De Gruyter Studies in Mathematics 38

Vassili N. Kolokoltsov

Markov Processes, Semigroups and Generators

De Gruyter

Mathematical Subject Classification 2010: Primary: 60-01, 60-02; Secondary: 60G51, 60G52, 60H10, 60H30, 60G50, 60J25, 60K35, 60F17, 81Q05, 81S40.

ISBN 978-3-11-025010-7
e-ISBN 978-3-11-025011-4
ISSN 0179-0986

Library of Congress Cataloging-in-Publication Data

Kolokol'tsov, V. N. (Vasilii Nikitich)
Markov processes, semigroups, and generators / by Vassili N. Kolokoltsov.
p. cm. – (De Gruyter studies in mathematics ; 38)
Includes bibliographical references and index.
ISBN 978-3-11-025010-7 (alk. paper)
1. Markov processes. 2. Semigroups. 3. Group theory – Generators. I. Title.
QA274.7.K65 2011
519.2'33–dc22
2010050783

Bibliographic information published by the Deutsche Nationalbibliothek

The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at http://dnb.d-nb.de.

Typesetting: Da-TeX Gerd Blumenstein, Leipzig, www.da-tex.de
Printing and binding: Hubert & Co. GmbH & Co. KG, Göttingen
∞ Printed on acid-free paper

Printed in Germany

www.degruyter.com

To the memory of my parents,
Nikita and Svetlana Kolokoltsov,
who helped me much to become a scientist

Preface

Markov processes represent a universal model for a large variety of real life random evolutions. The wide flow of new ideas, tools, methods and applications constantly pours into the ever-growing stream of research on Markov processes that rapidly spreads over new fields of natural and social sciences, creating new streamlined logical paths to its turbulent boundary. Even if a given process is not Markov, it can be often inserted into a larger Markov one (Markovianization procedure) by including the key historic parameters into the state space.

Markov processes are described probabilistically by the distributions on their trajectories (often specified by stochastic differential equations) and analytically by the Markov semigroups that specify the evolution of averages and arise from the solutions to a certain class of integro-differential (or pseudo-differential) equations, which is distinguished by the preservation of the positivity property (probabilities are positive). Thus the whole development stands on two legs: stochastic analysis (with tools such as the martingale problem, stochastic differential equations, convergence of measures on Skorohod spaces), and functional analysis (weighted Sobolev spaces, pseudo-differential operators, operator semigroups, methods of Hilbert and Fock spaces, Fourier analysis).

The aim of the monograph is to give a concise (but systematic and self-contained) exposition of the essentials of Markov processes (highly nontrivial, but conceptually excitingly rich and beautiful), together with recent achievements in their constructions and analysis, stressing specially the interplay between probabilistic and analytic tools. The main point is in the construction and analysis of Markov processes from the 'physical picture' – a formal pre-generator that specifies the corresponding evolutionary equation (here the analysis really meets probability) paying particular attention to the universal models (analytically – general positivity-preserving evolutions), which go above standard cases (e.g. diffusions and jump-type processes).

The introductory Part I is an enlarged version of the one-semester course on Brownian motion and its applications given by the author to the final year mathematics and statistics students of Warwick University. In this course, Browninan motion was studied not only as the simplest continuous random evolution, but as a basic continuous component of complex processes with jumps. Part I contains mostly well-known material, though written and organized with a point of view that anticipates further developments. In some places it provides more general formulations than usual (as with the duality theorem in Section 1.9 or with the Holtzmark distributions in Section 1.5) and new examples (as in Section 2.11).

Part II is based mainly on the author's research. To facilitate the exposition, each chapter of Part II is composed in such a way that it can be read almost independently of others, and it ends with a section containing comments on bibliography and related topics. The main results concern:

(i) various constructions and basic continuity properties of Markov processes, including processes stopped or killed at the boundary (as well as related boundary points classification and sensitivity analysis),

(ii) in particular, heat kernel estimates for stable-like processes,

(iii) limiting processes for position-dependent continuous time random walks (obtained by a random time change from the Markov processes) and related fractional (in time) dynamics,

(iv) the rigorous Feynman path-integral representation for the solutions of the basic equations of quantum mechanics, via jump-type Markov processes.

We also touch upon the theory of stochastic monotonicity, stochastic scattering, stochastic quasi-classical (also called small diffusion) asymptotics, and stochastic control. An important development of the methods discussed here is given by the theory of nonlinear Markov processes (including processes on manifolds) presented in the author's monograph [196]. They are briefly introduced at the end of Chapter 5.

It is worth pointing out the directions of research closely related to the main topic of this book, but not touched here. These are Dirichlet forms, which can be used for constructing Markov processes instead of generators, Mallivin calculus, which is a powerful tool for proving various regularity properties for transition probabilities, log-Sobolev inequalities, designed to systematically analyze the behavior of the processes for large times, and processes on manifolds. There exists an extensive literature on each of these subjects.

The book is meant to become a textbook and a monograph simultaneously, taking more features of the latter as the exposition advances. I include some exercises, their weight being much more sound at the beginning. The exercises are supplied with detailed hints and are meant to be doable with the tools discussed in the book. The exposition is reasonably self-contained, with pre-requisites being just the standard math culture (basic analysis and linear algebra, metric spaces, Hilbert and Banach spaces, Lebesgue integration, elementary probability). We shall start slowly from the prerequisites in probability and stochastic processes, omitting proofs if they are well presented in university text books and not very instructive for our purposes, but stressing ideas and technique that are specially relevant. Streamlined logical paths are followed to the main ideas and tools for the most important models, by-passing wherever possible heavy technicalities (say, by working with Lévy processes instead of general semi-martingales, or with left-continuous processes instead of predictable ones).

A methodological aspect of the presentation consists in often showing various perspectives for key topics and giving several proofs of main results. For example, we begin the analysis of random processes with several constructions of the Brownian motion: 1) via binary subdivisions anticipating the later given Itô approach to constructing Markov evolutions, 2) via tightness of random-walk approximations, anticipating the later given LLN for non-homogeneous random walks, 3) via Hilbert-space methods leading to Wiener chaos that is crucial for various developments, for instance for Malliavin calculus and Feynman path integration, 4) via the Kolmogorov continuity theorem. Similarly, we give two constructions of the Poisson process, several constructions of basic stochastic integrals, several approaches to proving functional CLTs (via tightness of random walks, Skorohod embedding and the analysis of generators). Further on various probabilistic and analytic constructions of the main classes of Markov semigroups are given. Every effort was made to introduce all basic notions in the most clear and transparent way, supplying intuition, developing examples and stressing details and pitfalls that are crucial to grasp its full meaning in the general context of stochastic analysis. Whenever possible, we opt for results with the simplest meaningful formulation and quick direct proof.

As teaching and learning material, the book can be used on various levels and with different objectives. For example, short courses on an introduction to Brownian motion, Lévy and Markov processes, or on probabilistic methods for PDE, can be based on Chapters 2, 3 and 4 respectively, with chosen topics from other parts. Let us stress only that the celebrated Itô's lemma is not included in the monograph (it actually became a common place in the textbooks). More advanced courses with various flavors can be built on part II, devoted, say, to continuous-time random walks, to probabilistic methods for boundary value problems or for the Feynman path integral.

Finally, let me express my gratitude to Professor Niels Jacob from the University of Wales, Swansea, an Editor of the De Gryuter Studies in Mathematics Series, who encouraged me to write this book. I am also most grateful to Professor Nick Bingham from Imperial College, London, for reading the manuscript carefully and making lots of comments that helped me to improve the overall quality immensely.

Coventry, November 2010 V. N. Kolokoltsov

Notations

Numbers and sets

- $a \vee b = \max(a, b)$, $a \wedge b = \min(a, b)$
- $\mathbb{N}$ and $\mathbb{Z}$ are the sets of natural and integer numbers
- $\mathbb{C}^d$ and $\mathbb{R}^d$ are the complex and real d-dimensional spaces, $|x|$ or $\|x\|$ for a vector $x \in \mathbb{R}^d$ denotes its Euclidean norm, (x, y) or xy denotes the scalar product of the vectors $x, y \in \mathbb{R}^d$
- $\operatorname{Re} a$ and $\operatorname{Im} a$ are the real and imaginary part of a complex number a
- $B_r(x)$ (resp. B_r) is the ball of radius r centered at x (resp. at the origin)
- S^d is the d-dimensional unit sphere in $\mathbb{R}^{d+1}$
- $\mathbb{R}_+$ (resp. $\bar{\mathbb{R}}_+$) is the the set of positive (resp. non-negative) numbers
- $\bar{\Omega}$ and $\partial\Omega$ are the closure and the boundary respectively of the subset Ω in a metric space
- Ω^S is the set of all mappings $S \to \Omega$
- $[x]$ is the integer part of a real number x

Functions

- $B(S)$ (resp. $C(S)$ or $C_b(S)$) for a complete metric space (S, ρ) (usually $S = \mathbb{R}^d$, $\rho(x, y) = \|x - y\|$) is the Banach space of bounded Borel measurable (resp. bounded continuous) functions on S equipped with the sup-norm

$$\|f\| = \sup_{x \in S} |f(x)|$$

- $C_c(S) \subset C(S)$ consists of functions with a compact support
- $C_{\text{Lip}}(S) \subset C(S)$ consists of Lipschitz continuous functions f, i.e. $|f(x) - f(y)| \leq \kappa\rho(x, y)$ with a constant κ; $C_{\text{Lip}}(S)$ is a Banach space under the norm

$$\|f\|_{\text{Lip}} = \sup_x |f(x)| + \sup_{x \neq y} |f(x) - f(y)|/|x - y|$$

- $C_\infty(S) \subset C(S)$ consists of f such that $\lim_{x\to\infty} f(x) = 0$, i.e. $\forall \epsilon\ \exists$ a compact set $K : \sup_{x \notin K} |f(x)| < \epsilon$ (it is a closed subspace of $C(S)$ if S is locally compact)

