Big(\int_{\Omega_2} |f(\omega_1, \omega_2)| \, d\mu_2(\omega_2) \Big) \, d\mu_1(\omega_1) < +\infty$$

or

$$\int_{\Omega_2} \Big(\int_{\Omega_1} |f(\omega_1, \omega_2)| \, d\mu_1(\omega_1) \Big) \, d\mu_2(\omega_2) < +\infty.$$

(e) *If f is positive or $\mu_1 \otimes \mu_2$–integrable, then the following equalities hold true:*

$$\begin{aligned} \int_{\Omega_1 \times \Omega_2} & f(\omega_1, \omega_2) \, d(\mu_1 \otimes \mu_2)(\omega_1, \omega_2) \\ &= \int_{\Omega_1} \Big(\int_{\Omega_2} f(\omega_1, \omega_2) \, d\mu_2(\omega_2) \Big) \, d\mu_1(\omega_1) \qquad \text{(A.40)} \\ &= \int_{\Omega_2} \Big(\int_{\Omega_1} f(\omega_1, \omega_2) \, d\mu_1(\omega_1) \Big) \, d\mu_2(\omega_2). \end{aligned}$$

A.11 The Theorem of Radon–Nikodym

We introduce two new notions, and do this in a somewhat informal manner. Given a measurable space $(\Omega, \mathcal{A})$, a *signed measure* ν thereon is a countably additive set function with values in $\overline{\mathbb{R}} = [-\infty, +\infty]$, so that $\nu(\emptyset) = 0$, and such that at most one of the values $\pm\infty$ is assumed. The decomposition theorems of *Hahn* and of *Jordan* state that if ν is a signed measure on $(\Omega, \mathcal{A})$, then Ω has a decomposition $\Omega = \Omega_+ \uplus \Omega_-$ in $\mathcal{A}$, so that ν is positive (negative) on Ω_+ (Ω_-, respectively), and furthermore ν can correspondingly be written uniquely (ν–a.e.) as the difference of two measures on Ω.

Suppose that $(\Omega, \mathcal{A}, \mu)$ is a measure space, and that ν is a signed measure on $(\Omega, \mathcal{A})$. ν is called *absolutely continuous with respect to* μ, if every μ–null set is also a ν–null set.

A.72 Theorem (Radon–Nikodym Theorem) *Assume that $(\Omega, \mathcal{A}, \mu)$ is a measure space, and that ν is a signed measure on $(\Omega, \mathcal{A})$ which is absolutely continuous with respect to μ. Then there exists a (μ–a.e.) unique function $f \in \mathcal{L}^1(\Omega, \mathcal{A}, \mu)$, so that*

$$\nu(A) = \int_A f \, d\mu, \qquad A \in \mathcal{A}, \tag{A.41}$$

holds true.

A.12 Jensen's Inequality

In this section we specialize to the case of a probability space $(\Omega, \mathcal{A}, P)$. As usual in this context, for a positive or integrable random variable X we write $E(X)$ for its P–integral over Ω.

Suppose that φ is a convex function on $\mathbb{R}$, that is, for all $x, y \in \mathbb{R}$, and $\alpha \in [0, 1]$, we have

$$\varphi\big(\alpha x + (1-\alpha)y\big) \leq \alpha\varphi(x) + (1-\alpha)\varphi(y). \tag{A.42}$$

Of course, this means that every line segment between two points of the graph of φ lies above the graph of φ. It is well-known (and proved in many analysis textbooks, a detailed proof can also be found in [5, Chapter 3], where the reader can also find a complete discussion of the material of this section), that every convex function is continuous, admits everywhere right and left derivatives, and that at every point of its graph the right and left tangent lie under the graph of φ. The last statement means that

$$\varphi(x) \geq \varphi(\xi) + \varphi'(\xi+)(x-\xi), \qquad x, \xi \in \mathbb{R}, \tag{A.43}$$

where $\varphi'(\xi+)$ denotes the derivative of φ from the right at ξ. Since at $x = \xi$ the last inequality turns into an equality, we also get

$$\varphi(x) = \sup_{\xi\in\mathbb{R}}\big(\varphi(\xi) + \varphi'(\xi+)(x-\xi)\big), \qquad x \in \mathbb{R}. \tag{A.44}$$

Replacing x by X in (A.43) and integrating of both sides with respect to P, we obtain the following inequality

$$E(\varphi \circ X) \geq \varphi(\xi) + \varphi'(\xi+)\big(E(X) - \xi\big),$$

which holds true for all $\xi \in \mathbb{R}$. But this implies

$$\begin{aligned} E(\varphi \circ X) &\geq \sup_{\xi\in\mathbb{R}}\Big(\varphi(\xi) + \varphi'(\xi+)\big(E(X) - \xi\big)\Big) \\ &= \varphi\big(E(X)\big), \end{aligned}$$

where we used (A.44). Thus we have proved the following

A.73 Theorem (Jensen's Inequality) *Suppose that X is a real valued, integrable random variable. Assume furthermore that φ is a convex function, such that $\varphi \circ X$ is integrable, too. Then*

$$E(\varphi \circ X) \geq \varphi\big(E(X)\big). \tag{A.45}$$

Appendix B

Metric Space of Sequences

In this appendix we study the space of real sequences as a metric space. Before we enter this, it will be useful to discuss the well-known diagonal sequence trick which we will use in this context, but which is also of independent interest.

B.1 The Diagonal Sequence Trick

The diagonal sequence trick is often used to prove the sequential compactness of a given set in an infinite dimensional setting. For example, it is essential to the standard proof of the theorem of Arzèla–Ascoli, and in the present lectures it serves to prove Helly's theorem, theorem 6.4.

Suppose that $(a^n, n \in \mathbb{N})$ is a sequence of sequences $(a^n_m, m \in \mathbb{N})$, $n \in \mathbb{N}$, in some metric space, which is such that for every $m \in \mathbb{N}$ the following holds true:

B.1 Hypothesis *For every $m \in \mathbb{N}$, every subsequence of $(a^n_m, n \in \mathbb{N})$ has a convergent subsequence.*

For example, if the sequences in question take their values in $\mathbb{R}$, and if for every $m \in \mathbb{N}$, the sequence $(a^n_m, n \in \mathbb{N})$ is bounded, then it satisfies this assumption by the theorem of Bolzano–Weierstraß . More generally this is so, if for every $m \in \mathbb{N}$ the sequence $(a^n_m, n \in \mathbb{N})$ belongs to a relatively compact subset of the underlying metric space.

B.2 Lemma *There exists a subsequence $(a^{n(l)}, l \in \mathbb{N})$ of $(a^n, n \in \mathbb{N})$ which is such that for every $m \in \mathbb{N}$, the sequence $(a^{n(l)}_m, l \in \mathbb{N})$ converges.*

Proof Our hypothesis B.1 successively implies the following statements:

$m = 1$:
There is a strictly increasing subsequence $(n_{1l}, l \in \mathbb{N})$ of $\mathbb{N}$ (considered as a sequence), and an $a_1 \in \mathbb{R}$ so that

$$a_1^{n_{1l}} \to a_1, \quad l \to \infty.$$

$m = 2$:
There is a strictly increasing subsequence $(n_{2l}, l \in \mathbb{N})$ of $(n_{1l}, l \in \mathbb{N})$, and an $a_2 \in \mathbb{R}$ so that

$$a_2^{n_{2l}} \to a_2, \quad l \to \infty,$$

and iteratively
$m = k$:
There is a strictly increasing subsequence $(n_{kl}, l \in \mathbb{N})$ of $(n_{k-1\,l}, l \in \mathbb{N})$, and an $a_k \in \mathbb{R}$ so that

$$a_k^{n_{kl}} \to a_k, \quad l \to \infty.$$

Thus we get the scheme

$$\begin{array}{cccccccc}
n_{11} & n_{12} & n_{13} & \cdots & n_{1k} & \cdots & n_{1l} & \cdots \\
n_{21} & n_{22} & n_{23} & \cdots & n_{2k} & \cdots & n_{2l} & \cdots \\
n_{31} & n_{32} & n_{33} & \cdots & n_{3k} & \cdots & n_{3l} & \cdots \\
\vdots & \vdots & \vdots & \ddots & \vdots & & \vdots & \\
n_{k1} & n_{k2} & n_{k3} & \cdots & n_{kk} & \cdots & n_{kl} & \cdots \\
\vdots & \vdots & \vdots & & \vdots & \ddots & \vdots & \\
n_{l1} & n_{l2} & n_{l3} & \cdots & n_{lk} & \cdots & n_{ll} & \cdots \\
\vdots & \vdots & \vdots & & \vdots & & \vdots & \ddots
\end{array}$$

Now define the *diagonal sequence* $(n(l), l \in \mathbb{N})$ by $n(l) = n_{ll}$, $l \in \mathbb{N}$. Then for every $k \in \mathbb{N}$, the sequence $(n(l), l \geq k)$ is a subsequence of $(n_{kl}, l \in \mathbb{N})$. Since by construction

$$a_k^{n_{kl}} \to a_k, \quad l \to \infty,$$

we also have that

$$a_k^{n(l)} \to a_k, \quad l \to \infty.$$

Therefore the subsequence $(a^{n(l)}, l \in \mathbb{N})$ is as desired. □

B.2 The Metric Space of Real Sequences

Consider the space $\mathbb{R}^{\mathbb{N}}$ of all real valued sequences $a = (a_k, k \in \mathbb{N})$. For $a, b \in \mathbb{R}^{\mathbb{N}}$ define

$$\rho(a, b) = \sum_{k=1}^{\infty} 2^{-k} \frac{|a_k - b_k|}{1 + |a_k - b_k|}. \tag{B.1}$$

B.3 Exercise Show that ρ defines a bounded metric on $\mathbb{R}^{\mathbb{N}}$. (*Hint:* Show first that $x \mapsto |x|/(1 + |x|)$ is subadditive on $\mathbb{R}$, where a function φ on $\mathbb{R}$ is called *subadditive*, if $\varphi(x + y) \leq \varphi(x) + \varphi(y)$ for all $x, y \in \mathbb{R}$.)

We recall that a subset $A \subset M$ of a metric space (M, d) is said to be *relatively compact*, if its closure $\overline{A}$ is a compact subset of M. Moreover, in a metric space the notions of compactness and sequential compactness coincide. Thus A is relatively compact, if and only if every sequence in A has a subsequence which converges in M with respect to d (not necessarily in A — if this is so, then A is actually compact).

B.4 Lemma *Consider the metric space $(\mathbb{R}^{\mathbb{N}}, \rho)$. The following statements hold true:*

(a) *A sequence $(a^n,\ n \in \mathbb{N})$ converges with respect to ρ to $a \in \mathbb{R}^{\mathbb{N}}$, if and only if for every $k \in \mathbb{N}$, $a_k^n \to a_k$, as $n \to \infty$.*

(b) *$(\mathbb{R}^{\mathbb{N}}, \rho)$ is complete.*

(c) *The space $\mathbb{Q}_0^{\mathbb{N}}$ of all rational sequences with at most finitely many nonzero members is dense in $(\mathbb{R}^{\mathbb{N}}, \rho)$.*

(d) *A subset $A \subset \mathbb{R}^{\mathbb{N}}$ is relatively compact in $\mathbb{R}^{\mathbb{N}}$, if and only if for every $k \in \mathbb{N}$ the set $A_k = \{a_k,\ a \in A\}$ bounded in $\mathbb{R}$.*

B.5 Remark The Heine–Borel–Lebesgue–theorem states that a subset of $\mathbb{R}^d$, $d \in \mathbb{N}$, is relatively compact, if and only if it is bounded. Therefore statement (d) could have been formulated equivalently as: *A subset $A \subset \mathbb{R}^{\mathbb{N}}$ is relatively compact in $\mathbb{R}^{\mathbb{N}}$, if and only if for every $k \in \mathbb{N}$ the set A_k is relatively compact in $\mathbb{R}$.*

Proof (of lemma B.4)
Statement (a):
"$\Leftarrow$" This can be proved most easily with an application of the dominated convergence theorem (theorem A.56): The terms under the sum

$$\sum_{k=1}^{\infty} 2^{-k} \frac{|a_k^n - a_k|}{1 + |a_k^n - a_k|}$$

converge to zero as $n \to \infty$, and they are uniformly in n bounded by 2^{-k}, which is summable. Therefore the series converges to zero with $n \to \infty$.
"$\Rightarrow$" Given $k \in \mathbb{N}$ and $\varepsilon > 0$. Choose $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$,

$$\rho(a^n, a) \leq \frac{\varepsilon}{1+\varepsilon} 2^{-k}.$$

Then we get for $n \geq n_0$ the estimate

$$\frac{|a_k^n - a_k|}{1 + |a_k^n - a_k|} \leq \frac{\varepsilon}{1+\varepsilon},$$

from which we directly conclude that $|a_k^n - a_k| \leq \varepsilon$.