- $C^k(\mathbb{R}^d)$ or $C^k_b(\mathbb{R}^d)$ (sometimes shortly C^k) is the Banach space of k times continuously differentiable functions with bounded derivatives on $\mathbb{R}^d$ with the norm being the sum of the sup-norms of the function itself and all its partial derivative up to and including order k
- $C^k_{\text{Lip}}(\mathbb{R}^d)$ is the subspace of $C^k(\mathbb{R}^d)$ with all derivative up to and including order k being Lipschitz continuous; it is a Banach space equipped with the norm

$$\|f\|_{C^k_{\text{Lip}}} = \|f\|_{C^{k-1}} + \|f^{(k)}\|_{\text{Lip}}$$

- $C^k_c(\mathbb{R}^d) = C_c(\mathbb{R}^d) \cap C^k(\mathbb{R}^d)$
- $\nabla f = (\nabla_1 f, \ldots, \nabla_d f) = (\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_d})$, $f \in C^1(\mathbb{R}^d)$
- $L^p(\Omega, \mathcal{F}, \mu)$, $p \geq 1$, is the usual Banach space of (the equivalence classes of) measurable functions f on the measure space Ω such that

$$\|f\|_p = \left(\int |f|^p(x)\mu(dx) \right)^{1/p} < \infty$$

- $L^\infty(\Omega, \mathcal{F}, \mathbf{P})$ is the Banach space of (the equivalence classes of) measurable functions f on the measure space Ω with a finite sup-norm

$$\|f\| = \text{ess}\sup_{x \in \Omega} |f(x)|$$

- $S(\mathbb{R}^d) = \{f \in C^\infty(\mathbb{R}^d) : \forall k, l \in \mathbb{N}, |x|^k \nabla^l f \in C_\infty(\mathbb{R}^d)\}$ is the Schwartz space of rapidly deceasing functions
- $(f, g) = \int f(x)g(x)\,dx$ denotes the scalar product for functions f, g on $\mathbb{R}^d$ or on a general measure space
- $\mathbf{1}_M$ is the indicator function of a set M (equals one or zero according to whether its argument is in M or otherwise)
- sgn is the sign function taking values $+1, 0, -1$ for positive, vanishing and negative values of the argument respectively
- $f = O(g)$ means $|f| \leq Cg$ for some constant C
- $f = o(g)_{n \to \infty} \iff \lim_{n \to \infty}(f/g) = 0$

Measures

- $\mathcal{M}(S)$ (resp. $\mathcal{P}(S)$) is the set of finite (positive) Borel measures (resp. probability measures) on a metric space S
- $\mathcal{M}^{\text{signed}}(S)$ defines the Banach space of finite signed Borel measures on a metric space S
- $|\nu|$ for a signed measure ν is its (positive) total variation measure
- $(f, \mu) = \int_S f(x)\mu(dx)$ for $f \in C(S)$, $\mu \in \mathcal{M}(S)$

Matrices and linear operators

- A^T is the transpose to a matrix A
- Ker(A), Sp(A), tr(A) are the kernel, spectrum and trace of the operator A
- $\|A\|_B$ is the norm of the operator A in a Banach space B
- $\|A\|_{B\to C}$ is the norm of the operator A as a mapping between Banach spaces B and C
- $C([0,t],B)$ is the Banach space of continuous functions on $[0,t]$ with values in the Banach space B equipped with the sup-norm $\|f\| = \sup_{s\in[0,t]} \|f(s)\|$

Probability

- $\mathbf{E}$ and $\mathbf{P}$ define the expectation and probability, $\mathbf{E}_x, \mathbf{P}_x$ for $x \in S$ (respectively $\mathbf{E}_\mu, \mathbf{P}_\mu$ for $\mu \in \mathcal{P}(S)$) are the expectation and probability with respect to an S-valued process started at x (respectively with the initial distribution μ)

Standard abbreviations

a.s.	almost sure
i.i.d.	independent identically distributed
l.h.s.	left-hand side
r.h.s.	right-hand side
r.v.	random variable
BM	Brownian motion
CLT	central limit theorem
CTRW	continuous time random walk
LLN	law of large numbers
ODE	ordinary differential equation
OU	Ornstein–Uhlenbeck (process)
PDE	partial differential equation
PDO	partial differential operator
ΨDE	pseudo differential equation
ΨDO	pseudo differential operator
SDE	stochastic differential equation

Contents

Preface vii

Notations xi

Standard abbreviations xiv

I Introduction to stochastic analysis **1**

1 Tools from probability and analysis 3
1.1 Essentials of measure and probability . . . 3
1.2 Characteristic functions . . . 13
1.3 Conditioning . . . 16
1.4 Infinitely divisible and stable distributions . . . 21
1.5 Stable laws as the Holtzmark distributions . . . 27
1.6 Unimodality of probability laws . . . 29
1.7 Compactness for function spaces and measures . . . 35
1.8 Fractional derivatives and pseudo-differential operators . . . 42
1.9 Propagators and semigroups . . . 48

2 Brownian motion (BM) 58
2.1 Random processes: basic notions . . . 58
2.2 Definition and basic properties of BM . . . 62
2.3 Construction via broken-line approximation . . . 66
2.4 Construction via Hilbert-space methods . . . 69
2.5 Construction via Kolmogorov's continuity . . . 71
2.6 Construction via random walks and tightness . . . 73
2.7 Simplest applications of martingales . . . 76
2.8 Skorohod embedding and the invariance principle . . . 78
2.9 More advanced Hilbert space methods: Wiener chaos and stochastic integral . . . 81
2.10 Fock spaces, Hermite polynomials and Malliavin calculus . . . 87
2.11 Stationarity: OU processes and Holtzmark fields . . . 91

3 Markov processes and martingales 94
3.1 Definition of Lévy processes . . . 94
3.2 Poisson processes and integrals . . . 96

3.3 Construction of Lévy processes 103
3.4 Subordinators 108
3.5 Markov processes, semigroups and propagators 110
3.6 Feller processes and conditionally positive operators 115
3.7 Diffusions and jump-type Markov processes 125
3.8 Markov processes on quotient spaces and reflections 130
3.9 Martingales 132
3.10 Stopping times and optional sampling 138
3.11 Strong Markov property; diffusions as Feller processes with continuous paths 143
3.12 Reflection principle and passage times 147

4 SDE, ΨDE and martingale problems 152
4.1 Markov semigroups and evolution equations 152
4.2 The Dirichlet problem for diffusion operators 158
4.3 The stationary Feynman–Kac formula 162
4.4 Diffusions with variable drift, Ornstein–Uhlenbeck processes 165
4.5 Stochastic integrals and SDE based on Lévy processes 167
4.6 Markov property and regularity of solutions 172
4.7 Stochastic integrals and quadratic variation for square-integrable martingales 178
4.8 Convergence of processes and semigroups 187
4.9 Weak convergence of martingales 193
4.10 Martingale problems and Markov processes 195
4.11 Stopping and localization 199

II Markov processes and beyond **203**

5 Processes in Euclidean spaces 204
5.1 Direct analysis of regularity and well-posedness 205
5.2 Introduction to sensitivity analysis 213
5.3 The Lie–Trotter type limits and T-products 213
5.4 Martingale problems for Lévy type generators: existence 221
5.5 Martingale problems for Lévy type generators: moments 226
5.6 Martingale problems for Lévy type generators: unbounded coefficients 228
5.7 Decomposable generators 231
5.8 SDEs driven by nonlinear Lévy noise 240
5.9 Stochastic monotonicity and duality 250
5.10 Stochastic scattering 255
5.11 Nonlinear Markov chains, interacting particles and deterministic processes 257

5.12 Comments . . . 262

6 Processes in domains with a boundary 270
6.1 Stopped processes and boundary points . . . 270
6.2 Dirichlet problem and mixed initial-boundary problem . . . 274
6.3 The method of Lyapunov functions . . . 280
6.4 Local criteria for boundary points . . . 282
6.5 Decomposable generators in $\mathbb{R}^d_+$ . . . 286
6.6 Gluing boundary . . . 290
6.7 Processes on the half-line . . . 292
6.8 Generators of reflected processes . . . 293
6.9 Application to interacting particles: stochastic LLN . . . 295
6.10 Application to evolutionary games . . . 304
6.11 Application to finances: barrier options, credit derivatives, etc. . . . 307
6.12 Comments . . . 308

7 Heat kernels for stable-like processes 310
7.1 One-dimensional stable laws: asymptotic expansions . . . 310
7.2 Stable laws: asymptotic expansions and identities . . . 314
7.3 Stable laws: bounds . . . 319
7.4 Stable laws: auxiliary convolution estimates . . . 323
7.5 Stable-like processes: heat kernel estimates . . . 328
7.6 Stable-like processes: Feller property . . . 335
7.7 Application to sample-path properties . . . 336
7.8 Application to stochastic control . . . 340
7.9 Application to Langevin equations driven by a stable noise . . . 345
7.10 Comments . . . 348

8 CTRW and fractional dynamics 351
8.1 Convergence of Markov semigroups and processes . . . 351
8.2 Diffusive approximations for random walks and CLT . . . 354
8.3 Stable-like limits for position-dependent random walks . . . 355
8.4 Subordination by hitting times and generalized fractional evolutions . 361
8.5 Limit theorems for position dependent CTRW . . . 369
8.6 Comments . . . 371

9 Complex Markov chains and Feynman integral 373
9.1 Infinitely-divisible complex distributions and complex Markov chains 373
9.2 Path integral and perturbation theory . . . 380
9.3 Extensions . . . 385
9.4 Regularization of the Schrödinger equation by complex time or mass, or continuous observation . . . 390

9.5 Singular and growing potentials, magnetic fields and curvilinear state spaces . . . 393
9.6 Fock-space representation . . . 398
9.7 Comments . . . 400

Bibliography 403

Index 425

Part I

Introduction to stochastic analysis

Chapter 1
Tools from probability and analysis

This chapter is meant to supply the preliminary material needed for reading the book. Though we do give most of the proofs (sometimes sketchy), some fundamental facts are only formulated. The criterion used for the omission of the proofs was two-fold. On the one hand, these proofs are not deeply connected with (nor very instructive for the understanding of) the main body of this text, and are non-trivial, so that their proper exposition would be time and space consuming; and on the other hand, they are quite standard by now and are widely represented in university textbooks. To set the ground for probability, we recall the notion of a measure space, but we do assume readers to be acquainted with the definition and basic properties of integrals with respect to an abstract measure including dominated and monotone convergence theorems. In the next three sections we collect the basic facts from standard probability texts, see e.g. Applebaum [19], Jacod and Protter [146], Shiryaev [293] and Kallenberg [154], so that references are not given to each formulated result separately. Afterwards we introduce infinitely divisible and stable distributions. Then we recall the basic topologies used routinely in stochastic analysis. And finally we introduce the analytic tools (fractional derivatives, pseudo-differential operators and semigroups) used in what follows. Readers with a sound background in probability and/or analysis may wish to skip some or all sections of this introductory chapter.

1.1 Essentials of measure and probability

A collection $\mathcal{F}$ of subsets of a given set S is called a *σ-algebra* if

(i) $S \in \mathcal{F}$;

(ii) $A \in \mathcal{F} \Rightarrow S \setminus A \in \mathcal{F}$;

(iii) (σ-additivity) $\bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$ whenever $A_n \in \mathcal{F}$ for any $n \in \mathbb{N}$.

The pair $(S, \mathcal{F})$ is called a *measurable space*.

A *measure* on $(S, \mathcal{F})$ is a mapping $\mu : \mathcal{F} \mapsto [0, \infty]$ such that $\mu(\emptyset) = 0$ and *σ-additivity* holds:

$$\mu\Big(\bigcup_{n=1}^{\infty} A_n\Big) = \sum_{n=1}^{\infty} \mu(A_n)$$

for any sequence A_n of mutually disjoint sets in $\mathcal{F}$. The triple $(S, \mathcal{F}, \mu)$ is called a *measure space*. A measure μ is called *finite* if its *total mass* $\mu(S)$ is finite, *σ-finite* if there exists a sequence A_n, $n \in \mathbb{N}$, of subsets of $\mathcal{F}$ such that $S = \bigcup_{n=1}^{\infty} A_n$ and $\mu(A_n) < \infty$ for all n.

A measure space $(\Omega, \mathcal{F}, \mu)$ is called a *probability space* whenever $\mu(\Omega) = 1$. In this case μ is called a *probability measure* and the subsets from $\mathcal{F}$ are called *events*.

An extension of the notion of a measure that does not assume positivity is sometimes useful as well. Namely a *signed measure* (respectively a *complex measure*) of *finite variation* can be defined as a set function ϕ on a measurable space $(S, \mathcal{F})$ that is given by the integral

$$\phi(A) = \int_S f(x)\mu(dx), \quad A \in \mathcal{F}, \tag{1.1}$$

where μ is a (positive) finite measure on $(S, \mathcal{F})$ and f is a real (respectively complex-valued) function on S integrable with respect to μ.[1] The *total variation norm* of ϕ is defined as

$$\|\phi\| = \int |f(x)|\mu(dx).$$

The set of signed (respectively complex) measures of finite variation on $(S, \mathcal{F})$ is easily seen to be a real (respectively complex) Banach space when equipped with this norm. The *total variation measure* of ϕ is defined as the measure

$$|\phi|(dx) = |f(x)|\mu(dx),$$

so that

$$\phi(dx) = \sigma(x)|\phi|(dx)$$

with σ taking only three values $0, 1, -1$. As for usual measures, the following extension is sometimes useful. A set function ϕ is called a σ-*finite* signed (or complex) measure if it has a representation (1.1) with a σ-finite measure μ and a bounded measurable real (respectively complex) function f.

For a metric space S, e.g. a subset of $\mathbb{R}^d$, the smallest σ-algebra $\mathcal{B}(S)$ containing all its open subsets is called the *Borel σ-algebra* of S. Its elements are called *Borel sets* and any measure on $(S, \mathcal{B}(S))$ is called a *Borel measure*. The simplest example of a Borel measure is given by Lebesgue measure on $\mathbb{R}^d$. A Borel measure is called a *Radon measure* if it is finite on any compact set. One can also define a *signed or complex Radon measure* as a set function that becomes a signed or complex measure of finite variation when reduced to any compact set.

Throughout this book our processes will live in Euclidean spaces $\mathbb{R}^d$. However, the distribution of a $\mathbb{R}^d$-valued process is a distribution on a certain space of trajectories of such a process, and the latter space is often specified as a rather nontrivial infinite-dimensional metric space (Skorohod space). Hence the necessity to work with measures on general metric spaces, even when analyzing finite-dimensional processes.

[1] Alternatively signed measures can be defined axiomatically, in which case this representation follows from the so-called Hahn decomposition theorem.

For a collection Γ of the subsets of a set Ω the σ-algebra $\sigma(\Gamma)$ *generated by* Γ is the minimal σ-algebra containing all sets from Γ.

An important method of constructing measures is via the products. Namely, for a finite or a countable family of measure spaces $(S_i, \mathcal{F}_i, \mu_i)$, $i = 1, 2, \dots$, the *product measure space* $(S, \mathcal{F}, \mu)$ is defined, where $S = S_1 \times S_2 \times \cdots$, $\mathcal{F} = \mathcal{F}_1 \otimes \mathcal{F}_2 \otimes \cdots$ – the σ-algebra generated by the sets $A_1 \times \cdots \times A_n$, $A_i \in \mathcal{F}_i$, $n \in \mathbb{N}$, and $\mu = \mu_1 \times \mu_2 \times \cdots$ is the *product measure* uniquely specified by the prescription $\mu(A_1 \times \cdots \times A_n) = \mu_1(A_1) \cdots \mu_n(A_n)$.

For a measure space $(S, \mathcal{F}, \mu)$ a subset of S is called *negligible* or a *null set* if it is a subset of a $N \in \mathcal{F}$ with $\mu(N) = 0$. The σ-algebra $\bar{\mathcal{F}}$ of the subsets of S of the form $A \cup B$, with $A \subset \mathcal{F}$, B negligible and the measure $\bar{\mu}$ on it defined on these sets as $\bar{\mu}(A \cup B) = \mu(A)$, are called respectively the *completion* of $\mathcal{F}$ and μ (with respect to μ). In particular, for $S \subset \mathbb{R}^d$ the completion of $\mathcal{B}(S)$ with respect to Lebesgue measure is called the σ-algebra of *Lebesgue measurable sets* in S.

For a probability space $(\Omega, \mathcal{F}, \mu)$ one says that some property depending on $\omega \in \Omega$ holds *almost surely* (briefly *a.s.*) or *with probability* 1 if there exists a negligible set $N \in \mathcal{F}$ such that this property holds for all $\omega \in \Omega \setminus N$.

A handy tool of probability theory is given by the following famous result called the *Borel–Cantelli lemma*.

Theorem 1.1.1. *If a sequence of events A_n, $n \in \mathbb{N}$, on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ is such that $\sum_n \mathbf{P}(A_n) < \infty$, then a.s. only a finite number of A_n can occur.*

Proof. Let $B = \{\omega \in \Omega : \text{infinite number } A_n \text{ occur}\}$. Then

$$B = \bigcap_n \Big(\bigcup_{k \ge n} A_k \Big)$$

and

$$\mathbf{P}(B) \le \mathbf{P}\Big(\bigcup_{k \ge n} A_k \Big) \le \sum_{k \ge n} \mathbf{P}(A_k) \to 0,$$

as $n \to \infty$. Hence $\mathbf{P}(B) = 0$. □

If $(S_i, \mathcal{F}_i)$, $i = 1, 2$, are measurable spaces, a mapping $f : S_1 \to S_2$ is called $(\mathcal{F}_1, \mathcal{F}_2)$-*measurable* if $f^{-1}(A) \in \mathcal{F}_1$ whenever $A \in \mathcal{F}_2$. Two measurable spaces are called *Borel isomorphic*, or just *isomorphic*, if there exists a bijection $f : S \to T$ such that both f and f^{-1} are measurable. A measurable space S is called a *Borel space* if it is isomorphic to a Borel subset of $[0, 1]$. A deep result of measure theory states that a complete metric space is a Borel space. This result is very convenient, as it allows one to establish certain general facts by proving them only for the real line (see below the randomization lemma). In our book we shall use only Borel spaces and measures.

If S_1, S_2 are metric spaces equipped with their Borel σ-algebras, such a mapping is said to be *Borel measurable* or briefly *Borel*. Speaking about measurable mapping with values in $\mathbb{R}^d$ one usually means that $\mathbb{R}^d$ is equipped with its Borel σ-algebra.

For a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ the measurable mappings $X : \Omega \to \mathbb{R}^d$ are called *random variables* (briefly r.v.), or sometimes random vectors in case $d > 1$. More generally, for a metric space S the Borel measurable mappings $\Omega \to S$ are called *S-valued random variables* or *random elements on S*. The σ-algebra $\sigma(X)$ *generated by a r.v. X* is the smallest σ-algebra containing the sets $\{X \subset B\}$ for all Borel sets B.

The *law* (or the *distribution*) of a random variable is the Borel probability measure p_X on S defined as $p_X = \mathbf{P} \circ X^{-1}$. In other words,

$$p_X(A) = \mathbf{P}(X^{-1}(A)) = \mathbf{P}(\omega \in \Omega : X(\omega) \in A) = \mathbf{P}(X \in A).$$

For example, if X takes only finite number of values, then the law p_X is a sum of Dirac δ-measures.

Clearly, if μ is a probability measure on $\mathbb{R}^d$, then the identical mapping in $\mathbb{R}^d$ defines a $\mathbb{R}^d$-valued random vector with the law μ defined on the probability space $(\mathbb{R}^d, \mathcal{B}, \mu)$. It turns out that for a family of laws depending measurably on a parameter one can specify a family of random variables defined on a single probability space and depending measurably on this parameter. This is shown in the following *randomization lemma*:

Lemma 1.1.1. *Let $\mu(x, dz)$ be a family of probability measures on a Borel space Z depending measurably on a parameter x from another measurable space X (such a family is called a probability kernel from X to Z). Then there exists a measurable function $f : X \times [0, 1] \to Z$ such that if θ is uniformly distributed on* $[0, 1]$*, then $f(X, \theta)$ has distribution $\mu(x, .)$ for every $x \in X$.*

Proof. Since Z is a Borel space, it is sufficient to prove the statement for $Z = [0, 1]$. In this case f can be defined by the explicit formula, the *probability integral transformation*, that represents a standard method (widely used in practical simulations), for obtaining a random variable from a given one-dimensional distribution:

$$f(s, t) = \sup\{x \in [0, 1] : \mu(s, [0, x]) < t\}.$$

Clearly this mapping depends measurably on s. Moreover, the events $\{f(s, t) \leq y\}$ and $\{t \leq \mu(s, [0, y])\}$ coincide. Hence, for a uniform θ

$$\mathbf{P}(f(s, \theta) \leq x) = \mathbf{P}(t \leq \mu(s, [0, y])) = \mu(s, [0, x]). \qquad \square$$

Two r.v. X and Y are called *identically distributed* if they have the same probability law. For a real (i.e. one-dimensional) r.v. X its *distribution function* is defined by

$F_X(x) = p_X((-\infty, x])$. A real r.v. X has a *continuous distribution* with a *probability density function* f if $p_X(A) = \int_A f(x)dx$ for all Borel sets A.

For an $\mathbb{R}^d$-valued r.v. X on a probability space $(\Omega, \mathcal{F}, \mu)$ and a Borel measurable function $f : \mathbb{R}^d \mapsto \mathbb{R}^m$ the *expectation* $\mathbf{E}$ of $f(X)$ is defined as

$$\mathbf{E}f(X) = \mathbf{E}(f(X)) = \int_\Omega f(X(\omega))\mu(d\omega) = \int_{\mathbb{R}^d} f(x)p_X(dx). \tag{1.2}$$

X is called *integrable* if $\mathbf{E}(|X|) < \infty$.

Exercise 1.1.1. Convince yourself that the two integral expressions in (1.2) really coincide. Hint: first choose f to be an indicator, then use linearity and approximation.

For two $\mathbb{R}^d$-valued r.v. $X = (X_1, \dots, X_d)$ and $Y = (Y_1, \dots, Y_d)$ the $d \times d$-matrix with the entries $\mathbf{E}[(X_i - \mathbf{E}(X_i))(Y_j - \mathbf{E}(Y_j))]$ is called the *covariance* of X and Y and is denoted $\mathrm{Cov}(X, Y)$. In case $d = 1$ and $X = Y$ the number $\mathrm{Cov}(X, Y)$ is called the *variance* of X and is denoted by $\mathrm{Var}(X)$ and sometimes also by σ_X^2. Expectation and variance supply two basic numeric characteristics of a random variable.