Statement (b):
Suppose that $(a^n,\ n \in \mathbb{N})$ is a Cauchy sequence with respect to ρ. With an argument

similar to the one we used for the proof of (a), this is the case, if and only if for every $k \in \mathbb{N}$ the sequence $(a_k^n,\, n \in \mathbb{N})$ is a Cauchy sequence in $\mathbb{R}$. Thus for every $k \in \mathbb{N}$ there exists an $a_k \in \mathbb{R}$ so that $a_k^n \to a_k$ as $n \to \infty$, because $\mathbb{R}$ is complete. But then we get from (a) that $\rho(a^n, a) \to 0$, $n \to \infty$, where $a = (a_k,\, k \in \mathbb{N}) \in \mathbb{R}^{\mathbb{N}}$.

Statement (c):
Let $a \in \mathbb{R}^{\mathbb{N}}$ and $\varepsilon > 0$ be given. Choose $k_0 \in \mathbb{N}$ such that

$$\sum_{k=k_0+1}^{\infty} 2^{-k} \frac{|a_k|}{1+|a_k|} \le \frac{\varepsilon}{2}.$$

Now choose $q_1, \dots, q_{k_0} \in \mathbb{Q}$ so that

$$\max_{k=1,\dots,k_0} |a_k - q_k| \le \frac{\varepsilon}{2}.$$

Put $q_k = 0$, $k \ge k_0 + 1$, and $q = (q_k,\, k \in \mathbb{N}) \in \mathbb{Q}_0^{\mathbb{N}}$. Then $\rho(a, q) \le \varepsilon$.

Statement (d):
"$\Rightarrow$" Suppose that there exists $k \in \mathbb{N}$ so that A_k is an unbounded subset of $\mathbb{R}$. Then for every $n \in \mathbb{N}$ there exists an element $a_k^n \in A_k$ with $|a_k^n| \ge n$. Consider the sequence $(a^n,\, n \in \mathbb{N})$ in A defined this way. By hypothesis A is relatively compact, so that $(a^n,\, n \in \mathbb{N})$ has a subsequence $(a^{n_l},\, l \in \mathbb{N})$ which converges in $(\mathbb{R}^{\mathbb{N}}, \rho)$. From (a) we find that the sequence $(a_k^{n_l},\, l \in \mathbb{N})$ converges in $\mathbb{R}$, which is in contradiction to the construction of $(a_k^n,\, n \in \mathbb{N})$.
"$\Leftarrow$" Let $(a^n,\, n \in \mathbb{N})$ be a sequence in A. Under the hypothesis that for every $k \in \mathbb{N}$ the set A_k is bounded, we have to show that there exists a subsequence $(a^{n_l},\, l \in \mathbb{N})$ of $(a^n,\, n \in \mathbb{N})$ which converges. Since for every $k \in \mathbb{N}$ the set A_k is a bounded set, the sequence $(a_k^n,\, n \in \mathbb{N})$ — and consequently also every subsequence of it — is bounded. The theorem of Bolzano–Weierstraß implies that every subsequence of $(a_k^n,\, n \in \mathbb{N})$ has a convergent subsequence. But this means that the hypothesis B.1 is valid in the present context. Thus the diagonal sequence trick, lemma B.2, provides the existence of a subsequence $(a^{n_l},\, l \in \mathbb{N})$ of $(a^n,\, n \in \mathbb{N})$ and an $a \in \mathbb{R}^{\mathbb{N}}$, which are such that for every $k \in \mathbb{N}$, $a_k^{n_l} \to a_k$, as $l \to \infty$. By statement (a) we get that $\rho(a^{n_l}, a) \to 0$, $l \to \infty$. Thus A is relatively compact. $\square$

Appendix C

A Theorem of Weierstraß

In this appendix we prove a famous theorem of Weierstraß about the approximation of continuous functions on a compact interval, which is different from the one finds in many analysis courses, viz. theorem 4.5 which treats the approximation by polynomials. Here we consider another class of approximating functions:

C.1 Definition A complex valued function φ on $[a,b]$, $a, b \in \mathbb{R}$, $a < b$, of the form

$$\varphi(x) = \sum_{k=-n}^{n} a_{n,k} \exp\Bigl(i\,\frac{2\pi}{b-a}\,kx\Bigr), \qquad x \in [a,b], \tag{C.1}$$

with $a_{n,k} \in \mathbb{C}$, $n \in \mathbb{N}$, $k = -n, -(n-1), \dots, n$, is called a *trigonometric polynomial* on $[a,b]$.

C.2 Theorem (Weierstraß) *For every complex valued continuous function f on $[a,b]$, a, $b \in \mathbb{R}$, $a < b$, with $f(a) = f(b)$, there exists a sequence $(\varphi_n,\ n \in \mathbb{N})$ of trigonometric polynomials on $[a,b]$, which converges uniformly to f.*

Let us simplify our discussion by first arguing that it is sufficient to consider the interval $[0,2\pi]$: Indeed for given f on $[a,b]$ as above, set

$$\tilde f(x) = f\Bigl(\frac{b-a}{2\pi}x + a\Bigr), \qquad x \in [0,2\pi],$$

then $\tilde f$ is continuous on $[0,2\pi]$ with $\tilde f(0) = \tilde f(2\pi)$. Suppose that the theorem has been proved for the interval $[0,2\pi]$, and that $(\tilde\varphi_n,\ n \in \mathbb{N})$ is a sequence of trigonometric polynomials on $[0,2\pi]$ which converges uniformly to $\tilde f$ on $[0,2\pi]$. Set

$$\varphi_n(x) = \tilde\varphi_n\Bigl(2\pi\,\frac{x-a}{b-a}\Bigr), \qquad x \in [a,b].$$

Then this defines a sequence of trigonometric polynomials which converges uniformly on $[a,b]$ to f.

Weierstraß' theorem C.2 actually follows directly from the powerful Stone–Weierstraß–theorem, which is a generalization of the classical theorem of Weierstraß cited at the beginning of the appendix, and it can be found in many texts on functional analysis or harmonic analysis. A (quite old but) nice reference is the book by Loomis [26]. In order to apply this theorem, one identifies f with a continuous function defined on the unit circle $\mathcal{S}^1$ in $\mathbb{R}^2$ (or the complex plane — observe that for this the condition $f(a) = f(b)$ is essential), and considers the algebra of functions given by the trigonometric polynomials, considered as well as continuous functions on the circle. Since $\mathcal{S}^1$ is compact, one can apply the Stone–Weierstraß–theorem which gives theorem C.2.

In order to make these lecture notes self–contained, and since theorem C.2 plays a key role in chapter 6, we give another, more classical and calculative proof below. To this end, we bring in the following sequence $(K_n,\ n \in \mathbb{N})$ of functions on $\mathbb{R}$, called *Fejér kernels*:

$$K_n(x) = \sum_{k=-n}^{n} \Bigl(1 - \frac{|k|}{n+1}\Bigr) e^{ikx}, \qquad x \in \mathbb{R},\ n \in \mathbb{N}. \tag{C.2}$$

In figure C.1 the Fejér kernels for $n = 3$, 5, 10, and 20 are depicted. (The reader might wonder about these graphs, since by the definition in (C.2) K_n, $n \in \mathbb{N}$, looks like a *complex* valued functions. However we shall show below that K_n is real valued.)

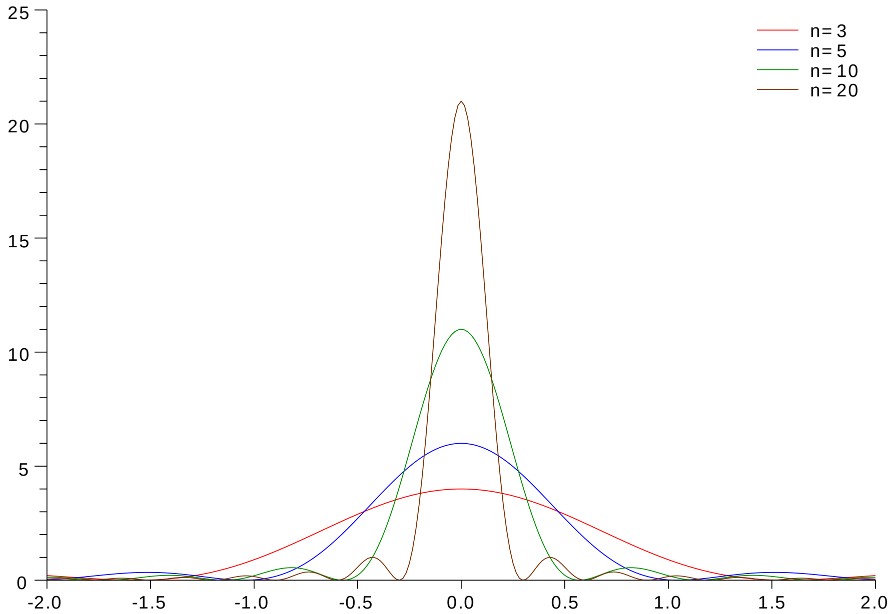

Figure C.1: The Fejér kernels for $n = 3$, 5, 10, and 20

C.3 Lemma *For every $n \in \mathbb{N}$,*

$$K_n(m2\pi) = n + 1, \qquad m \in \mathbb{Z}, \tag{C.3a}$$

and

$$K_n(x) = \frac{1}{n+1}\Big(\frac{\sin\big((n+1)x/2\big)}{\sin(x/2)}\Big)^2, \qquad x \in \mathbb{R},\ x \neq m2\pi,\ m \in \mathbb{Z}. \tag{C.3b}$$

Moreover, for all $n \in \mathbb{N}$,

$$\frac{1}{2\pi}\int_0^{2\pi} K_n(x)\,dx = 1 \tag{C.4}$$

holds true.

Proof The proof of this lemma is by elementary computation. First we observe that for every complex sequence $(a_n,\ n \in \mathbb{Z})$ the following holds for every $n \in \mathbb{N}$:

$$\sum_{k=0}^{n}\sum_{l=0}^{n} a_{k-l} = \sum_{m=-n}^{n} (n+1-|m|)\,a_m. \tag{C.5}$$

Indeed, for $m \in \{0, 1, \dots, n\}$, there are $n+1-m$ many possibilities to write $m = k-l$ for $k, l \in \{0, 1, \dots, n\}$, namely those which are given by $k \in \{m, m+1, \dots, n\}$. A similar argument takes care of the case $m \in \{-n, \dots, -1\}$. Now we can compute as follows

$$\begin{aligned}
\sum_{m=-n}^{n}\Big(1-\frac{|m|}{n+1}\Big)e^{imx} &= \frac{1}{n+1}\sum_{m=-n}^{n}\big(n+1-|m|\big)e^{imx} \\
&= \frac{1}{n+1}\sum_{k=0}^{n}\sum_{l=0}^{n} e^{i(k-l)x} \\
&= \frac{1}{n+1}\Big|\sum_{k=0}^{n} e^{ikx}\Big|^2,
\end{aligned}$$

where we used (C.5). If $x = m2\pi$ for some $m \in \mathbb{Z}$, $\exp(ikx) = 1$, and (C.3a) follows. In the sequel we assume that for all $m \in \mathbb{Z}$, $x \neq m2\pi$, so that $\exp(ix) \neq 1$. Then

$$\sum_{k=0}^{n} e^{ikx} = \frac{e^{i(n+1)x}-1}{e^{ix}-1},$$

and we obtain

$$\begin{aligned}
&\sum_{m=-n}^{n}\Big(1-\frac{|m|}{n+1}\Big)e^{imx} \\
&= \frac{1}{n+1}\Big|\frac{e^{i(n+1)x}-1}{e^{ix}-1}\Big|^2 \\
&= \frac{1}{n+1}\Big|\frac{e^{i(n+1)x/2}\big(e^{i(n+1)x/2}-e^{-i(n+1)x/2}\big)}{e^{ix/2}\big(e^{ix/2}-e^{-ix/2}\big)}\Big|^2 \\
&= \frac{1}{n+1}\Big(\frac{\sin\big((n+1)x/2\big)}{\sin(x/2)}\Big)^2.
\end{aligned}$$