The random variables X and Y are called *uncorrelated* whenever $\mathrm{Cov}(X, Y) = 0$. Random variables $X_1, \dots, X_n$ are called *independent* whenever

$$\mathbf{P}(X_1 \in A_1, X_2 \in A_2, \dots, X_n \in A_n) = \mathbf{P}(X_1 \in A_1)\mathbf{P}(X_2 \in A_2) \cdots \mathbf{P}(X_n \in A_n)$$

for all Borel A_j. Clearly in this case $\mathrm{Cov}(X_i, X_j) = 0$ for all $i \neq j$ (i.e. independent variables are uncorrelated) and

$$\mathrm{Var}(X_1 + \cdots + X_n) = \mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n). \tag{1.3}$$

As an easy consequence of the definition of the expectation constitute the following inequalities whose importance to the probability analysis is difficult to overestimate.

Theorem 1.1.2. *Markov's inequality: If X is a non-negative random variable, then for any $\epsilon > 0$*

$$\mathbf{P}(X \geq \epsilon) \leq \frac{\mathbf{E}X}{\epsilon}.$$

Chebyshev's inequality: For any $\epsilon > 0$ and a random variable Y

$$\mathbf{P}(|Y - \mathbf{E}Y| > \epsilon) \leq \frac{\mathrm{Var}(Y)}{\epsilon^2}.$$

Jensen's inequality: If g is a convex (respectively concave) function, then

$$g(\mathbf{E}(X)) \leq \mathbf{E}(g(X))$$

(respectively vice versa) whenever X and $g(X)$ are both integrable.

Proof. Evident inequalities

$$\mathbf{E}X \geq \mathbf{E}(X\mathbf{1}_{X\geq\epsilon}) \geq \epsilon\mathbf{E}\mathbf{1}_{X\geq\epsilon} = \epsilon\mathbf{P}(X \geq \epsilon)$$

imply Markov's one. Applying Markov's inequality with $X = |Y - \mathbf{E}Y|^2$ yields Chebyshev's one. Finally, if g is convex, then for any x_0 there exists a $\lambda(x_0)$ such that $g(x) \geq g(x_0) + (x - x_0)\lambda(x_0)$ for all x. Choosing $x_0 = \mathbf{E}X$ and $x = X$ yields

$$g(X) \geq g(\mathbf{E}X) + (X - \mathbf{E}X)\lambda(\mathbf{E}X).$$

Passing to the expectations leads to Jensen's inequality. Concave g are analyzed similarly. □

Exercise 1.1.2. Let X, Y be a random variable and a d-dimensional random vector respectively on a probability space. Show that for a continuous $g : \mathbb{R}^d \to \mathbb{R}$

$$\mathbf{E}(Xg(Y)) = \int_{\mathbb{R}^d} g(y)\nu(dy),$$

where ν is the signed measure $\nu(B) = \mathbf{E}(X\mathbf{1}_B(Y))$. Hint: start with indicator functions g.

A more complicated inequality that we are going to mention here is the following *Kolmogorov's inequality* that states that for the sums $S_m = \xi_1 + \xi_2 + \cdots + \xi_m$ of independent zero mean random variables $\xi_1, \xi_2, \ldots$ one has

$$\mathbf{P}(\max_{1\leq m\leq n} |S_m| > \epsilon) \leq \frac{\mathbf{E}|S_n|^2}{\epsilon^2}. \tag{1.4}$$

We shall not prove it here, but we shall establish later on its far-reaching extension: the Doob maximum inequality (notice only that both proofs, as well as other modification like Ottaviani's maximal inequality from Theorem 2.6.2, are based on the same idea of stopping at a point where the maximum is achieved). Doob's maximum inequality implies directly the following more general form of Kolmogorov's inequality (under the same assumptions as in (1.4)):

$$\mathbf{P}(\max_{1\leq m\leq n} |S_m| > \epsilon) \leq \frac{\mathbf{E}|S_n|^p}{\epsilon^p} \tag{1.5}$$

for any $p \in [0, 1]$. In order to appreciate the beauty of this estimate it is worth noting that they give precisely the same estimate for $\max |S_m|$ as one would get for S_n itself via the rough Markov–Chebyshev inequality.

Let us recall now the four basic notions of the convergence of random variables. Let X and X_n, $n \in \mathbb{N}$, be S-valued random variables, where (S, ρ) is a metric space with the distance ρ. One says that X_n converges to X

1. *almost surely* or *with probability* 1 if $\lim_{n\to\infty} X_n(\omega) = X(\omega)$ almost surely;
2. *in probability* if for any $\epsilon > 0$ $\lim_{n\to\infty} \mathbf{P}(\rho(X_n, X) > \epsilon) = 0$;
3. *in distribution* if p_{X_n} *weakly converges* to p_X, i.e. if

$$\lim_{n\to\infty} \int_{\mathbb{R}^d} f(x) p_{X_n}(dx) = \int_{\mathbb{R}^d} f(x) p_X(dx)$$

for all bounded continuous functions f.

If S is $\mathbb{R}^d$, or a Banach space, X is said to converge *in* L^p ($1 \le p < \infty$) if $\lim_{n\to\infty} \mathbf{E}(|X_n - X|^p) = 0$.

To visualize these notions, let us start with two examples.

1. Consider the following sequence of indicator functions $\{X_n\}$ on $[0,1]$: $\mathbf{1}_{[0,1]}$, $\mathbf{1}_{[0,1/2]}$, $\mathbf{1}_{[1/2,1]}$, $\mathbf{1}_{[0,1/3]}$, $\mathbf{1}_{[1/3,2/3]}$, $\mathbf{1}_{[2/3,1]}$, $\mathbf{1}_{[0,1/4]}$, $\mathbf{1}_{[1/4,2/4]}$, etc. Then $X_n \to 0$ as $n \to \infty$ in probability and in all L^p, $p \ge 1$, but not a.s. In fact $\limsup X_n(x) = 1$ and $\liminf X_n(x) = 0$ for each x so that $X_n(x) \to X(x)$ nowhere.
2. Choosing $X_n = X'$ for all n with X' distributed like X but independent of it, shows that $X_n \to X$ in distribution does not imply in general $X_n - X \to 0$.

The following statement gives instructive criteria for convergence in probability and a.s. and establish the link between them.

Proposition 1.1.1. 1. *$X_n \to X$ in probability if and only if*

$$\lim_{n\to\infty} \mathbf{E}\left(\frac{\rho(X_n, X)}{1 + \rho(X_n, X)}\right) = 0 \Longleftrightarrow \lim_{n\to\infty} \mathbf{E}(1 \wedge \rho(X_n, X)) = 0. \tag{1.6}$$

2. *$X_n \to X$ a.s. if and only if*

$$\lim_{m\to\infty} \mathbf{P}(\sup_{n\ge m} \rho(X_n, X) > \epsilon) = 0 \tag{1.7}$$

for all $\epsilon > 0$.

3. *Almost sure convergence implies convergence in probability.*
4. *Any sequence converging in probability has a subsequence converging a.s.*

Proof. 1. Convergence in probability follows from (1.6), because by Chebyshev's inequality

$$\mathbf{P}(\rho(X_n, X) > \epsilon) = \mathbf{P}(1 \wedge \rho(X_n, X) > \epsilon) \le \frac{1}{\epsilon}\mathbf{E}(1 \wedge \rho(X_n, X))$$

for $\epsilon \in (0, 1)$. The converse statement follows from the inequalities

$$\mathbf{E}\left(\frac{\rho(X_n, X)}{1 + \rho(X_n, X)}\right) \le \mathbf{E}(1 \wedge \rho(X_n, X)) \le \epsilon + \mathbf{P}(\rho(X_n, X) > \epsilon).$$

2. The event $X_n \to X$ is the complement of the event

$$B = \bigcup_{r \in \mathbb{Q}} B_r, \quad B_r = \bigcap_{m \in \mathbb{Q}} \{\sup_{n \geq m} |X_n - X| > 1/r\},$$

i.e., a.s. convergence is equivalent to $\mathbf{P}(B) = 0$ and hence to $\mathbf{P}(B_r) = 0$ for all r.

3. This is an obvious consequence of either of statements 1 or 2.

4. If X_n converge in probability, using statement 2, we can choose a subsequence X_k such that

$$\mathbf{E} \sum_k (1 \wedge \rho(X_k, X)) = \sum_k \mathbf{E}(1 \wedge \rho(X_k, X)) < \infty,$$

implying that $\sum_k (1 \wedge \rho(X_k, X)) < \infty$ a.s. and hence $\rho(X_k, X) \to 0$ a.s. □

Proposition 1.1.2. *L^p-convergence $\Rightarrow$ convergence in probability $\Rightarrow$ weak convergence. Finally, weak convergence to a constant implies convergence in probability.*

Proof. The first implication follows from Chebyshev's inequality.

For the second one assume S is $\mathbb{R}^d$. Decompose the integral $\int |f(X_n(\omega)) - f(X(\omega))| \mathbf{P}(d\omega)$ into the sum $I_1 + I_2 + I_3$ of three terms over the sets $\{|X_n - X| > \delta\}$, $\{|X_n - X| \leq \delta, |X| > K\}$ and $\{|X_n - X| \leq \delta, |X| \leq K\}$. First choose K such that $\mathbf{P}(|X| > K + 1) < \epsilon$. Next, by the uniform integrability of f on the ball of radius K (here we use its compactness), choose δ such that $|f(x) - f(y)| < \epsilon$ for $|x - y| < \delta$. By the convergence in probability, choose N such that $\mathbf{P}(|X_n - X| > \delta) < \epsilon$ for $n > N$. Then

$$I_1 + I_2 + I_3 \leq 3\epsilon \|f\| + \epsilon.$$

For general metric spaces S a proof can be obtained from statements 1 and 4 of Proposition 1.1.1.

Finally, the last statement follows from (1.6). □

Exercise 1.1.3. If probability measures p_n on $\mathbb{R}^d$ converge weakly to a measure p as $n \to \infty$, then the sequence $p_n(A)$ converges to $p(A)$ for any open or closed set A such that $p(\partial A) = 0$ (where ∂A is the boundary of A).