Finally, equality (C.4) is due to

$$\int_0^{2\pi} e^{ikx}\, dx = 0,$$

for all $k \in \mathbb{Z}$, $k \neq 0$, and the proof is complete. □

For a complex valued, continuous function f defined on $[0, 2\pi]$, with $f(0) = f(2\pi)$, and $x \in [0, 2\pi]$ define

$$K_n * f(x) = \frac{1}{2\pi} \int_0^{2\pi} K_n(x-y) f(y)\, dy. \tag{C.6}$$

C.4 Theorem (Fejér) *If f is a complex valued continuous function on $[0, 2\pi]$ with $f(0) = f(2\pi)$, then $(K_n * f,\ n \in \mathbb{N})$ converges uniformly on $[0, 2\pi]$ to f.*

Proof We identify f with its 2π–periodic extension to all of $\mathbb{R}$, and recall that we had already noted that K_n, $n \in \mathbb{N}$, is 2π–periodic. Moreover, the extended function f is uniformly continuous on $\mathbb{R}$: On each interval of the form $[(m-1)2\pi, m2\pi]$, $m \in \mathbb{Z}$, this is true by hypothesis, and at the boundary points of the intervals we get (uniform) continuity due to the assumption that $f(0) = f(2\pi)$, which entails for the extended f that $f(m2\pi) = f(m'2\pi)$ for all $m, m' \in \mathbb{Z}$. Using this extension of f, we can also write

$$K_n * f(x) = \frac{1}{2\pi} \int_0^{2\pi} K_n(y) f(x-y).$$

Given $\varepsilon > 0$, we have to find an $n_0 \in \mathbb{N}$ so that for all $n \geq n_0$ we get

$$\sup_{0 \leq x \leq 2\pi} |K_n * f(x) - f(x)| < \varepsilon. \tag{C.7}$$

Since f is uniformly continuous there exists a $\delta > 0$, $\delta \leq 2\pi$, so that $|x - y| < \delta$, x, $y \in \mathbb{R}$, implies that $|f(x) - f(y)| < \varepsilon/4$. Choose $n_0 \in \mathbb{N}$ in such a way that

$$\frac{2}{n_0 + 1} \|f\|_\infty \sup_{\delta < x < 2\pi - \delta} \frac{1}{\sin(x/2)^2} < \frac{\varepsilon}{2}. \tag{C.8}$$

Consider $x \in [0, 2\pi]$, and let $n \geq n_0$. Then we can estimate in the following way:

$$\begin{aligned} &|K_n * f(x) - f(x)| \\ &\qquad = \frac{1}{2\pi} \Big| \int_0^{2\pi} K_n(y) \big(f(x-y) - f(x) \big)\, dy \Big| \\ &\qquad \leq \frac{1}{2\pi} \int_\delta^{2\pi - \delta} K_n(y) \big| f(x-y) - f(x) \big|\, dy \\ &\qquad\qquad + \frac{1}{2\pi} \int_{-\delta}^{\delta} K_n(y) \big| f(x-y) - f(x) \big|\, dy, \end{aligned}$$

because formula (C.3) shows in particular that K_n is positive. Moreover, for all $y \in (\delta, 2\pi - \delta)$ we get from (C.3b) the estimate

$$K_n(y) \le \frac{1}{n+1}\,\frac{1}{\sin(y/2)^2} \le \frac{1}{n_0+1}\,\frac{1}{\sin(y/2)^2}.$$

We use this for the first of the integrals above, and in combination with inequality (C.8) we get

$$\begin{aligned}
|K_n * f(x) - f(x)| &< \frac{\varepsilon}{2} + \frac{\varepsilon}{4}\,\frac{1}{2\pi}\int_{-\delta}^{\delta} K_n(y)\,dy \\
&= \frac{\varepsilon}{2} + \frac{\varepsilon}{2}\,\frac{1}{2\pi}\int_{0}^{\delta} K_n(y)\,dy \\
&\le \frac{\varepsilon}{2} + \frac{\varepsilon}{2}\,\frac{1}{2\pi}\int_{0}^{2\pi} K_n(y)\,dy \\
&= \varepsilon,
\end{aligned}$$

where we used the symmetry $K_n(y) = K_n(-y)$, $y \in \mathbb{R}$, in the second step, and in the last step we made use of (C.4). □

Now we can come to the

Proof (of theorem C.2) As remarked above, we only need to consider the case where the interval $[a, b]$ is given as $[0, 2\pi]$. If f is continuous on $[0, 2\pi]$ with $f(0) = f(2\pi)$, then $(K_n * f,\ n \in \mathbb{N})$ converges uniformly on $[0, 2\pi]$ to f. But by (C.6)

$$\begin{aligned}
K_n * f(x) &= \frac{1}{2\pi}\int_0^{2\pi} K_n(x-y) f(y)\,dy \\
&= \sum_{k=-n}^{n} \Big(1 - \frac{|k|}{n+1}\Big)\,\frac{1}{2\pi}\Big(\int_0^{2\pi} e^{-iky} f(y)\,dy\Big)\, e^{ikx},
\end{aligned}$$

which shows that for every $n \in \mathbb{N}$, $K_n * f$ is indeed a trigonometric polynomial. □

Appendix D

Bochner's Theorem

In this appendix we prove Bochner's theorem, e.g., [8, Supplement, III.§8], which also has been stated in chapter 6, see theorem 6.18. The proof we give below is a version of the one which can be found in [21]. First we prepare some background material.

D.1 Preparations

D.1 Definition A *complex measure* μ on a measurable space $(\Omega, \mathcal{A})$ is a σ–additive mapping from $\mathcal{A}$ into $\mathbb{C}$. The *total variation measure* $|\mu|$ of μ is defined by

$$|\mu|(A) = \sup \sum_{n=1}^{\infty} |\mu(A_n)|,$$

where the supremum is over all pairwise disjoint sequences $(A_n,\ n \in \mathbb{N})$ in $\mathcal{A}$ so that $A = \uplus_n A_n$.

Suppose that μ is a complex measure on $(\Omega, \mathcal{A})$, and that $(A_n,\ n \in \mathbb{N})$ is a pairwise disjoint sequence in $\mathcal{A}$ with $A = \uplus_n A_n$. Then the condition of σ–additivity reads

$$\mu(A) = \sum_{n=1}^{\infty} \mu(A_n).$$

Observe that this condition has implicitly built in the hypothesis that the series on the right hand side converges, because — in contrast to the case of a positive measure — the sequence $(\mu(A_n),\ n \in \mathbb{N})$ is a general sequence in $\mathbb{C}$, and the associated sequence of partial sums is no longer increasing. Actually, if it converges one can argue that it converges absolutely. The interested reader is referred to [32, Chapter 6] for a proof of this result, as well as of those concerning the theory of complex measures which are stated hereafter.

The fact that for every complex measure μ its total variation measure $|\mu|$ is indeed a (positive) *measure* on $(\Omega, \mathcal{A})$ is quite non-trivial. It turns out that $|\mu|$ is always finite:

$|\mu|(\Omega) < +\infty$. It is clear that if μ is a positive measure, then $\mu = |\mu|$. Basically as a consequence of the Radon–Nikodym–theorem (which also holds true for complex measures), one has that for every complex measure μ on $(\Omega, \mathcal{A})$

$$\mu(A) = \int_A h(x)\, d|\mu|(x)$$

for every $A \in \mathcal{A}$, where h is a measurable function of modulus 1. Upon decomposition of h into its real and imaginary parts, and by decomposition of the latter into their respective positive and negative parts, one arrives at the result that a complex measure can be written as a complex linear combination of finite positive measures on $(\Omega, \mathcal{A})$. In fact, it is convenient to think of a complex measure in this way.

The set of complex measures on $(\Omega, \mathcal{A})$ will be denoted by $\mathcal{M}(\Omega, \mathcal{A})$, and if the σ–algebra is understood from the context we shall also simply write $\mathcal{M}(\Omega)$. Obviously, with the natural way to define complex linear combinations of complex measures, $\mathcal{M}(\Omega, \mathcal{A})$ is a complex vector space. With the definition $\|\mu\|_{\mathcal{M}(\Omega,\mathcal{A})} = |\mu|(\Omega)$, $\mu \in \mathcal{M}(\Omega, \mathcal{A})$, it becomes a normed vector space.

A central result of functional analysis is the following representation theorem by F. Riesz. Suppose that $(X, \mathcal{T})$ is a locally compact Hausdorff topological space, and denote by $\mathcal{B}(X)$ its Borel–σ–algebra $\sigma(\mathcal{T})$. Let $C_0(X)$ denote the Banach space of complex valued continuous functions on X which vanish at infinity, equipped with the supremum norm $\|\cdot\|_\infty$. That f vanishes at infinity means here that for every $\varepsilon > 0$ there exists a compact subset K of X so that for all $x \in \complement K$ we have $|f(x)| \le \varepsilon$. A complex measure μ on $(X, \mathcal{B}(X))$ is called *regular*, if $|\mu|$ is regular in the sense that

$$|\mu|(A) = \sup\{|\mu|(K),\ K \subset A,\ K \text{ compact}\}$$

for all open $A \subset X$, and all $A \in \mathcal{B}(X)$ with $|\mu|(A) < +\infty$. $\mathcal{M}_{\mathrm{reg}}(X)$ denotes the set of all regular complex measures on $(X, \mathcal{B}(X))$. We remark that if X is a polish space, then every finite positive Borel measure is regular (e.g., [4, Satz 26.3]), and therefore every complex measure is regular. In particular, every complex measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ or on an interval in $\mathbb{R}$ — and this is the only case which interest us in this appendix — is regular.

D.2 Theorem (Riesz) *The dual space of* $(C_0(X), \|\cdot\|_\infty)$ *is equal to* $\mathcal{M}_{reg}(X)$. *That is, if* Φ *is a continuous linear functional on* $C_0(X)$*, then there exists a unique* $\mu \in \mathcal{M}_{reg}(X)$ *such that*

$$\Phi(f) = \int_X f\, d\mu, \qquad f \in C_0(X).$$

Moreover, $\|\Phi\| = \|\mu\|_{\mathcal{M}(X)}$ *holds, where*

$$\|\Phi\| = \sup\{\Phi(f),\ f \in C_0(X),\ \|f\|_\infty = 1\}.$$

Proofs of Riesz' representation theorem can be found in many textbooks, for example in [3,4,29,32]. For the case $X = \mathbb{R}$ a rather intuitive proof is given in [31].

If $(\mu_n,\, n \in \mathbb{N})$ is a sequence in $\mathcal{M}_{\mathrm{reg}}(X)$ such that for every $f \in C_0(X)$,

$$\lim_{n\to\infty} \int_X f\, d\mu_n = \int_X f\, d\mu$$

for some $\mu \in \mathcal{M}_{\mathrm{reg}}(X)$, then we say that μ is the *weak–$*$ limit* of $(\mu_n,\, n \in \mathbb{N})$. For later purposes we remark here that it follows from the *principle of uniform boundedness* that in this case the sequence $(\mu_n,\, n \in \mathbb{N})$ is bounded (see, e.g., [20, Theorem III.1.28]): There exists $C \geq 0$ so that for all $n \in \mathbb{N}$, $\|\mu_n\|_{\mathcal{M}(X)} \leq C$.

D.2 Herglotz' Theorem

As a first step towards the proof of Bochner's theorem, we prove its analog for positive definite sequences (see below) arising as the Fourier transform of positive measures on a finite interval: *Herglotz' theorem.* First we have to set up some elementary facts about the Fourier transform of functions defined on an interval, which we shall choose throughout as $[0, 2\pi]$.

$\mathbb{T}$ denotes the torus $\mathbb{R}/2\pi\mathbb{Z}$, and often it will be convenient to consider $\mathbb{T}$ as the interval $[0, 2\pi]$ with its endpoints being identified. In a similar vein, we consider a function $f \in C(\mathbb{T})$ also as a continuous function on $[0, 2\pi]$, satisfying the condition $f(0) = f(2\pi)$, or as a function in $C_b(\mathbb{R})$ which is periodic with period 2π.

$\mathcal{B}(\mathbb{T})$ is the Borel–σ–algebra over $\mathbb{T}$. For $f \in C(\mathbb{T})$, $\hat{f} = (\hat{f}(n),\, n \in \mathbb{Z})$ denotes its *Fourier transform*: [1]

$$\hat{f}(n) = \frac{1}{2\pi} \int_{\mathbb{T}} f(t)\, e^{int}\, dt, \qquad n \in \mathbb{Z}. \tag{D.1}$$

Similarly, for $\mu \in \mathcal{M}(\mathbb{T})$ we set

$$\hat{\mu}(n) = \frac{1}{2\pi} \int_{\mathbb{T}} e^{int}\, d\mu(t), \qquad n \in \mathbb{Z}. \tag{D.2}$$

Let $P(\mathbb{T})$ denote the space of trigonometric polynomials f on $\mathbb{T}$:

$$f(t) = \sum_{n\in\mathbb{Z}} a_n\, e^{-int}, \qquad t \in \mathbb{T},$$

for some sequence $a = (a_n,\, n \in \mathbb{Z})$ in $\mathbb{C}^{\mathbb{Z}}$ which is such that at most finitely many elements are nonzero. It is an easy exercise to calculate the Fourier transform $\hat{f}$ of such an $f \in P(\mathbb{T})$, and we get that $\hat{f}(n) = a_n$ for all $n \in \mathbb{Z}$. In particular, the Fourier transform of a trigonometric polynomial is a complex sequence with at most finitely many non-vanishing elements.

[1] There are many different conventions for the Fourier transform in the literature. This should be taken into account when results and formulas from different sources are compared.

For two functions $f, g \in C(\mathbb{T})$ we define their *convolution* $f * g \in C(\mathbb{T})$ by

$$f * g(t) = \frac{1}{2\pi} \int_{\mathbb{T}} f(t-s)g(s)\,ds, \qquad t \in \mathbb{T}. \tag{D.3}$$

It is straightforward to check that $f * g = g * f$, and that

$$(f * g)\hat{}(n) = \hat{f}(n)\,\hat{g}(n), \qquad n \in \mathbb{Z}, \tag{D.4}$$

holds true: *Exercise!* (for this it is convient to consider f and g as 2π–periodic functions in $C_b(\mathbb{R})$).

Recall the Féjer kernel (C.2) $K_n \in P(\mathbb{T})$, $n \in \mathbb{N}$, from appendix C, which we prefer to write in this section as

$$K_n(t) = \sum_{k=-n}^{k} \Big(1 - \frac{|k|}{n+1}\Big) e^{-ikt}. \tag{D.5}$$

Thus we have

$$\hat{K}_n(k) = \begin{cases} 1 - \frac{|k|}{n+1}, & \text{if } |k| \le n, \\ 0, & \text{otherwise,} \end{cases} \tag{D.6}$$

and for $f \in C(T)$ we find

$$K_n * f(t) = \sum_{k=-n}^{n} \Big(1 - \frac{|k|}{n+1}\Big) \hat{f}(k) e^{-ikt},$$

and therefore

$$(K_n * f)\hat{}(k) = \begin{cases} \big(1 - \frac{|k|}{n+1}\big) \hat{f}(k), & \text{if } |k| \le n, \\ 0, & \text{otherwise.} \end{cases}$$

Similarly, for a measure $\mu \in \mathcal{M}(\mathbb{T})$, we define $K_n * \mu \in P(\mathbb{T})$ by ($t \in \mathbb{T}$)

$$K_n * \mu(t) = \frac{1}{2\pi} \int_{\mathbb{T}} K_n(t-s)\,d\mu(s) = \sum_{k=-n}^{n} \Big(1 - \frac{|k|}{n+1}\Big) \hat{\mu}(k)\, e^{-ikt}, \tag{D.7}$$

and get

$$(K_n * \mu)\hat{}(k) = \begin{cases} \big(1 - \frac{|k|}{n+1}\big)\, \hat{\mu}(k), & \text{if } |k| \le n, \\ 0, & \text{otherwise.} \end{cases}$$

Féjer's theorem, theorem C.4, states that for every $f \in C(\mathbb{T})$,

$$\|K_n * f - f\|_\infty \to 0, \quad n \to \infty.$$

For $f \in C(\mathbb{T})$, $\mu \in \mathcal{M}(\mathbb{T})$ define

$$\langle \mu, f \rangle = \frac{1}{2\pi} \int_{\mathbb{T}} f(t)\, d\overline{\mu}(t). \tag{D.8}$$

Hence

$$\begin{aligned}
\langle \mu, f \rangle &= \lim_{n\to\infty} \frac{1}{2\pi} \int_0^{2\pi} K_n * f(t)\, d\overline{\mu}(t) \\
&= \lim_{n\to\infty} \sum_{k=-n}^{n} \Big(1 - \frac{|k|}{n+1}\Big) \hat{f}(k) \frac{1}{2\pi} \int_0^{2\pi} e^{-ikt}\, d\overline{\mu}(t) \\
&= \lim_{n\to\infty} \sum_{k=-n}^{n} \overline{\hat{\mu}(k)} \Big(1 - \frac{|k|}{n+1}\Big) \hat{f}(k),
\end{aligned}$$

and we have proved the first statement of

D.3 Theorem (Parseval's Formula) *For all $f \in C(\mathbb{T})$, $\mu \in \mathcal{M}(\mathbb{T})$,* Parseval's formula

$$\langle \mu, f \rangle = \lim_{n\to\infty} \sum_{k=-n}^{n} \overline{\hat{\mu}(k)} \Big(1 - \frac{|k|}{n+1}\Big) \hat{f}(k) \tag{D.9}$$

holds true. In particular, if $\sum_{k\in\mathbb{Z}} \overline{\hat{\mu}(k)} \hat{f}(k)$ converges absolutely , then

$$\langle \mu, f \rangle = \sum_{k\in\mathbb{Z}} \overline{\hat{\mu}(k)} \hat{f}(k). \tag{D.10}$$

Proof It remains to prove the last statement. Suppose that $\sum_{k\in\mathbb{Z}} |\hat{f}(k)\hat{\mu}(k)| < +\infty$. Then $|\hat{f}(k)\hat{\mu}(k)|$, $k \in \mathbb{Z}$, is a summable majorant of

$$\overline{\hat{\mu}(k)} \Big(1 - \frac{|k|}{n+1}\Big) \hat{f}(k), \qquad k \in \mathbb{Z},$$

which is uniform in $n \in \mathbb{N}$. For example, we may now use the dominated convergence theorem A.56 to take the limit $n \to +\infty$ under the sum, and the last statement of the theorem follows. □

In particular if f is a trigonometric polynomial we obtain formula (D.10), because then the series on the right hand side of (D.10) has at most a finite number of non-zero terms, and therefore converges absolutely. Thus for $\mu \in \mathcal{M}(\mathbb{T})$, $f \in C(\mathbb{T})$, $n \in \mathbb{N}$, we can compute as follows

$$\langle \mu, K_n * f \rangle = \sum_{k\in\mathbb{Z}} \overline{\hat{\mu}(k)} \Big(1 - \frac{|k|}{n+1}\Big) \hat{f}(k) = \langle K_n * \mu, f \rangle, \tag{D.11}$$

where we identify the trigonometric polynomial $K_n * \mu$ with the measure defined by $K_n * \mu(t)\, dt$, $t \in \mathbb{T}$. Thus we obtain

D.4 Corollary *Suppose that $\mu \in \mathcal{M}(\mathbb{T})$. Then the sequence $(\mu_n,\ n \in \mathbb{N})$ of measures in $\mathcal{M}(\mathbb{T})$ defined by*

$$\mu_n(dt) = K_n * \mu(t)\, dt, \qquad t \in \mathbb{T}, \tag{D.12}$$

converges in weak–$*$ sense to μ.

D.5 Definition A sequence $a = (a_n,\ n \in \mathbb{Z})$ is called *positive definite*, if for every $n \in \mathbb{N}$, and all choices of $z_1, \dots, z_n \in \mathbb{C}$,

$$\sum_{k,l=0}^{n} a_{k-l}\, z_k \overline{z_l} \geq 0 \tag{D.13}$$

holds true.

D.6 Theorem (Herglotz) *A complex sequence $a = (a_n,\ n \in \mathbb{N})$ is positive definite, if and only if, there exists a positive measure $\mu \in \mathcal{M}(\mathbb{T})$ so that a is the Fourier transform of μ: $a_n = \hat{\mu}(n)$ for all $n \in \mathbb{Z}$.*

The proof is done in a sequence of lemmas. In the first two of them we characterize those complex sequences a which are Fourier transforms of complex measures on the torus $\mathbb{T}$.

D.7 Lemma *A complex sequence $a = (a_n,\ n \in \mathbb{Z})$ is the Fourier transform of a measure $\mu \in \mathcal{M}(\mathbb{T})$, if and only if there exists a constant $C \geq 0$ so that for every trigonometric polynomial f the inequality*

$$\Big|\sum_{n\in\mathbb{Z}} \overline{a_n}\, \hat{f}(n)\Big| \leq C\|f\|_\infty \tag{D.14}$$

holds.

Proof Suppose first that a is the Fourier transform of $\mu \in \mathcal{M}(\mathbb{T})$: $a_n = \hat{\mu}(n), n \in \mathbb{Z}$. Riesz' representation theorem, theorem D.2, gives for all $f \in C(\mathbb{T})$,

$$|\langle \mu, f\rangle| \leq \|\mu\|_{\mathcal{M}(\mathbb{T})}\|f\|_\infty.$$

So we set $C = \|\mu\|_{\mathcal{M}(\mathbb{T})}$. This holds true in particular for f being a trigonometric polynomial, but for such an f we immediately get from formula (D.10) the estimate (D.14).

For the converse implication assume that $(a_n,\ n \in \mathbb{Z})$ is such that (D.14) holds for some $C \geq 0$ and all trigonometric polynomials f. Consider the mapping Φ from the space of trigonometric polynomials into $\mathbb{C}$ defined by

$$\Phi(f) = \sum_{n\in\mathbb{Z}} \overline{a_n}\, \hat{f}(n).$$

Then this is a linear mapping bounded by C relative to the supremum norm. Since by Weierstraß' theorem C.2 the trigonometric polynomials are dense in $C(\mathbb{T})$ with respect to $\|\cdot\|_\infty$, Φ has a unique extension to $C(\mathbb{T})$ with norm less than C. [2] Now

[2] For example, this follows from the so-called *B.L.T.–theorem*, e.g., [29]. But this is also easily seen directly: Let $f \in C(\mathbb{T})$, and let $(f_n,\ n \in \mathbb{N})$ be a sequence of trigonometric polynomials converging to f. Thus they form a Cauchy sequence with respect to $\|\cdot\|_\infty$. Thus $\Phi(f_n - f_m)$ converges to zero as m, $n \to \infty$. Since $\mathbb{C}$ is complete, the limit of $(\Phi(f_n),\ n \in \mathbb{N})$ exists, and we set $\Phi(f) = \lim_n \Phi(f_n)$. In an *exercise* the reader checks that the extension is well-defined, unique and satisfies $\|\Phi\| \leq C$. In fact, the same argument gives the proof of the B.L.T.–theorem in the general case.

we apply the representation theorem of Riesz, theorem D.2, to conclude that there exists a measure $\mu \in \mathcal{M}(\mathbb{T})$ so that for every trigonometric polynomial f

$$\langle \mu, f \rangle = \frac{1}{2\pi} \int_{\mathbb{T}} f \, d\overline{\mu} = \sum_{n \in \mathbb{Z}} \overline{a_n} \, \hat{f}(n).$$

Choosing in particular $f(t) = \exp(-int)$, $t \in \mathbb{T}$, we get that $\hat{\mu}(n) = a_n$ for every $n \in \mathbb{Z}$. □

Let a complex sequence $a = (a_k, k \in \mathbb{Z})$ be given. For every $n \in \mathbb{N}$ define a complex measure $\mu_n(a)$ by

$$\mu_n(a)(dt) = \sum_{k=-n}^{n} \Big(1 - \frac{|k|}{n+1}\Big) a_k \, e^{-ikt} \, dt. \tag{D.15}$$

Suppose that a is such that the sequence $(\|\mu_n(a)\|_{\mathcal{M}(\mathbb{T})}, n \in \mathbb{N})$ is uniformly bounded, say by $C \geq 0$. Then from Parseval's formulas (D.9), (D.10), theorem D.3, we obtain for every trigonometric polynomial $f \in P(\mathbb{T})$,

$$\begin{aligned} \lim_{n\to\infty} \langle \mu_n(a), f \rangle &= \lim_{n\to\infty} \sum_{k\in\mathbb{Z}} \overline{a_k} \, \hat{K}_n(k) \hat{f}(k) \\ &= \sum_{k\in\mathbb{Z}} \overline{a_k} \hat{f}(k). \end{aligned}$$

Consequently

$$\Big| \sum_{k\in\mathbb{Z}} \overline{a_k} \hat{f}(k) \Big| \leq C \|f\|_\infty.$$

An application of lemma D.7 gives

D.8 Lemma *If the complex sequence $a = (a_n, n \in \mathbb{Z})$ is such that there exists a constant $C \geq 0$ so that $\|\mu_n(a)\|_{\mathcal{M}(\mathbb{T})} \leq C$ for all $n \in \mathbb{N}$, where $\mu_n(a)$ is the measure defined in* (D.15)*, then it is the Fourier transform of a measure $\mu \in \mathcal{M}(\mathbb{T})$.*

D.9 Remark Actually the converse statement is also true: Suppose that $a = (a_k, k \in \mathbb{Z})$ is the Fourier transform of $\mu \in \mathcal{M}(\mathbb{T})$: $a_k = \hat{\mu}(k)$, $k \in \mathbb{Z}$. Let $\mu_n(a)$ be defined as in (D.15). Then actually, $\mu_n(a) = K_n * \mu$, and corollary D.4 states that μ is the weak–$*$ limit of $(\mu_n(a), n \in \mathbb{N})$. Therefore, as already mentioned in section D.1, the uniform boundedness principle implies that the sequence $(\|\mu(a)_n\|_{\mathcal{M}(\mathbb{T})}, n \in \mathbb{N})$ is bounded (e.g., [20, Theorem III.1.28]).