A family H of $L^1(\Omega, \mathcal{F}, \mu)$ is called *uniformly integrable* if

$$\lim_{c \to \infty} \sup_{X \in H} \mathbf{E}(|X| \mathbf{1}_{|X| > c}) = 0.$$

Proposition 1.1.3. *If either* (i) $\sup_{X \in H} \mathbf{E}(|X|^p) < \infty$ *for a* $p > 1$, *or* (ii) *there exists an integrable r.v. Y s.t. $|X| \leq Y$ for all $X \in H$, then H is uniformly integrable.*

Proof. Follows from the inequalities.

$$\text{(i)} \qquad \mathbf{E}(|X|\mathbf{1}_{|X|>c}) < \frac{1}{c^{p-1}}\mathbf{E}(|X|^p\mathbf{1}_{|X|>c}) < \frac{1}{c^{p-1}}\mathbf{E}(|X|^p),$$

$$\text{(ii)} \qquad \mathbf{E}(|X|\mathbf{1}_{|X|>c}) < \mathbf{E}(Y\mathbf{1}_{Y>c}). \qquad \square$$

Proposition 1.1.4. *If $X_n \to X$ a.s. and $\{X_n\}$ is uniformly integrable, then $X_n \to X$ in L^1.*

Proof. Decompose the integral $\int |X_n - X| p(d\omega)$ into the sum of the three integrals over the domains $\{|X_n - X| > \epsilon\}$, $\{|X_n - X| \le \epsilon, |X| \le c\}$ and $\{|X_n - X| \le \epsilon, |X| > c\}$. These can be made small respectively because $X_n \to X$ in probability (as it holds a.s.), by dominated convergence and by uniform integrability. □

Two famous theorems of integration theory, the dominated and monotone convergence theorems, give easy-to-use criteria for a.s. convergence to imply convergence in L_1.

The following famous result allows one to transfer weak convergence to a.s. convergence by an appropriate coupling.

Theorem 1.1.3 (Skorohod coupling). *Let $\xi, \xi_1, \xi_2, \dots$ be a sequence of random variables with values in a separable metric space S such that $\xi_n \to \xi$ weakly as $n \to \infty$. Then there exists a probability space with some S-valued random variables $\eta, \eta_1, \eta_2, \dots$ distributed as $\xi, \xi_1, \xi_2, \dots$ respectively and such that $\eta_n \to \eta$ a.s. as $n \to \infty$.*

The following celebrated convergence result is one of the oldest in probability theory.

Theorem 1.1.4 (Weak law of large numbers). *If $\xi_1, \xi_2, \dots$ is a collection of i.i.d. random variables with $\mathbf{E}\xi_j = m$ and $\operatorname{Var}\xi_j < \infty$, then the means $(\xi_1 + \dots + \xi_n)/n$ converge to m in probability and in L^2.*

Proof. By (1.3)

$$\operatorname{Var}\left(\frac{\xi_1 + \dots + \xi_n}{n} - m\right) = \operatorname{Var}\frac{(\xi_1 - m) + \dots + (\xi_n - m)}{n} = \frac{\operatorname{Var}\xi_1}{n},$$

implying convergence in L_2. Hence by Chebyshev's inequality

$$\mathbf{P}\left(\left|\frac{\xi_1 + \dots + \xi_n}{n} - m\right| > \epsilon\right) \le \frac{\operatorname{Var}\xi_1}{n\epsilon^2},$$

implying convergence in probability. □

Using the stronger Kolmogorov's inequality allows one to get the following improvement.

Theorem 1.1.5 (Strong law of large numbers). *Let S_n denote the sums $\xi_1 + \xi_2 + \cdots + \xi_n$ for a sequence $\xi_1, \xi_2, \dots$ of independent zero-mean random variables such that $\mathbf{E}|\xi_j|^2 = \sigma^2 < \infty$ for all j. Then the means S_n/n converge to 0 a.s.*

Proof. By (1.7) we have to show that

$$\lim_{m\to\infty} \mathbf{P}\left(\sup_{n\geq m} \left| \frac{S_n}{n} \right| > \epsilon \right) = 0.$$

Denote by A_k the events

$$A_k = \left\{ \max_{2^{k-1}\leq n<2^k} \left| \frac{S_n}{n} \right| > \epsilon \right\}.$$

Then by (1.4)

$$\mathbf{P}(A_k) \leq \mathbf{P}(\max_{2^{k-1}\leq n<2^k} |S_n| > \epsilon 2^{k-1}) \leq 2^{-k} \frac{4\sigma^2}{\epsilon^2}.$$

Hence the sum $\sum_k \mathbf{P}(A_k)$ converges. Consequently

$$\mathbf{P}\left(\sup_{n\geq 2^{m-1}} \left| \frac{S_n}{n} \right| > \epsilon \right) \leq \sum_{k=m}^{\infty} \mathbf{P}(A_k) \to 0$$

as $m \to \infty$ for any ϵ. □

Remark 1. Using (1.5) instead of (1.4) allows us to prove the above theorem under a weaker assumption that $\mathbf{E}|\xi_j|^p = \omega < \infty$ for some $p > 1$. It is instructive to see where this proof breaks down in case $p = 1$. By more involved arguments one can still prove the strong LLN if only $\mathbf{E}|\xi_j| < \infty$, but assuming that ξ_j are i.i.d.

The following result is routinely used in stochastic analysis to check a validity of a certain property for elements of $\sigma(\Gamma)$, where Γ is a collection of subsets closed under intersection. According to the theorem it is sufficient to check that the validity of this property is preserved under set subtraction and countable unions.

Theorem 1.1.6 (Monotone class theorem). *Let $\mathcal{S}$ be a collection of subsets of a set Ω s.t.*

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

(ii) $A, B \in \mathcal{S} \Rightarrow A \setminus B \in \mathcal{S}$,

(iii) $A_1 \subset A_2 \subset \cdots \subset A_n \in \mathcal{S} \Rightarrow \bigcup_n A_n \in \mathcal{S}$.

If a collection of subsets Γ belongs to S and is closed under pairwise intersection, then $\sigma(\Gamma) \in \mathcal{S}$.

Exercise 1.1.4. For $S \subset \mathbb{R}^d$ the *universal* σ-field $\mathcal{U}(S)$ is defined as the intersection of the completions of $\mathcal{B}(S)$ with respect to all probability measures on S. The $(\mathcal{U}(S), \mathcal{B}(S))$-measurable functions are called *universally measurable.* Show that a real valued function f is universally measurable if and only if for every probability measure μ on S there exists a Borel measurable function g_μ such that $\mu\{x : f(x) \neq g_\mu(x)\} = 0$. Hint for "only if" part: show that

$$f(x) = \inf\{r \in \mathbb{Q} : x \in U(r)\}, \quad \text{where } U(r) = \{x \in S : f(x) \leq r\}.$$

Since $U(r)$ belong to the completion of the Borel σ-algebra with respect to μ there exist $B(r)$, $r \in \mathbb{Q}$, such that

$$\mu\Big(\bigcup_{r \in \mathbb{Q}} (B(r) \Delta U(r))\Big) = 0.$$

Define

$$g_\mu(x) = \inf\{r \in \mathbb{Q} : x \in B(r)\}.$$

1.2 Characteristic functions

As we already mentioned, expectation and variance supply two basic numeric characteristics of a random variable. Some additional information on its behavior can be obtained from higher moments. A complete analytical description of a random variable is given by the characteristic function, which we recall briefly in this section.

If p is a probability measure on $\mathbb{R}^d$ its *characteristic function* is the function $\phi_p(y) = \int e^{i(y,x)} p(dx)$. For a $\mathbb{R}^d$-valued r.v. X its *characteristic function* is defined as the characteristic function $\phi_X = \phi_{p_X}$ of its law p_X, i.e.

$$\phi_X(y) = \mathbf{E} e^{i(y,X)} = \int_{\mathbb{R}^d} e^{i(y,x)} p_X(dx).$$

Any characteristic function is a continuous function, which clearly follows from the inequalities

$$|\phi_X(y+h) - \phi_X(y)| \leq \mathbf{E}|e^{ihX} - 1| \leq \max_{|x| \leq a} |e^{ihx} - 1| + 2\mathbf{P}(|X| > a). \tag{1.8}$$

Theorem 1.2.1 (Riemann–Lebesgue lemma). *If a probability measure p has a density, then ϕ_p belongs to $C_\infty(\mathbb{R}^d)$. In other words, the inverse Fourier transform*

$$f \mapsto F^{-1} f(y) = (2\pi)^{-d/2} \int e^{i(y,x)} f(x) dx$$

is a bounded linear operator $L^1(\mathbb{R}^d) \to C_\infty(\mathbb{R}^d)$.

Sketch of the proof. Reduce to the case, when f is a continuously differentiable function with a compact support. For this case use integration by parts.

For a vector $m \in \mathbb{R}^d$ and a positive definite $d \times d$-matrix A, a r.v. X is called *Gaussian* (or has *Gaussian distribution*) *with mean* m *and covariance* A, denoted by $N(m, A)$, whenever its characteristic function is

$$\phi_{N(m,A)}(y) = \exp\left\{i(m, y) - \frac{1}{2}(y, Ay)\right\}.$$

It is easy to deduce that $m = \mathbf{E}(X)$ and $A_{ij} = \mathbf{E}((X_i - m_i)(X_j - m_j))$ and that if A is non-degenerate, $N(m, A)$ random variables have distribution with the pdf

$$f(x) = \frac{1}{(2\pi)^{d/2}\sqrt{\det(A)}} \exp\left\{-\frac{1}{2}(x - m, A^{-1}(x - m))\right\}.$$

It is useful to observe that if X_1 and X_2 are independent $\mathbb{R}^d$-valued random variables with laws μ_1, μ_2 and characteristic functions ϕ_1 and ϕ_2, then $X_1 + X_2$ has the characteristic function $\phi_1\phi_2$ and the law given by the *convolution* $\mu_1 \star \mu_2$ defined by

$$(\mu_1 \star \mu_2)(A) = \int_{\mathbb{R}^d} \mu_1(A - x)\mu_2(dx).$$

Similarly, for independent random variables $X_1, \dots, X_n$ with the laws $\mu_1, \dots, \mu_n$ and characteristic functions $\phi_1, \dots, \phi_n$, the sum $X_1 + \dots + X_n$ has the characteristic function $\phi_1 \cdots \phi_n$ and the law $\mu_1 \star \dots \star \mu_n$. In particular, if $X_1, \dots, X_n$ are independent identically distributed (common abbreviation *i.i.d.*) random variables, then the sum $X_1 + \dots + X_n$ has the characteristic function ϕ_1^n and the law $\mu_1 \star \dots \star \mu_1$.

The next exercise anticipates the discussion of weak compactness or measures given at the end of this chapter.

Exercise 1.2.1. Show that if probability distributions p_n on $\mathbb{R}^d$, $n \in \mathbb{N}$, converge weakly to a probability distribution p, then

(i) the family p_n is *tight*, i.e.

$$\forall \epsilon > 0 \, \exists K > 0 : \forall n, \; p_n(|x| > K) < \epsilon;$$

(ii) their characteristic functions ϕ_n converge uniformly on compact sets.