D.10 Lemma *A complex sequence $a = (a_n, n \in \mathbb{Z})$ is the Fourier transform of a positive measure $\mu \in \mathcal{M}(\mathbb{T})$, if and ony if, for every $n \in \mathbb{N}$ the measure $\mu_n(a)$ defined by* (D.15) *is positive.*

Proof Assume that $a = \hat{\mu}$ for a positive measure $\mu \in \mathcal{M}(\mathbb{T})$. Then the measure $\mu_n(a)$ in (D.15) actually is equal to $K_n * \mu$, see (D.7). Let $f \in C(\mathbb{T})$ be positive, then obtain with (D.11)

$$\begin{aligned}\langle \mu_n(a), f\rangle &= \langle K_n * \mu, f\rangle \\ &= \langle \mu, K_n * f\rangle \\ &= \frac{1}{2\pi}\int_{\mathbb{T}} K_n * f(t)\, d\mu(t) \\ &\geq 0,\end{aligned}$$

because $K_n * f$ is a positive function: By hypothesis f is positive, and the Féjer kernel K_n is positive, too, see (C.3a) and (C.3b). Hence also their convolution is positive.

For the converse suppose that $\mu_n(a)$ is a positive measure for every $n \in \mathbb{N}$. Then

$$\|\mu_n(a)\|_{\mathcal{M}(\mathbb{T})} = \int_0^{2\pi} \sum_{k=-n}^{n} \Big(1 - \frac{|k|}{n+1}\Big)\, a_k\, e^{-ikt}\, dt = a_0.$$

Therefore the sequence $(\|\mu_n(a)\|_{\mathcal{M}(\mathbb{T})},\ n \in \mathbb{N})$ is bounded (by a_0), and lemma D.8 implies that $(a_k,\ k \in \mathbb{Z})$ is the Fourier transform of a measure $\mu \in \mathcal{M}(\mathbb{T})$. It remains to show that μ is positive. Since $a = \hat{\mu}$, we have again that $\mu_n(a) = K_n * \mu$, $n \in \mathbb{N}$, so that μ is the weak–$*$ limit of the sequence $(\mu_n(a),\ n \in \mathbb{N})$. Thus for every $f \in C(\mathbb{T})$ which is positive we find

$$\langle \mu, f\rangle = \lim_{n\to\infty} \langle \mu_n(a), f\rangle \geq 0.$$

Consequently μ is a positive measure. □

Finally we come to the proof of Herglotz' theorem:

Proof (of theorem D.6) Suppose that $a = (a_k,\ k \in \mathbb{Z})$ is the Fourier transform $\hat{\mu}$ of a positive measure $\mu \in \mathcal{M}(\mathbb{T})$. Let $n \in \mathbb{N}$, $z_1, \dots, z_n \in \mathbb{C}$ be given, then

$$\begin{aligned}\sum_{k,l=1}^{n} a_{k-l}\, z_k\, \overline{z_l} &= \sum_{k,l=1}^{n} \hat{\mu}_{k-l}\, z_k\, \overline{z_l} \\ &= \frac{1}{2\pi}\sum_{k,l=1}^{n} z_k\, \overline{z_l} \int_{\mathbb{T}} e^{i(k-l)t}\, d\mu(t) \\ &= \frac{1}{2\pi}\int_{\mathbb{T}} \Big|\sum_{k=1}^{n} z_k e^{ikt}\Big|^2\, d\mu(t) \\ &\geq 0.\end{aligned}$$

Therefore a is positive definite.

Now assume that a is positive definite. For $n \in \mathbb{N}$, $z_k = \exp(-ikt)$, $k = 0, 1, \ldots, n$, and $t \in \mathbb{T}$ one has

$$\begin{aligned} 0 &\le \sum_{k,l=0}^{n} a_{k-l}\, z_k\, \overline{z_l} \\ &= \sum_{k=-n}^{n} (n+1-|k|)\, a_k\, e^{-ikt} \\ &= (n+1) \sum_{k=-n}^{n} \Big(1 - \frac{|k|}{n+1}\Big)\, a_k\, e^{-ikt}, \end{aligned}$$

where we used formula (C.5). Therefore the measures $\mu_n(a)$, $n \in \mathbb{N}$, defined by (D.15) are positive measures. An application of lemma D.10 concludes the proof. □

D.3 Bochner's Theorem

We begin this section by collecting without proofs some basic facts about the Fourier transform on $\mathbb{R}$. For proofs and more results the interested reader may consult, e.g., [21], [30, Chapter IX] or [14].

For $\mathcal{L}^1(\mathbb{R}, \mathcal{B}(\mathbb{R}), \lambda)$, λ the Lebesgue measure, we shall simply write $\mathcal{L}^1(\mathbb{R})$ in the sequel. For $f \in \mathcal{L}^1(\mathbb{R})$ define its *Fourier transforms*

$$\begin{aligned} \mathcal{F} f(\xi) &= \int_{\mathbb{R}} f(x)\, e^{i\xi x}\, dx, \\ \mathcal{F}^{-1} f(\xi) &= \frac{1}{2\pi} \int_{\mathbb{R}} f(x)\, e^{-i\xi x}\, dx. \end{aligned}$$

$\mathcal{S}(\mathbb{R})$ denotes the *Schwartz space* of rapidly decreasing, smooth functions f on $\mathbb{R}$: f is infinitely differentiable, and for all $m, n \in \mathbb{N}_0$,

$$\lim_{|x| \to +\infty} x^m f^{(n)}(x) = 0,$$

where $f^{(n)}$ denotes the n–th derivative of f. Clearly, $C_c(\mathbb{R}) \subset \mathcal{S}(\mathbb{R})$.

D.11 Theorem *The mappings $\mathcal{F}$ and $\mathcal{F}^{-1}$ are bijections from $\mathcal{S}(\mathbb{R})$ onto itself. Moreover, on $\mathcal{S}(\mathbb{R})$ $\mathcal{F}$ and $\mathcal{F}^{-1}$ are inverses of each other: $\mathcal{F} \circ \mathcal{F}^{-1} = \mathcal{F}^{-1} \circ \mathcal{F} = id$.*

D.12 Remark The Schwartz space $\mathcal{S}(\mathbb{R})$ is usually equipped with a natural (Fréchet) topology, and one can show that $\mathcal{F}$ and $\mathcal{F}^{-1}$ are *bicontinuous* bijections with respect to this topology. However, we will not need this here.

Notation In the sequel we shall write $\hat{f}$ for $\mathcal{F} f$.

$\mathcal{M}(\mathbb{R})$ denotes the space of all finite complex measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Thus, by Riesz' theorem D.2, $\mathcal{M}(\mathbb{R})$ is the dual space of $(C_0(\mathbb{R}), \|\cdot\|_\infty)$, and we have

$$|\langle \mu, f\rangle| \le \|\mu\|_{\mathcal{M}(\mathbb{R})} \|f\|_\infty,$$

for all $\mu \in \mathcal{M}(\mathbb{R})$, $f \in C_0(\mathbb{R})$, where

$$\langle \mu, f\rangle = \int_{\mathbb{R}} f(x)\, d\overline{\mu}(x).$$

For $\mu \in \mathcal{M}(\mathbb{R})$ its *Fourier transform* is defined by

$$\hat{\mu}(\xi) = \int_{\mathbb{R}} e^{i\xi x}\, d\mu(x),$$

Let $f \in \mathcal{S}(\mathbb{R})$, and $\mu \in \mathcal{M}(\mathbb{R})$, then

$$\int f\, d\overline{\mu} = \frac{1}{2\pi} \int \Big(\int \hat{f}(\xi) e^{-i\xi x}\, d\xi \Big)\, d\overline{\mu}(x) = \frac{1}{2\pi} \int \hat{f}(\xi)\, \overline{\hat{\mu}(\xi)}\, d\xi.$$

Thus in the context of the real line we get *Parseval's formula* in the form

$$\langle \mu, f\rangle = \frac{1}{2\pi} \int \hat{f}(\xi)\, \overline{\hat{\mu}(\xi)}\, d\xi. \tag{D.16}$$

D.13 Lemma *For every $\mu \in \mathcal{M}(\mathbb{R})$, $\hat{\mu}$ is bounded and uniformly continuous on $\mathbb{R}$.*

Proof Since μ is a complex linear combination of finite positive measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, the same arguments as in the proof of statement (a) of theorem 6.15 gives the continuity. Moreover, for $\xi \in \mathbb{R}$ we find that

$$|\hat{\mu}(\xi)| = \Big| \int_{\mathbb{R}} e^{i\xi x}\, d\mu(x) \Big| \le |\mu|(\mathbb{R}) = \|\mu\|_{\mathcal{M}(\mathbb{R})},$$

and therefore $\hat{\mu}$ is bounded. □

Recall from definition 6.13, chapter 6, that a complex valued function φ on $\mathbb{R}$ is called *positive definite*, if

$$\sum_{k,l=1}^{n} \varphi(x_k - x_l)\, z_k\, \overline{z_l} \ge 0,$$

for every choice of $n \in \mathbb{N}$, $x_1, \dots, x_n \in \mathbb{R}$, $z_1, \dots, z_n \in \mathbb{C}$.

D.14 Theorem (Bochner) *A complex valued function φ on $\mathbb{R}$ is the Fourier transform of a finite positive measure, if and only if it is positive definite and continuous.*

We reduce the proof of Bochner's theorem through a sequence of lemmas to an application of Herglotz' theorem D.6.

D.15 Lemma *$\varphi \in C_b(\mathbb{R})$ is the Fourier transform of a measure $\mu \in \mathcal{M}(\mathbb{R})$, if and only if there exists a constant $C \geq 0$, so that*

$$\left|\frac{1}{2\pi}\int \hat{f}(\xi)\,\overline{\varphi(\xi)}\,d\xi\right| \leq C\|f\|_\infty \tag{D.17}$$

for every $f \in \mathcal{S}(\mathbb{R})$.

Proof Assume that $\varphi = \hat{\mu}$, $\mu \in \mathcal{M}(\mathbb{R})$. Clearly, we have that $|\varphi(\xi)| \leq \|\mu\|_{\mathcal{M}(\mathbb{R})}$ for all $\xi \in \mathbb{R}$, so that φ is bounded. From Parseval's formula (D.16) we get

$$\begin{aligned}\left|\frac{1}{2\pi}\int \hat{f}(\xi)\,\overline{\varphi(\xi)}\,d\xi\right| &= \left|\frac{1}{2\pi}\int \hat{f}(\xi)\,\overline{\hat{\mu}(\xi)}\,d\xi\right| \\ &= \left|\int f\,d\mu\right| \\ &\leq \|\mu\|_{\mathcal{M}(\mathbb{R})}\,\|f\|_\infty.\end{aligned}$$

So (D.17) holds with $C = \|\mu\|_{\mathcal{M}(\mathbb{R})}$.