Hint: for (ii) use tightness and representation (1.8) to show that the family ϕ_n is equicontinuous, i.e.

$$\forall \epsilon \, \exists \delta : |\phi_n(y + h) - \phi(y)| < \epsilon \quad \forall h < \delta, n \in \mathbb{N},$$

which implies uniform convergence.

Theorem 1.2.2 (Glivenko's theorem). *If ϕ_n, $n \in \mathbb{N}$, and ϕ are the characteristic functions of probability distributions p_n and p on $\mathbb{R}^d$, then* $\lim_{n\to\infty} \phi_n(y) = \phi(y)$ *for each $y \in \mathbb{R}^d$ if and only if p_n converge to p weakly.*

Theorem 1.2.3 (Lévy's theorem). *If ϕ_n, $n \in \mathbb{N}$, is a sequence of characteristic functions of probability distributions on $\mathbb{R}^d$ and* $\lim_{n\to\infty} \phi_n(y) = \phi(y)$ *for each $y \in \mathbb{R}^d$ for some function ϕ, which is continuous at the origin, then ϕ is itself a characteristic function (and so the corresponding distributions converge weekly as above).*

The following exercise suggests using Lévy's theorem to prove a particular case of the fundamental Prohorov criterion for tightness.

Exercise 1.2.2. Show that if a family of probability measures p_α on $\mathbb{R}^d$ is tight, then it is *relatively weakly compact*, i.e. any sequence of this family has a weakly convergent subsequence. Hint: tight $\Rightarrow$ family of characteristic functions is equicontinuous (by (1.8)), and hence is relatively compact in the topology of uniform convergence on compact sets. Finally use Lévy's theorem.

Exercise 1.2.3. (i) Show that a finite linear combination of $\mathbb{R}^d$-valued Gaussian random variables is again a Gaussian r.v.

(ii) Show that if a sequence of $\mathbb{R}^d$-valued Gaussian random variables converges in distribution to a random variable, then the limiting random variable is again Gaussian.

(iii) Show that if (X, Y) is a $\mathbb{R}^2$-valued Gaussian random variables, then X and Y are uncorrelated if and only if they are independent.

Theorem 1.2.4 (Bochner's criterion). *A function $\phi : \mathbb{R}^d \mapsto \mathbb{C}$ is a characteristic function of a probability distribution if and only if it satisfies the following three properties:*

(i) $\phi(0) = 1$;

(ii) *ϕ is continuous at the origin;*

(iii) *ϕ is positive definite, which means that*

$$\sum_{j,k=1}^{d} c_j \bar{c}_k \phi(y_j - y_k) \geq 0$$

for all real $y_1, \dots, y_d$ and all complex $c_1, \dots, c_d$.

Remark 2. To prove the "only if" part of Bochner's theorem is easy. In fact:

$$\sum_{j,k=1}^{d} c_j \bar{c}_k \phi_X(y_j - y_k) = \int_{\mathbb{R}^d} \sum_{j,k=1}^{d} c_j \bar{c}_k e^{i(y_j - y_k, x)} p_X(dx)$$
$$= \int_{\mathbb{R}^d} \Big(\sum_{j=1}^{d} c_j e^{i(y_j, x)}\Big)^2 p_X(dx) \geq 0.$$

1.3 Conditioning

Formally speaking, probability can be considered as a part of measure theory. What actually makes it special and fills it with new intuitive and practical content is *conditioning*. On the one hand, conditioning is a method for updating our perception of the probability of an event based on the information received (conditioning on an event). On the other hand, it is a method for characterizing random variables from their coarse description that neglects certain irrelevant details, like increasing the scale of an atlas or an image (conditioning with respect to a partition or subalgebra).

Assume a finite *partition* $\mathcal{A} = \{A_i\}$ of our probability space $(\Omega, \mathcal{F}, \mathbf{P})$ is given, i.e. it is decomposed into the union of non-intersecting measurable subsets $A_1, \ldots, A_m$. Assume that for certain purposes we do not need to distinguish the points belonging to the same element of the partition. In other words, we would like to reduce our original probability space to the simpler one $(\Omega, \mathcal{F}_{\mathcal{A}}, \mathbf{P})$, where $\mathcal{F}_{\mathcal{A}}$ is the finite σ-algebra generated by the partition $\mathcal{A}$ (that consists of all unions of the elements of this partition). Now, if we have a random variable X on $(\Omega, \mathcal{F}, \mathbf{P})$, how should we reasonably project it on the reduced probability space $(\Omega, \mathcal{F}_{\mathcal{A}}, \mathbf{P})$? Clearly such a projection $\tilde{X}$ should be measurable with respect to $\mathcal{F}_{\mathcal{A}}$, meaning that it should be constant on each A_i. Moreover, we want the averages of X and $\tilde{X}$ to coincide on each A_i. This implies that the value of $\tilde{X}$ on A_i should equal the average value of X on A_i. The random variable $\tilde{X}$, obtained in this way, is denoted by $\mathbf{E}(X|\mathcal{F}_{\mathcal{A}})$ and is called the *conditional expectation* of X given the σ-algebra $\mathcal{F}_{\mathcal{A}}$ (or equivalently, given the partition $\mathcal{A}$). Hence, by definition,

$$\mathbf{E}(X|\mathcal{F}_{\mathcal{A}})(\omega) = \int_{A_i} X(\omega)\mathbf{P}(d\omega)/\mathbf{P}(A_i), \quad \omega \in A_i, \tag{1.9}$$

for all $i = 1, \ldots, m$. Equivalently, $\mathbf{E}(X|\mathcal{F}_{\mathcal{A}})(\omega)$ is defined as a random variable on $(\Omega, \mathcal{F}_{\mathcal{A}}, \mathbf{P})$ such that

$$\int_A \mathbf{E}(X|\mathcal{F}_{\mathcal{A}})(\omega)\mathbf{P}(d\omega) = \int_A X(\omega)\mathbf{P}(d\omega)$$

for any $A \in \mathcal{F}_{\mathcal{A}}$.

This definition can be straightforwardly extended to arbitrary subalgebras. Namely, for a random variable X on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ and a σ-subalgebra $\mathcal{G}$ of $\mathcal{F}$, the *conditional expectation* of X given $\mathcal{G}$ is defined as a random variable $\mathbf{E}(X|\mathcal{G})$ on the probability space $(\Omega, \mathcal{G}, \mathbf{P})$ such that

$$\int_A \mathbf{E}(X|\mathcal{G})(\omega)\mathbf{P}(d\omega) = \int_A X(\omega)\mathbf{P}(d\omega)$$

for any $A \in \mathcal{G}$. Clearly, if such a random variable exists it is uniquely defined (up to the natural equivalence of random variables in $(\Omega, \mathcal{G}, \mathbf{P})$), because the difference of any two random variables with the required property has vanishing integrals over any measurable set in $(\Omega, \mathcal{G}, \mathbf{P})$, and hence this difference vanishes a.s. However, the existence of conditional expectation is not so obvious for infinite subalgebras. In fact, the defining equation (1.9) does not make sense in case $\mathbf{P}(A_i) = 0$. Hence in the general case, another approach to the construction of conditional expectation is needed, which we now describe.

For a given measure space $(S, \mathcal{F}, \mu)$, a measure ν on $(S, \mathcal{F})$ is called *absolutely continuous* with respect to μ if $\nu(A) = 0$ whenever $A \in \mathcal{F}$ and $\mu(A) = 0$. Two measures are called *equivalent* if they are mutually absolutely continuous.

Theorem 1.3.1 (Radon–Nikodym theorem). *If μ is σ-finite and ν is finite and absolutely continuous with respect to μ, then there exists a unique (up to almost sure equality) non-negative measurable function g on S such that for all $A \in \mathcal{F}$*

$$\nu(A) = \int_A g(x)\mu(dx).$$

This g is called the Radon–Nikodym derivative of ν with respect to μ and is often denoted $d\nu/d\mu$.

Let X be an integrable random variable on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ and let $\mathcal{G}$ be a sub-σ-algebra of $\mathcal{F}$. If $X \geq 0$ everywhere, the formula

$$Q_X(A) = \mathbf{E}(X\mathbf{1}_A) = \int_A X(\omega)\mathbf{P}(d\omega)$$

for $A \in \mathcal{G}$ defines a measure Q_X on $(\Omega, \mathcal{G})$ that is obviously absolutely continuous with respect to $\mathbf{P}$. The r.v. $dQ_X/d\mathbf{P}$ on $(\Omega, \mathcal{G}, \mathbf{P})$ is called the *conditional expectation of X with respect to $\mathcal{G}$*, and is usually denoted $\mathbf{E}(X|\mathcal{G})$. If X is not supposed to be positive one defines the *conditional expectation* as $\mathbf{E}(X|\mathcal{G}) = \mathbf{E}(X^+|\mathcal{G}) - \mathbf{E}(X^-|\mathcal{G})$. Clearly this new definition complies with the previous one, as so defined $Y = \mathbf{E}(X|\mathcal{G})$ is a r.v. on $(\Omega, \mathcal{G}, \mathbf{P})$ satisfying

$$\int_A Y(\omega)\mathbf{P}(d\omega) = \int_A X(\omega)\mathbf{P}(d\omega) \tag{1.10}$$

for all $A \in \mathcal{G}$ or, equivalently,

$$\mathbf{E}(YZ) = \mathbf{E}(XZ) \tag{1.11}$$

for any bounded $\mathcal{G}$-measurable Z.

If $X = (X_1, \dots, X_d) \in \mathbb{R}^d$, then

$$\mathbf{E}(X|\mathcal{G}) = (\mathbf{E}(X_1|\mathcal{G}), \dots, \mathbf{E}(X_n|\mathcal{G})).$$

The following result collects the basic properties of the conditional expectation.

Theorem 1.3.2. (i) $\mathbf{E}(\mathbf{E}(X|\mathcal{G})) = \mathbf{E}(X)$*;*

(ii) *if Y is $\mathcal{G}$-measurable, then $\mathbf{E}(XY|\mathcal{G}) = Y\mathbf{E}(X|\mathcal{G})$ a.s.;*

(iii) *if Y is $\mathcal{G}$-measurable and X is independent of $\mathcal{G}$, then a.s.*

$$\mathbf{E}(XY|\mathcal{G}) = Y\mathbf{E}(X),$$

and more generally

$$\mathbf{E}(f(X,Y)|\mathcal{G}) = G_f(Y) \tag{1.12}$$

a.s. for any bounded Borel function f, where $G_f(y) = \mathbf{E}(f(X,y))$ a.s.;

(iv) *if $\mathcal{H}$ is a sub-σ-algebra of $\mathcal{G}$ then $\mathbf{E}(\mathbf{E}(X|\mathcal{G})|\mathcal{H}) = \mathbf{E}(X|\mathcal{H})$ a.s. (this property is called the chain rule for conditioning);*

(v) *the mapping $X \mapsto \mathbf{E}(X|\mathcal{G})$ is an orthogonal projection $L^2(\Omega, \mathcal{F}, \mathbf{P}) \to L^2(\Omega, \mathcal{G}, \mathbf{P})$;*

(vi) $X_1 \le X_2 \Rightarrow \mathbf{E}(X_1|\mathcal{G}) \le \mathbf{E}(X_2|\mathcal{G})$ *a.s.;*

(vii) *the mapping $X \mapsto \mathbf{E}(X|\mathcal{G})$ is a linear contraction $L^1(\Omega, \mathcal{F}, \mathbf{P}) \to L^1(\Omega, \mathcal{G}, \mathbf{P})$.*

Exercise 1.3.1. Prove the above theorem. Hint: (ii) consider first the case with Y being an indicator function of a $\mathcal{G}$-measurable set; (v) assume $X = Y + Z$ with Y from $L^2(\Omega, \mathcal{G}, \mathbf{P})$ and Z from its orthogonal complement and show that $Y = \mathbf{E}(X|\mathcal{G})$. (vi) Follows from an obvious remark that $X \ge 0 \Rightarrow \mathbf{E}(X|\mathcal{G}) \ge 0$.