For the converse assume that φ is continuous, and that (D.17) holds true for all $f \in \mathcal{S}(\mathbb{R})$. Then the mapping

$$f \mapsto \frac{1}{2\pi}\int \hat{f}(\xi)\,\overline{\varphi(\xi)}\,d\xi$$

defines a linear functional on $\mathcal{S}(\mathbb{R})$, which is continuous relative to the supremum norm $\|\cdot\|_\infty$ on $\mathcal{S}(\mathbb{R})$. Since $\mathcal{S}(\mathbb{R})$ is dense in $C_0(\mathbb{R})$ with respect to $\|\cdot\|_\infty$, [3] this functional extends with the same bound to all of $C_0(\mathbb{R})$. By Riesz representation theorem D.2 there exists $\mu \in \mathcal{M}(\mathbb{R})$ such that for all $f \in \mathcal{S}(\mathbb{R})$ we get the equality

$$\int f\,d\mu = \frac{1}{2\pi}\int \hat{f}(\xi)\,\overline{\varphi(\xi)}\,d\xi.$$

Parseval's formula (D.16) yields

$$\frac{1}{2\pi}\int \hat{f}(\xi)\,\overline{\hat{\mu}(\xi)}\,d\xi = \frac{1}{2\pi}\int \hat{f}(\xi)\,\overline{\varphi(\xi)}\,d\xi,$$

for all $f \in \mathcal{S}(\mathbb{R})$. But since φ and $\hat{\mu}$ are bounded and continuous (see lemma D.13), we find that $\varphi(\xi) = \hat{\mu}(\xi)$ for all $\xi \in \mathbb{R}$. (For example, we may choose for f the density of the normal distribution $\mathcal{N}(\eta, 1/n)$, $n \in \mathbb{N}$, $\eta \in \mathbb{R}$. Then exercise 5.2.(c) shows that $\varphi(\eta) = \hat{\mu}(\eta)$.) □

D.16 Lemma *$\varphi \in C_b(\mathbb{R})$ is the Fourier transform of a positive measure μ in $\mathcal{M}(\mathbb{R})$, if and only if*

$$\int_{\mathbb{R}} \hat{f}(\xi)\,\overline{\varphi(\xi)}\,d\xi \geq 0, \tag{D.18}$$

for all $f \in \mathcal{S}(\mathbb{R})$ with $f \geq 0$.

[3] The density of $\mathcal{S}(\mathbb{R})$ in $C_0(\mathbb{R})$ with respect to $\|\cdot\|_\infty$ can be proved with the standard techniques of functional analysis, see, e.g., [1, Chapter II]. Details are left as an *exercise* to the interested reader.

Proof If $\varphi = \hat{\mu}$ for a positive finite measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, then Parseval's formula (D.16) shows that for every $f \in \mathcal{S}(\mathbb{R})$ with $f \geq 0$,

$$\frac{1}{2\pi} \int_{\mathbb{R}} \hat{f}(\xi)\, \overline{\varphi(\xi)}\, d\xi = \frac{1}{2\pi} \int_{\mathbb{R}} \hat{f}(\xi)\, \overline{\hat{\mu}(\xi)}\, d\xi = \int_{\mathbb{R}} f(x)\, d\mu(x) \geq 0$$

holds true.

For the proof of the converse statement choose $f_n \in \mathcal{S}(\mathbb{R})$ as given by $\gamma_n(x) = \exp(-x^2/2n)$, $x \in \mathbb{R}$, $n \in \mathbb{N}$, so that $\hat{\gamma}_n$ is the density of $\mathcal{N}(0, 1/n)$. Then we obtain from exercise 5.2 that

$$\varphi(0) = \lim_{n\to\infty} \int_{\mathbb{R}} \hat{\gamma}_n(\xi)\, \overline{\varphi(\xi)}\, d\xi \geq 0.$$

On the other hand, γ_n converges to 1 as $n \to +\infty$, uniformly on every compact. Thus for $f \in C_c(\mathbb{R})$, $f \geq 0$, and $\varepsilon > 0$ we find for all $n \in \mathbb{N}$ large enough that

$$g_n(x) = (\varepsilon + \|f\|_\infty)\gamma_n(x) - f(x) \geq 0, \qquad x \in \mathbb{R},$$

and clearly g_n belongs to $\mathcal{S}(\mathbb{R})$. We use g_n in the hypothesis and find

$$\begin{aligned} 0 &\leq \int_{\mathbb{R}} \hat{g}_n(\xi)\, \overline{\varphi(\xi)}\, d\xi \\ &= \int_{\mathbb{R}} \big((\varepsilon + \|f\|_\infty)\hat{\gamma}(\xi) - \hat{f}(\xi)\big)\, \overline{\varphi(\xi)}\, d\xi \end{aligned}$$

which implies

$$\int_{\mathbb{R}} \hat{f}(\xi)\, \overline{\varphi(\xi)}\, d\xi \leq (\varepsilon + \|f\|_\infty) \int_{\mathbb{R}} \hat{\gamma}_n(\xi)\, \overline{\varphi(\xi)}\, d\xi,$$

and for all n large enough therefore

$$\int_{\mathbb{R}} \hat{f}(\xi)\, \overline{\varphi(\xi)}\, d\xi \leq \varphi(0)(\varepsilon + \|f\|_\infty) + \varepsilon.$$

Repeating the same argument for f interchanged with $-f$ and observing that $\varepsilon > 0$ was arbitrary, we get

$$\left| \int_{\mathbb{R}} \hat{f}(\xi)\, \overline{\varphi(\xi)}\, d\xi \right| \leq \varphi(0)\|f\|_\infty.$$

Since $C_c(\mathbb{R})$ is dense in $\mathcal{S}(\mathbb{R})$ with respect to the supremum norm (cf. footnote 3), we get the last inequality for all real $f \in \mathcal{S}(\mathbb{R})$, and then with a factor 2 on the right hand side for all $f \in \mathcal{S}(\mathbb{R})$. Thus lemma D.15 implies that φ is the Fourier transform of a measure $\mu \in \mathcal{M}(\mathbb{R})$, and Parseval's formula (D.16) entails that μ is positive. □

Now we reformulate the criteria of the lemmas D.15, D.16 on the torus $\mathbb{T}$. To this end, we first prove the following elementary, but rather technical lemma. For $f \in C_0(\mathbb{R})$, $\lambda > 0$, set

$$f_{\mathbb{T},\lambda}(x) = \sum_{k\in\mathbb{Z}} f\Big(\frac{x + 2\pi k}{\lambda}\Big), \qquad x \in \mathbb{R}, \tag{D.19}$$

and for $\mu \in \mathcal{M}(\mathbb{R})$ set

$$\mu_{\mathbb{T},\lambda}(A) = \sum_{k\in\mathbb{Z}} \mu\big(\lambda^{-1}(A + 2\pi k)\big), \qquad A \in \mathcal{B}(\mathbb{T}). \tag{D.20}$$

Then we have for every $A \in \mathcal{B}(\mathbb{T})$,

$$\mu_{\mathbb{T},\lambda}(A) = \mu\Big(\biguplus_{k\in\mathbb{Z}} \lambda^{-1}(A + 2\pi k\Big),$$

and therefore $\mu_{\mathbb{T},\lambda}$ is a measure in $\mathcal{M}(\mathbb{T})$. Moreover,

$$\|\mu_{\mathbb{T},\lambda}\|_{\mathcal{M}(\mathbb{T})} = |\mu_{\mathbb{T},\lambda}|(\mathbb{T}) = |\mu(\mathbb{R})| = \|\mu\|_{\mathcal{M}(\mathbb{R})}. \tag{D.21}$$

D.17 Lemma *Suppose that $f \in C_0(\mathbb{R})$ is such that $x^2 f(x) \to 0$ as $|x| \to +\infty$.*

(a) *For every $\lambda \in (0, 1]$ the series* (D.19) *converges absolutely, uniformly in x on every compact. Therefore $f_{\mathbb{T},\lambda}$ is a continuous function which is periodic with period 2π, that is, a continuous function on $\mathbb{T}$. Furthermore, for $\varepsilon > 0$ there exists a $\lambda_0 \in (0, 1]$ so that for every $\lambda \in (0, \lambda_0]$,*

$$\|f_{\mathbb{T},\lambda}\|_\infty \le \|f\|_\infty + \varepsilon. \tag{D.22}$$

(b) *The Fourier transform of $f_{\mathbb{T},\lambda}$ on the torus is given by*

$$(f_{\mathbb{T},\lambda})\hat{}(n) = \frac{\lambda}{2\pi} \hat{f}(n\lambda), \qquad n \in \mathbb{Z} \tag{D.23}$$

and the Fourier transform of $\mu_{\mathbb{T},\lambda}$ on $\mathbb{T}$ is equal to

$$(\mu_{\mathbb{T},\lambda})\hat{}(n) = \frac{1}{2\pi} \hat{\mu}(n\lambda). \tag{D.24}$$

(c) *For $\lambda > 0$ set*

$$R_\lambda(f) = \lambda \sum_{k\in\mathbb{Z}} f(\lambda k). \tag{D.25}$$

Then $R_\lambda(f)$ is a Riemann approximation to the integral of f over $\mathbb{R}$, that is, for every $\varepsilon > 0$ there exists a $\lambda_0 > 0$ so that for every $\lambda \in (0, \lambda_0]$,

$$\Big|\int_{\mathbb{R}} f(x)\,dx - R_\lambda(f)\Big| \le \varepsilon \tag{D.26}$$

holds.

Proof For the convenience of the reader rather detailed arguments are provided, even though they are based on rather basic analysis.

By hypothesis, for every $\delta > 0$ there exists $n \in \mathbb{N}$ so that

$$|f(y)| \le \delta y^{-2}, \qquad y \in \mathbb{R},\ |y| \ge n\pi. \tag{D.27}$$

For the proof of (a) choose the compact set $K = [-m\pi, m\pi]$, $m \in \mathbb{N}$, and assume that $x \in K$. Let $\varepsilon > 0$ be given, and choose $\delta = \varepsilon$ in (D.27). Furthermore choose $k_0 \in \mathbb{N}$ such that $2k_0 \geq m + n$. Then we have for all $x \in K$, $k \in \mathbb{N}$ with $k \geq k_0$, $\lambda \in (0, 1]$, that $\lambda^{-1}|x + 2k\pi| \geq n\pi$, and therefore

$$\sum_{k\in\mathbb{Z},\,|k|\geq k_0} \Big| f\Big(\frac{x+2\pi k}{\lambda}\Big)\Big| \leq 2\varepsilon \sum_{k=k_0}^{\infty} \frac{\lambda^2}{(x+2\pi k)^2} \leq \frac{2\varepsilon}{\pi^2} \sum_{k=k_0}^{\infty} \frac{1}{(2k-m)^2}.$$

In order to estimate the last sum, we use the following (rough, but for our purposes sufficient) bound

$$\sum_{k=k_0}^{\infty} \frac{1}{(ak+b)^2} \leq \frac{a+1}{a}\,\frac{1}{ak_0+b} \leq \frac{a+1}{a}, \tag{D.28}$$

which is valid, whenever $a > 0$, $b \in \mathbb{R}$, $k_0 \in \mathbb{N}$, $ak_0 + b \geq 1$. This estimate is easy to see:

$$\begin{aligned}\sum_{k=k_0}^{\infty} \frac{1}{(ak+b)^2} &= \frac{1}{(ak_0+b)^2} + \sum_{k=k_0+1}^{\infty} \frac{1}{(ak+b)^2}\\ &\leq \frac{1}{(ak_0+b)^2} + \int_{k_0}^{\infty} \frac{1}{(ax+b)^2}\,dx\\ &\leq \frac{1}{(ak_0+b)^2} + \frac{1}{a}\,\frac{1}{ak_0+b},\end{aligned}$$

and (D.28) immediately follows. Since above we have $2k_0 - m \geq n \geq 1$ we can apply this to find

$$\sum_{k=k_0}^{\infty} \frac{1}{(2k-m)^2} \leq \frac{3}{2},$$

so that

$$\sum_{k\in\mathbb{Z},\,|k|\geq k_0} \Big| f\Big(\frac{x+2\pi k}{\lambda}\Big)\Big| \leq \varepsilon.$$

This shows the absolute convergence of the series on the right hand side of (D.19), uniformly in $x \in K$. Thus $f_{\mathbb{T},\lambda}$ is indeed a continuous, 2π–periodic function on $\mathbb{R}$. In particular, in order to estimate its supremum, it is sufficient to to do this on any interval of length 2π, and we choose $[-\pi, \pi]$. So let $x \in [-\pi, \pi]$, $\delta = \varepsilon$, and choose $\lambda_0 = 1/n$, where n is as above. Then for every $k \in \mathbb{Z}$, $k \neq 0$, and $\lambda \in (0, \lambda_0]$,

$\lambda^{-1}|x + 2\pi k| \geq n\pi$, so that we may use the bound (D.27):

$$
\begin{aligned}
\left|f_{\mathbb{T},\lambda}(x)\right| &= \Big|f\Big(\frac{x}{\lambda}\Big) + \sum_{k\in\mathbb{Z},\, k\neq 0} f\Big(\frac{x+2\pi k}{\lambda}\Big)\Big| \\
&\leq \|f\|_\infty + 2\varepsilon \sum_{k=1}^{\infty} \frac{\lambda^2}{(x+2\pi k)^2} \\
&\leq \|f\|_\infty + \frac{2\varepsilon}{\pi^2} \sum_{k=1}^{\infty} \frac{1}{(2k-1)^2} \\
&\leq \|f\|_\infty + \frac{2\varepsilon}{\pi^2}\,\frac{3}{2} \\
&\leq \|f\|_\infty + \varepsilon,
\end{aligned}
$$

where we made again use of the bound (D.28). This proves (D.22).

By the definition of the Fourier transform of a measure in $\mathcal{M}(\mathbb{T})$ on $\mathbb{T}$ (see (D.2) we get for $n \in \mathbb{Z}$, $\lambda > 0$,

$$
\begin{aligned}
(\mu_{\mathbb{T},\lambda})\hat{}\,(n) &= \frac{1}{2\pi} \int_0^{2\pi} e^{inx}\, d\mu_{\mathbb{T},\lambda}(x) \\
&= \frac{1}{2\pi} \sum_{k\in\mathbb{Z}} \int_{(2\pi k, 2\pi(k+1)]} e^{in\lambda x}\, d\mu(x) \\
&= \frac{1}{2\pi} \int_{\mathbb{R}} e^{in\lambda x}\, d\mu(x) \\
&= \frac{\lambda}{2\pi}\, \hat{\mu}(n\lambda).
\end{aligned}
$$

If we choose $\mu(dx) = f(x)\,dx$, $x \in \mathbb{R}$, then we obtain formula (D.23) (of course, the additional factor λ on the right hand side of (D.23) comes from the transformation of the Lebesgue measure).

For the proof of (c), first choose $n \in \mathbb{N}$, $n \geq 2$, large enough so that

$$
\Big|\int_{\complement[-n\pi, n\pi]} f(x)\,dx\Big| \leq \frac{\varepsilon}{3}.
$$

(This is possible because f is integrable, but of course one can also use the estimate (D.27).) Then — if necessary — increase n so that (D.27) holds also true with $\delta = \varepsilon/3$. Next choose $m_0 \in \mathbb{N}$ so that the Riemann approximation of the integral of h over the the interval $[-n\pi, n\pi]$ is closer to the integral than $\varepsilon/3$, that is, for all $m \in \mathbb{N}$ with $m \geq m_0$

$$
\Big|\int_{-n\pi}^{n\pi} f(x)\,dx - \frac{n\pi}{m} \sum_{k=-m}^{m} f\Big(\frac{kn\pi}{m}\Big)\Big| \leq \frac{\varepsilon}{3}.
$$

Put $\lambda_0 = n\pi/m_0$, $\lambda = n\pi/m$. Then estimate as follows

$$\Big|\lambda \sum_{k\in\mathbb{Z},\, k\geq|m+1|} f(k\lambda)\Big| \leq \frac{2\lambda\varepsilon}{3} \sum_{k=m+1} \frac{1}{k^2\lambda^2} \leq 4\,\frac{\varepsilon}{3}\,\frac{1}{m\lambda} \leq \frac{\varepsilon}{3}.$$

Thus we obtain the estimate (D.26) for all $\lambda > 0$ of the form $\lambda = n\pi/m$ with $m \geq m_0$, and by continuity of the expression $R_\lambda(f)$ for all $\lambda \in (0, \lambda_0]$. $\square$

From now on we shall consider the function $f_{\mathbb{T},\lambda}$ defined in (D.19) as a continuous function on the torus $\mathbb{T} = \mathbb{R}/2\pi\mathbb{Z}$.

D.18 Corollary *Suppose that $\varphi \in C_b(\mathbb{R})$, and that $f \in \mathcal{S}(R)$. Then for every $\varepsilon > 0$ there exists $\lambda_0 > 0$ so that for all $\lambda \in (0, \lambda_0]$ the following estimate holds:*

$$\Big|\frac{1}{2\pi}\int_{\mathbb{R}} \hat{f}(\xi)\,\overline{\varphi(\xi)}\,d\xi - \sum_{n\in\mathbb{Z}} (f_{\mathbb{T},\lambda})\hat{}\,(n\lambda)\,\overline{\varphi(n\lambda)}\Big| \leq \varepsilon. \tag{D.29}$$

In particular, if for every $\lambda \in (0, \lambda_0]$ the sequence $(\varphi(n\lambda),\ n \in \mathbb{Z})$ is the Fourier transform of a measure $\mu_{\mathbb{T},\lambda}$ in $\mathcal{M}(\mathbb{T})$, then

$$\Big|\frac{1}{2\pi}\int_{\mathbb{R}} \hat{f}(\xi)\,\overline{\varphi(\xi)}\,d\xi - \langle \mu_{\mathbb{T},\lambda}, f_{\mathbb{T},\lambda}\rangle\Big| \leq \varepsilon. \tag{D.30}$$

Proof We apply the lemma to the function $\hat{f}\,\overline{\varphi}$, so that (see (D.25))

$$\begin{aligned}\frac{1}{2\pi}\,R_\lambda(\hat{f}\,\overline{\varphi}) &= \frac{\lambda}{2\pi}\sum_{k\in\mathbb{Z}} \hat{f}(\lambda k)\,\overline{\varphi(\lambda k)} \\ &= \sum_{k\in\mathbb{Z}} (f_{\mathbb{T},\lambda})\hat{}\,(k)\,\overline{\varphi(\lambda k)},\end{aligned}$$

by formula (D.23). Then estimate (D.26) gives the first statement. If $(\varphi(\lambda n),\ n \in \mathbb{Z})$ is the Fourier transform of $\mu_{\mathbb{T},\lambda} \in \mathcal{M}(\mathbb{T})$, then the last expression reads

$$\sum_{k\in\mathbb{Z}} (f_{\mathbb{T},\lambda})\hat{}\,(k)\,\overline{\hat{\mu}_{\mathbb{T},\lambda}(k)} = \langle \mu_{\mathbb{T},\lambda}, f_{\mathbb{T},\lambda}\rangle,$$

where we used Parseval's formula (D.10) on the torus. $\square$

D.19 Lemma *$\varphi \in C_b(\mathbb{R})$ is the Fourier transform of a measure in $\mathcal{M}(\mathbb{R})$, if and only if there exists a constant $C \geq 0$, so that for every $\lambda > 0$ the sequence $(\varphi(\lambda n),\ n \in \mathbb{Z})$ is the Fourier transform of a measure $\mu_{\mathbb{T},\lambda}$ in $\mathcal{M}(\mathbb{T})$ of norm less than C.*

Proof Assume that φ is the Fourier transform of $\mu \in \mathcal{M}(\mathbb{R})$. The we get from (D.24) that

$$\varphi(\lambda n) = \hat{\mu}(\lambda n) = 2\pi(\mu_{\mathbb{T},\lambda})\hat{}\,(n), \qquad n \in \mathbb{Z},$$

where $\mu_{\mathbb{T},\lambda}$ is given as in (D.20). From (D.21) we see that we may choose $C = \|\mu\|_{\mathcal{M}(\mathbb{R})}$.

Suppose that for every $\lambda > 0$, $(\varphi_\lambda(n),\ n \in \mathbb{Z})$ is the Fourier transform of $\mu_{\mathbb{T},\lambda} \in \mathcal{M}(\mathbb{T})$ with $\|\mu_{\mathbb{T},\lambda}\|_{\mathcal{M}(\mathbb{T})} \leq C$ for all $\lambda > 0$. Let $f \in \mathcal{S}(\mathbb{R})$, and note that $\hat{f} \in \mathcal{S}(\mathbb{R})$ (cf. theorem D.11). Hence by corollary D.18

$$\begin{aligned}\left|\frac{1}{2\pi}\int_{\mathbb{R}} \hat{f}(\xi)\,\overline{\varphi(\xi)}\,d\xi\right| &= |\langle \mu_\lambda, f_{\mathbb{T},\lambda}\rangle| + \varepsilon \\ &\leq C\|f_{\mathbb{T},\lambda}\|_\infty + \varepsilon.\end{aligned}$$

Next we use the second part of statement (a) of lemma D.17, to conclude that for all $\lambda > 0$ small enough we have

$$\|f_{T,\lambda}\|_\infty \leq \|f\|_\infty + \varepsilon.$$

Inserting this above, we obtain the estimate

$$\left|\frac{1}{2\pi}\int_{\mathbb{R}} \hat{f}(\xi)\,\overline{\varphi(\xi)}\,d\xi\right| \leq C\|f\|_\infty + \varepsilon(C+1),$$

that is,

$$\left|\frac{1}{2\pi}\int_{\mathbb{R}} \hat{f}(\xi)\,\overline{\varphi(\xi)}\,d\xi\right| \leq C\|f\|_\infty,$$

because $\varepsilon > 0$ was arbitrary. Lemma D.15 implies that φ is the Fourier transform of a measure $\mu \in \mathcal{M}(\mathbb{R})$. □

D.20 Lemma *$\varphi \in C_b(\mathbb{R})$ is the Fourier transform of a positive measure $\mu \in \mathcal{M}(\mathbb{R})$, if and only if for all $\lambda > 0$, the sequence $(\varphi(\lambda n),\ n \in \mathbb{N})$ is the Fourier transform of a positive measure on the torus $\mathbb{T}$.*

Proof Assume that $\varphi = \hat{\mu}$ for a positive measure $\mu \in \mathcal{M}(\mathbb{R})$. Let $\lambda > 0$, and consider the measure $\mu_{\mathbb{T},\lambda}$ on $\mathbb{T}$ defined in (D.20), which obviously is positive. Then equation (D.24) shows that $(\varphi(\lambda n),\ n \in \mathbb{Z})$ is the Fourier transform of the positive measure $\mu_{\mathbb{T},\lambda}$ on $\mathbb{T}$.

Now suppose that $\varphi \in C_b(\mathbb{R})$ is such that for each $\lambda > 0$ the sequence $(\varphi(\lambda n),\ n \in \mathbb{N})$ is the Fourier transform of a positive measure $\mu_{\mathbb{T},\lambda} \in \mathcal{M}(\mathbb{T})$. Let $f \in \mathcal{S}(\mathbb{R})$ with $f \geq 0$ be given. From corollary D.18, inequality (D.30) we find that for all $\lambda > 0$ small enough

$$\frac{1}{2\pi}\int_{\mathbb{R}} \hat{f}(\xi)\,\overline{\varphi(\xi)}\,d\xi \geq \langle \mu_{\mathbb{T},\lambda}, f_{\mathbb{T},\lambda}\rangle - \varepsilon \geq -\varepsilon,$$

because $f_{\mathbb{T},\lambda}$ and $\mu_{\mathbb{T},\lambda}$ are positive. Since $\varepsilon > 0$ was arbitrary, we get

$$\frac{1}{2\pi}\int_{\mathbb{R}} \hat{f}(\xi)\,\overline{\varphi(\xi)}\,d\xi \geq 0.$$

Now we apply lemma D.16 to conlcude that φ is the Fourier transform of a positive measure in $\mathcal{M}(\mathbb{R})$. □

Finally we come to the proof of Bochner's theorem D.14:

Proof (of theorem D.14) Suppose that φ is the Fourier transform of a positive measure $\mu \in \mathcal{M}(\mathbb{R})$. Let $n \in \mathbb{N}$, $\xi_1, \ldots, \xi_n \in \mathbb{R}$, $z_1, \ldots, z_n \in \mathbb{C}$. Then

$$\begin{aligned}\sum_{k,l=1}^{n} \varphi(\xi_k - \xi_l)\, z_k\, \overline{z_l} &= \sum_{k,l=1}^{n} z_k\, \overline{z_l} \int_{\mathbb{R}} e^{i(\xi_k - \xi_l)x}\, d\mu(x) \\ &= \int_{\mathbb{R}} \Big| \sum_{k=1}^{n} z_k\, e^{i\xi_k x} \Big|^2 d\mu(x) \\ &\geq 0.\end{aligned}$$

Conversely, assume that $\varphi \in C_b(\mathbb{R})$ is positive definite. Then it follows that for every $\lambda > 0$ the sequence $(\varphi(\lambda n),\, n \in \mathbb{Z})$ is a positive definite sequence, and therefore by Herglotz theorem D.6 it follows that it is the Fourier transform of a positive measure $\mu_{\mathbb{T},\lambda}$ on $\mathbb{T}$. Lemma D.20 implies that φ is the Fourier transform of a positive measure on $\mathbb{R}$. $\square$

Appendix E

Taylor's Theorem

In this appendix we prove a version of Taylor's theorem, which is not quite the common one treated in analysis courses. In fact, the version below makes use of a remainder term which — for an approximation with a Taylor polynomial of order $n \in \mathbb{N}$ — involves the n–th derivative of the function, while the more common Cauchy or Lagrange remainders have the derivative of order $n + 1$. This difference pays off in the proof of the central limit theorem. Moreover, the method of proof is different: Commonly one uses the generalized mean value theorem for derivatives, while here we use the fundamental theorem of calculus.