Remark 3. Property (v) above can be used to give an alternative construction of conditional expectation by-passing the Radon–Nikodym theorem.

If Z is a r.v. on $(\Omega, \mathcal{F}, \mathbf{P})$ one calls $\mathbf{E}(X|\sigma(Z))$ the *conditional expectation of X with respect to Z* and denotes it briefly by $\mathbf{E}(X|Z)$.

The measurability of $\mathbf{E}(X|Z)$ with respect to $\sigma(Z)$ implies that $\mathbf{E}(X|Z)$ is a constant on any Z-level set $\{\omega : Z(\omega) = z\}$. One denotes this constant by $\mathbf{E}(X|Z = z)$ and calls it the *conditional expectation of X given $Z = z$*. From statement (iv) of

Theorem 1.3.2 it follows that

$$\mathbf{E}(X) = \int \mathbf{E}(X|Z)(\omega)\mathbf{P}(d\omega) = \int \mathbf{E}(X|Z=z)p_Z(dz) \tag{1.13}$$

(the second equality is obtained by applying (1.2) to $f(Z(\omega)) = \mathbf{E}(X|Z)(\omega)$).

Let X and Z be $\mathbb{R}^d$ and respectively $\mathbb{R}^m$-valued r.v. on $(\Omega, \mathcal{F}, \mathbf{P})$, and let $\mathcal{G}$ be a sub-σ-algebra of $\mathcal{F}$. The *conditional probability of* X *given* $\mathcal{G}$ and X *given* $Z = z$ respectively are defined as

$$\mathbf{P}_{X|\mathcal{G}}(B;\omega) \equiv \mathbf{P}(X \subset B|\mathcal{G})(\omega) = \mathbf{E}(\mathbf{1}_B(X)|\mathcal{G})(\omega), \quad \omega \in \Omega;$$
$$\mathbf{P}_{X|Z=z}(B) \equiv \mathbf{P}(X \subset B|Z=z) = \mathbf{E}(\mathbf{1}_B(X)|Z=z),$$

for Borel sets B, or equivalently through the equations

$$\mathbf{E}(f(X)|\mathcal{G})(\omega) = \int_{\mathbb{R}^d} f(x)\mathbf{P}_{X|\mathcal{G}}(dx;\omega),$$
$$\mathbf{E}(f(X)|Z=z) = \int_{\mathbb{R}^d} f(x)\mathbf{P}_{X|Z=z}(dx)$$

for bounded Borel functions f. Of course $\mathbf{P}_{X|Z=z}(B)$ is just the common value of $\mathbf{P}_{X|Z}(B;\omega)$ on the set $\{\omega : Z(\omega) = z\}$.

It is possible to show (though this is not obvious) that, for any $\mathbb{R}^d$-r.v. X, the *regular conditional probability of* X *given* $\mathcal{G}$ exists, i.e. such a version of conditional probability that $\mathbf{P}_{X|\mathcal{G}}(B,\omega)$ is a probability measure on $\mathbb{R}^d$ as a function of B for each ω (notice that from the above discussion the required additivity of conditional expectations hold a.s. only so that they may fail to define a probability even a.s.) and is $\mathcal{G}$-measurable as a function of ω. Hence one can define *conditional r.v.* $X_{\mathcal{G}}(\omega)$, $X_Z(\omega)$ and $X_{Z=z}$ as r.v. with the corresponding conditional distributions.

Proposition 1.3.1. *For a Borel function* h

$$\mathbf{E}h(X,Z) = \int h(x,z)\mathbf{P}_{X|Z=z}(dx)p_Z(dz) \tag{1.14}$$

whenever the l.h.s. is well defined.

Proof. It is enough to show this for the functions of the form $h(X,Z) = f(X)\mathbf{1}_{Z\in C}$ for a measurable C. And from (1.13) it follows that

$$\begin{aligned}\mathbf{E}f(X)\mathbf{1}_{Z\in C} &= \int_{Z\in C} \mathbf{E}(f(X)|Z)(\omega)\mathbf{P}(d\omega) = \int_C \mathbf{E}(f(X)|Z=z)p_Z(dz)\\ &= \int_C \int_{\mathbb{R}^d} f(x)\mathbf{P}_{X|Z=z}(dx)p_Z(dz).\end{aligned}$$

□

For instance, if X, Z are discrete r.v. with joint probability $\mathbf{P}(X = i, Z = j) = p_{ij}$, then the conditional probabilities $p(X = i|Z = j)$ are given by the usual formula $p_{ij}/\mathbf{P}(Z = j)$.

On the other hand, if X, Z are r.v. with a joint probability density function $f_{X,Z}(x,z)$, then the conditional r.v. $X_{Z=z}$ has a probability density function

$$f_{X_{Z=z}}(x) = f_{X,Z}(x,z)/f_Z(z)$$

whenever f_Z does not vanish. In order to see this, one has to compare (1.14) with the equation

$$\begin{aligned}\mathbf{E}h(X,Z) &= \int h(x,z) f(x,z)\,dxdz \\ &= \int h(x,z)\frac{f(x,z)}{f_Z(z)}\,dx f_Z(z)\,dz.\end{aligned}$$

Theorem 1.3.3. *Let X be a integrable variable on $(\Omega, \mathcal{F}, \mathbf{P})$ and let $\mathcal{G}_n$ be either*

(i) *an increasing sequence of sub-σ-algebras of $\mathcal{F}$ with $\mathcal{G}$ being the minimal σ-algebra containing all $\mathcal{G}_n$, or*

(ii) *a decreasing sequence of sub-σ-algebras of $\mathcal{F}$ with $\mathcal{G} = \bigcap_{n=1}^{\infty} \mathcal{G}_n$.*

Then a.s. and in L^1

$$\mathbf{E}(X|\mathcal{G}) = \lim_{n\to\infty} \mathbf{E}(X|\mathcal{G}_n). \tag{1.15}$$

Furthermore, if $X_n \to X$ a.s. and $|X_n| < Y$ for all n, where Y is an integrable random variable, then a.s. and in L^1

$$\mathbf{E}(X|\mathcal{G}) = \lim_{n\to\infty} \mathbf{E}(X_n|\mathcal{G}_n). \tag{1.16}$$

Proof. We shall sketch the proof of the convergence in L^1 (a.s. convergence is a bit more involved, and we shall neither prove, nor use it), say, for increasing sequences. Any r.v. of the form $\mathbf{1}_B$ with $B \in \mathcal{G}$ can be approximated in L^2 by a $\mathcal{G}_n$-measurable r.v. ξ_n. Hence the same holds for any r.v. from $L^2(\Omega, \mathcal{F}, \mathbf{P})$. As $\mathbf{E}(X|\mathcal{G}_n)$ is the best approximation (L^2-projection) for $\mathbf{E}(X|\mathcal{G})$ one obtains (1.15) for $X \in L^2(\Omega, \mathcal{F}, \mathbf{P})$, and hence for $X \in L^1(\Omega, \mathcal{F}, \mathbf{P})$ by density arguments. Next,

$$\mathbf{E}(X_n|\mathcal{G}_n) - \mathbf{E}(X|\mathcal{G}) = \mathbf{E}(X_n - X|\mathcal{G}_n) + (\mathbf{E}(X|\mathcal{G}_n) - \mathbf{E}(X|\mathcal{G})).$$

Since $|X_n| < Y$ and $X_n \to X$ a.s. one concludes that $X_n \to X$ in L^1 by dominated convergence. Hence as $n \to \infty$

$$\mathbf{E}(\mathbf{E}|X_n - X||\mathcal{G}_n) = \mathbf{E}|X_n - X| \to 0. \qquad \square$$

Theorem 1.3.4. *If $X \in L^1(\Omega, \mathcal{F}, \mathbf{P})$, the family of r.v. $\mathbf{E}(X|\mathcal{G})$, as $\mathcal{G}$ runs through all sub-σ-algebra of $\mathcal{F}$, is uniformly integrable.*

Proof.

$$\mathbf{1}_{|\mathbf{E}(X|\mathcal{G})|>c}\mathbf{E}(X|\mathcal{G}) = \mathbf{E}(X\mathbf{1}_{|\mathbf{E}(X|\mathcal{G})|>c}|\mathcal{G}),$$

because $\{|\mathbf{E}(X|\mathcal{G})| > c\} \in \mathcal{G}$. Hence

$$\begin{aligned}\mathbf{E}(\mathbf{1}_{|\mathbf{E}(X|\mathcal{G})|>c}\mathbf{E}(X|\mathcal{G})) &\le \mathbf{E}(\mathbf{1}_{|\mathbf{E}(X|\mathcal{G})|>c}|X|)\\ &\le \mathbf{E}(|X|\mathbf{1}_{|X|>d}) + d\mathbf{P}(|\mathbf{E}(X|\mathcal{G})| > c)\\ &\le \mathbf{E}(|X|\mathbf{1}_{|X|>d}) + \frac{d}{c}\mathbf{E}(|X|),\end{aligned}$$

where in the last inequality Markov's inequality was used. First choose d to make the first term small, then c to make the second one small. □

One says that two sigma algebra $\mathcal{G}_1$, $\mathcal{G}_2$ coincide on a set $A \in \mathcal{G}_1 \cap \mathcal{G}_2$ whenever $A \cap \mathcal{G}_1 = A \cap \mathcal{G}_2$.