E.1 Theorem (Taylor) *Assume that f is a real or complex valued function defined on $\mathbb{R}$, which is n times continuously differentiable, $n \in \mathbb{N}$. Then for every $x \in \mathbb{R}$,*

$$f(x) = \sum_{k=0}^{n} f^{(k)}(0)\frac{x^k}{k!} + \frac{x^n}{n!}\,\theta_n(x), \tag{E.1a}$$

where

$$\theta_n(x) = n\int_0^1 (1-\tau)^{n-1}\big(f^{(n)}(\tau x) - f^{(n)}(0)\big)\,d\tau. \tag{E.1b}$$

Furthermore, θ_n is continuous in $x = 0$ with $\theta_n(0) = 0$.

Proof Using integration by parts, we immediately get the following formula for every $k \in \mathbb{N}$:

$$\int_0^1 \frac{(1-\tau)^{k-1}}{(k-1)!}\Big(\int_0^\tau f(u)\,du\Big)\,d\tau = \int_0^1 \frac{(1-\tau)^k}{k!}\,f(\tau)\,d\tau. \tag{E.2}$$

By the fundamental theorem of calculus we also find for $u \in [0, 1]$, $x \geq 0$,

$$f(ux) - f(0) = x\int_0^u f'(\tau x)\,d\tau. \tag{E.3}$$

The proof of Taylor's formula (E.1) is now just an iteration of these two formulas: We use $u = 1$ in (E.3) so that

$$\begin{aligned} f(x) &= f(0) + x \int_0^1 f'(\tau x)\, d\tau \\ &= f(0) + x \int_0^1 \Big(f'(0) + x \int_0^\tau f''(ux)\, du \Big)\, d\tau \\ &= f(0) + x f'(0) + x^2 \int_0^1 \frac{(1-\tau)^0}{0!} \Big(\int_0^\tau f''(ux)\, du \Big)\, d\tau \\ &= f(0) + x f'(0) + x^2 \int_0^1 \frac{(1-\tau)^1}{1!} f''(\tau x)\, d\tau, \end{aligned}$$

where we used (E.3) again in the second step, and in the last one we applied (E.2). Insert again (E.3) under the integral, and proceed iteratively as above. The following formula results

$$f(x) = \sum_{k=0}^{n-1} f^{(k)}(0)\, \frac{x^k}{k!} + x^n \int_0^1 \frac{(1-\tau)^{n-1}}{(n-1)!} f^{(n)}(\tau x)\, d\tau.$$

Now we only have to note that

$$\int_0^1 \frac{(1-\tau)^{n-1}}{(n-1)!} f^{(n)}(0)\, d\tau = \frac{1}{n!} f^{(n)}(0)$$

to obtain formula (E.1). The case that $x < 0$ can be done in an analogous way. The last statement of the theorem follows from the continuity of $f^{(n)}$, and — for example — an application of the dominated convergence theorem A.56. □

Bibliography

[1] R. A. Adams, *Sobolev Spaces*, Academic Press, New York, 1975.

[2] R. B. Ash, *Measure, Integration, and Functional Analysis*, Academic Press, New York, 1972.

[3] ______, *Real Analysis and Probability*, Academic Press, New York and London, 1972.

[4] H. Bauer, *Maß- und Integrationstheorie*, de Gruyter, Berlin, New York, 1990.

[5] ______, *Wahrscheinlichkeitstheorie*, de Gruyter, Berlin, New York, 2002.

[6] E. Behrends, *Maß– und Integrationstheorie*, Springer, Berlin, Heidelberg, New York, 1987.

[7] P. Billingsley, *Convergence of Probability Measures*, Wiley, New York, London, 1968.

[8] S. Bochner, *Lectures on Fourier Integrals*, Princeton University Press, Princeton, New Jersey, 1959.

[9] V. I. Bogachev, *Measure Theory*, vol. I, Springer, Berlin, Heidelberg, New York, 2007.

[10] L. Breiman, *Probability*, Addison-Wesley, Reading, Massachusetts, 1968.

[11] K. L. Chung, *A Course in Probability Theory*, third ed., Academic Press, New York, London, 2001.

[12] H. Dinges and H. Rost, *Prinzipien der Stochastik*, Vieweg & Teubner, Stuttgart, 1982.

[13] J. L. Doob, *Measure Theory*, Springer, Berlin, Heidelberg, New York, 1994.

[14] H. Dym and H. P. McKean, *Fourier Series and Integrals*, Academic Press, New York, 1972.

[15] W. Feller, *An Introduction to Probability Theory and Its Applications*, 3rd ed., vol. 1, Wiley, New York, London, Sydney, 1968.

[16] ______, *An Introduction to Probability Theory and Its Applications*, 2nd ed., vol. 2, Wiley, New York, London, Sydney, 1971.

[17] P. Halmos, *Measure Theory*, Van Nostrand, Princeton, New Jersey, 1950.

[18] E. Henze, *Einführung in die Maßtheorie*, Bibliographisches Institut, Mannheim, 1971.

[19] D. Hilbert, *Mathematische Probleme*, Nachrichten der Königlichen Gesellschaft der Wissenschaften zu Göttingen (1900), 253–297.

[20] T. Kato, *Perturbation Theory for Linear Operators*, second ed., Springer-Verlag, Berlin, Heidelberg, New York, 1976.

[21] Y. Katznelson, *An Introduction to Harmonic Analysis*, Dover, New York, 1968.

[22] A. Kolmogoroff, *Grundbegriffe der Wahrscheinlichkeitsrechnung*, Springer, Berlin, Heidelberg, New York, 1933, reprinted 1973.

[23] A. N. Kolmogorov, *Foundations of the Theory of Probability*, Springer, Berlin, Heidelberg, New York, 1956, (English translation by N. Morrison).

[24] U. Krengel, *On the contribution of Georg Bohlmann to probability theory*, Journ@l Electronique d'Histoire des Probabilités et de la Statistique/ Electronic Journ@l for History of Probability and Statistics **7** (2011), 1–14.

[25] M. Loève, *Probability Theory*, 4th ed., vol. I, Springer, New York, Heidelberg, Berlin, 1978.

[26] L. H. Loomis, *An Introduction to Abstract Harmonic Analysis*, Van Nostrand, London, 1953.

[27] J. Potthoff, *Probability II*, 2013, Lecture Notes, Mannheim University, http://math5.math.uni-mannheim.de/.

[28] ______, *Einführung in die Wahrscheinlichkeitstheorie*, 2014, Lecture Notes, Mannheim University, http://math5.math.uni-mannheim.de/.

[29] M. Reed and B. Simon, *Methods of Modern Mathematical Physics*, vol. 1, *Functional Analysis*, Academic Press, New York, London, 1972.

[30] ______, *Methods of Modern Mathematical Physics*, vol. 2, *Fourier Analysis, Self-Adjointness*, Academic Press, New York, London, 1975.

[31] F. Riesz and B. Sz.-Nagy, *Functional Analysis*, Dover, New York, 1990.

[32] W. Rudin, *Real and Complex Analysis*, McGraw-Hill, London, New York, 1970.

[33] G. Shafer and V. Vovk, *The sources of Kolmogorov's Grundbegriffe*, arXiv:math/0606533v1, 2006.

[34] D. Williams, *Probability with Martingales*, Cambridge University Press, Cambridge, 1991.

Index

Index of Notations

$\mathcal{A}(\mathcal{M})$	algebra generated by $\mathcal{M}$
$\mathcal{A}_1 \otimes \mathcal{A}_2$	product σ–algebra
$\mathcal{B}(\mathbb{C})$	Borel σ–algebra over $\mathbb{C}$
$\mathcal{B}(\mathbb{R}^n)$	Borel σ–algebra over $\mathbb{R}^n$
$\mathcal{B}(A)$	Borel σ–algebra over $A \subset \mathbb{R}^n$
$\mathcal{B}(p)$	Bernoulli law with parameters $p \in [0, 1]$, $x_1 = 1$, $x_2 = 0$
$\mathcal{B}(p, n)$	binomial law with parameters $p \in [0, 1]$, $n \in \mathbb{N}$
$\mathcal{E}(\lambda)$	exponential law with parameter λ
$\mathcal{F}$	a generic family of distribution functions
$\mathcal{I}^n$	semiring of right semiclosed intervals in $\mathbb{R}^n$
$\mathcal{K}(\mathbb{R})$	set of all characteristic functions on $\mathbb{R}$
$\mathcal{L}(\Omega, \mathcal{A}, P)$	set of positive or integrable random variables
$\mathcal{L}^1(\Omega, \mathcal{A}, \mu)$	μ–integrable functions on Ω
$\mathcal{L}^p(\Omega, \mathcal{A}, \mu)$	p–fold μ–integrable functions on Ω
$\mathcal{L}^p(\Omega, \mathcal{A}, P)$	space of p–fold integrable random variables
$\mathcal{M}$	a generic family of probability measures
$\mathcal{M}(\Omega, \mathcal{A})$	set of all complex measures on $(\Omega, \mathcal{A})$
$\mathcal{M}_+(\Omega, \mathcal{A})$	positive, $\mathcal{A}$–measurable functions on Ω
$\mathcal{N}(\mu, \sigma^2)$	normal law with parameters μ, σ^2
$\mathrm{Cov}(X, Y)$	covariance of X and Y
$\mathcal{P}(\lambda)$	Poisson law with parameter $\lambda > 0$

$\mathcal{P}(\Omega)$	powerset of Ω
$\mathcal{R}(\mathcal{M})$	ring generated by $\mathcal{M}$
$\mathcal{U}([a,b])$	uniform law on $[a,b]$
$\mathcal{W}(\mathbb{R})$	set of all probability measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$
$\mathcal{W}_n(\mathbb{R})$	set of probability measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ with finite n–th moment
$\mathcal{X}$	a generic family of random variables
$\xrightarrow{\text{a.s.}}$	almost sure convergence
$\xrightarrow{\text{P}}$	convergence in probability
$\sigma(\mathcal{M})$	σ–algebra generated by $\mathcal{M}$
$\sigma(T_i,\ i \in I)$	σ–algebra generated by a family $(T_i,\ i \in I)$ of mappings
$\hat{\mu}$	characteristic function of the measure μ
$\hat{f}$	Fourier transform of f
$\langle \mu, f \rangle$	dual pairing of μ and f
$\mu_1 * \mu_2$	convolution of μ_1 and μ_2
$\mu_1 \otimes \mu_2$	product measure
$\mu_n \xrightarrow{\text{w}} \mu$	weak convergence of μ_n to μ
$\rho(X,Y)$	correlation of X and Y
$C_0(S)$	continuous functions on S vanishing at infinity
$C_b(S)$	bounded, continuous functions on S
$C_c(S)$	continuous functions on S with compact support
$E(X)$	expectation value of X
$L_n^X(\varepsilon)$	Lindeberg expression
$M_n(\mu)$	n–th moment of μ
$P(\mathbb{T})$	space of trigonometric polynomials on the torus
$X \sim \mathcal{L}$	X has the law $\mathcal{L}$
$X_n \xrightarrow{\text{L}} X$	convergence in law of X_n to X