Theorem 1.3.5 (Locality of conditional expectation). *Let the σ-algebras $\mathcal{G}_1, \mathcal{G}_2 \in \mathcal{F}$ and the random variables $X_1, X_2 \in L^1(\Omega, \mathcal{F}, \mathbf{P})$ be such that $\mathcal{G}_1 = \mathcal{G}_2$ and $X_1 = X_2$ on a set $A \in \mathcal{G}_1 \cap \mathcal{G}_1$. Then $\mathbf{E}(X_1|\mathcal{G}_1) = \mathbf{E}(X_2|\mathcal{G}_2)$ a.s. on A.*

Proof. The sets $\mathbf{1}_A\mathbf{E}(X_1|\mathcal{G}_1)$ and $\mathbf{1}_A\mathbf{E}(X_2|\mathcal{G}_2)$ are both $\mathcal{G}_1 \cap \mathcal{G}_2$-measurable, and for any $B \subset A$ such that $B \in \mathcal{G}_1$ (and hence $B \in \mathcal{G}_2$)

$$\int_B \mathbf{E}(X_1|\mathcal{G}_1)\mathbf{P}(d\omega) = \int_B X_1\mathbf{P}(d\omega) = \int_B X_2\mathbf{P}(d\omega) = \int_B \mathbf{E}(X_2|\mathcal{G}_2)\mathbf{P}(d\omega).$$ □

1.4 Infinitely divisible and stable distributions

Infinitely divisible laws studied here occupy an honored place in probability, because of their extreme modeling power in the variety of situations. They form the cornerstone for the theory of Lévy processes discussed later.

A probability measure μ on $\mathbb{R}^d$ with a characteristic function ϕ_μ is called *infinitely divisible* if, for all $n \in \mathbb{N}$, there exists a probability measure ν such that $\mu = \nu \star \cdots \star \nu$ (n times) or equivalently $\phi_\mu(y) = f^n(y)$ with f being a characteristic function of a probability measure.

A random variable X is called *infinitely divisible* whenever its law p_X is infinitely divisible. This is equivalent to the existence, for any n, of i.i.d. random variable Y_j, $j = 1, \dots, n$, such that $Y_1 + \cdots + Y_n$ has the law p_X.

For example, any Gaussian distribution is clearly infinitely divisible.

Another key example is given by a *Poisson random variable with mean (or parameter) $c > 0$*, which is a random variable N with the non-negative integers as range and law

$$\mathbf{P}(N = n) = \frac{c^n}{n!}e^{-c}.$$

One easily checks that $\mathbf{E}(N) = \mathrm{Var}(N) = c$ and that the characteristic function of N is $\phi_N(y) = \exp\{c(e^{iy} - 1)\}$. This implies that N is infinitely divisible.

Of importance is the following generalization. Let $Z(n)$, $n \in \mathbb{N}$, be a sequence of $\mathbb{R}^d$-valued i.i.d. random variables with the common law μ_Z. The random variable $X = Z(1) + \cdots + Z(N)$ is called a *compound Poisson random variable*. It represents a random walk (each step specified by a random variable distributed like $Z(1)$) with a random (Poisson) number of steps. Let us check that

$$\phi_X(y) = \exp\left\{\int_{\mathbb{R}^d} (e^{i(y,x)} - 1)c\mu_Z(dx)\right\}. \tag{1.17}$$

In fact,

$$\begin{aligned}
\phi_X(y) &= \sum_{n=0}^{\infty} \mathbf{E}(\exp\{i(y, Z(1) + \cdots + Z(N))\}|N = n)\mathbf{P}(N = n) \\
&= \sum_{n=0}^{\infty} \mathbf{E}(\exp\{i(y, Z(1) + \cdots + Z(n))\})\frac{c^n}{n!}e^{-c} \\
&= \sum_{n=0}^{\infty} \phi_Z^n(y)\frac{c^n}{n!}e^{-c} = \exp\{c(\phi_Z(y) - 1)\}.
\end{aligned}$$

A Borel measure ν on $\mathbb{R}^d$ is called a *Lévy measure* if $\nu(\{0\}) = 0$ and

$$\int_{\mathbb{R}^d} \min(1, x^2)\nu(dx) < \infty.$$

The major role played by these measures in the theory of infinite divisibility is revealed by the following fundamental result.

Theorem 1.4.1 (The Lévy–Khintchine formula). *For any $b \in \mathbb{R}^d$, a positive definite $d \times d$-matrix G and a Lévy measure ν the function*

$$\phi(u) = \exp\left\{i(b, u) - \frac{1}{2}(u, Gu) + \int_{\mathbb{R}^d} [e^{i(u,y)} - 1 - i(u, y)\mathbf{1}_{B_1}(y)]\nu(dy)\right\} \tag{1.18}$$

is a characteristic function of an infinitely divisible measure, where B_a denotes a ball of radius a in $\mathbb{R}^d$. Conversely, any infinite divisible distribution has a characteristic function of form (1.18).

Proof. We shall prove only the simpler first part. For the converse statement see e.g. [20], [302] and references therein. If any function of form (1.18) is a characteristic function, then it is infinitely divisible (as its roots have the same form). To show the latter we introduce the approximations

$$\phi_n(u) = \exp\left\{i\left(b - \int_{B_1 \setminus B_{1/n}} y\nu(dy), u\right) - \frac{1}{2}(u, Gu) + \int_{\mathbb{R}^d \setminus B_{1/n}} (e^{i(u,y)} - 1)\nu(dy)\right\}.$$

Each ϕ_n is a characteristic function (of the convolution of a normal distribution and an independent compound Poisson) and $\phi_n(u) \to \phi(u)$ for any u. By the Lévy theorem in order to conclude that ϕ is a characteristic function, one needs to show that ϕ is continuous at zero. This is easy (check it!). □

The function η appearing under the exponent in the representation $\phi(u) = e^{\eta(u)}$ of form (1.18) is called the *characteristic exponent* or *Lévy exponent* or *Lévy symbol* of ϕ (or of its distribution). The vector b in (1.18) is called the *drift vector* and G is called the matrix of *diffusion coefficients*.

Theorem 1.4.2. *Any infinitely divisible probability measure μ is a weak limit of a sequence of compound Poisson distributions.*

Proof. Let ϕ be a characteristic function of μ so that $\phi^{1/n}$ is the characteristic function of its "convolution root" μ_n. Define

$$\phi_n(u) = \exp\{n[\phi^{1/n}(u) - 1]\} = \exp\left\{\int_{\mathbb{R}^d} (e^{i(u,y)} - 1)n\mu_n(dy)\right\}.$$

Each ϕ_n is a ch.f. of a compound Poisson process and

$$\phi_n = \exp\{n(e^{(1/n)\ln\phi(u)} - 1)\} \to \phi(u), \quad n \to \infty.$$

The proof completes by Glivenko's theorem. □

An important class of infinitely divisible distributions constitute the so-called stable laws. A probability law in $\mathbb{R}^d$, its characteristic function ϕ and a random variable X with this law are called *stable (respectively strictly stable)* if for any integer n there exist a positive constant c_n and a real constant γ_n (resp. if additionally $\gamma_n = 0$) such that

$$\phi(y) = [\psi(y/c_n)\exp\{i\gamma_n y\}]^n.$$

In other words, the sum of any number of i.i.d. copies of X is distributed like X up to a shift and scaling. Obviously, it implies that ϕ is infinitely divisible and therefore $\log\phi$ can be presented in the Lévy–Khintchine form with appropriate b, A, ν.

Theorem 1.4.3. *If ϕ is stable, then there exists an $\alpha \in (0,2]$, called the index of stability such that:*

(i) *if $\alpha = 2$, then $\nu = 0$ in the representation* (1.18)*, i.e. the distribution is normal;*

(ii) *if $\alpha \in (0,2)$, then in the representation* (1.18) *the matrix G vanishes and the radial part of the Lévy measure ν has the form $|\xi|^{-(1+\alpha)}$, i.e.*

$$\log\phi_\alpha(y) = i(b,y) + \int_0^\infty \int_{S^{d-1}} \left(e^{i(y,\xi)} - 1 - \frac{i(y,\xi)}{1+\xi^2}\right)\frac{d|\xi|}{|\xi|^{1+\alpha}}\mu(ds), \quad (1.19)$$

where ξ is presented by its magnitude $|\xi|$ and the unit vector $s = \xi/|\xi| \in S^{d-1}$ in the direction ξ, and μ is some (finite) measure in S^{d-1}.

The classical proof can be found e.g. in Feller [111] or Samorodnitski and Taqqu [288].

The integration in $|\xi|$ in (1.19) can be carried out explicitly, as the following result shows.

Theorem 1.4.4. *The stable exponent* (1.19) *can be written in the form*

$$\log \phi_\alpha(y) = i(\tilde{b}, y) - \int_{S^{d-1}} |(y,s)|^\alpha \Big(1 - i \operatorname{sgn}((y,s)) \tan \frac{\pi\alpha}{2}\Big) \tilde{\mu}(ds), \quad \alpha \neq 1, \tag{1.20}$$

$$\log \phi_\alpha(y) = i(\tilde{b}, y) - \int_{S^{d-1}} |(y,s)| \Big(1 + i \frac{2}{\pi} \operatorname{sgn}((y,s)) \log |(y,s)|\Big) \tilde{\mu}(ds), \quad \alpha = 1, \tag{1.21}$$

where

$$\tilde{b} = b + a_\alpha \int_{S^{d-1}} s\mu(ds), \quad \tilde{\mu} = \begin{cases} \sigma_\alpha \cos(\pi\alpha/2)\mu, & \alpha \neq 1 \\ \pi\mu/2, & \alpha = 1 \end{cases} \tag{1.22}$$

with some constants a_α and σ_α specified below. The measure $\tilde{\mu}$ on S^{d-1} is called sometimes the spectral measure of a stable law.

Proof. For $\alpha \in (0,1)$ and a real p

$$\int_0^\infty (e^{irp} - 1) \frac{dr}{r^{1+\alpha}} = -\frac{\Gamma(1-\alpha)}{\alpha} e^{-i\pi\alpha \operatorname{sgn} p/2} |p|^\alpha, \tag{1.23}$$

where $\operatorname{sgn} p$ is of course the sign of p. In fact, one presents the integral on the r.h.s. of (1.23) as the limit as $\epsilon \to 0_+$ of

$$\int_0^\infty (e^{-(\epsilon - ip)r} - 1) \frac{dr}{r^{1+\alpha}}. \tag{1.24}$$

Let, say, $p > 0$. Then

$$\epsilon - ip = (\epsilon^2 + p^2)^{1/2} e^{-i\theta}$$

with $\tan\theta = p/\epsilon$. So by the Cauchy theorem one can rotate the contour of integration in (1.24) through the angle θ. Changing the variable r to $s = e^{-i\theta} r$ in the integral thus obtained yields for (1.24) the expression

$$e^{-i\theta\alpha} \int_0^\infty (e^{-(\epsilon^2+p^2)^{1/2}s} - 1) \frac{ds}{s^{1+\alpha}},$$

which equals (by integration by parts)

$$-e^{-i\theta\alpha} \frac{(\epsilon^2 + p^2)^{1/2}}{\alpha} \int_0^\infty e^{-(\epsilon^2+p^2)^{1/2}s} s^{-\alpha}\, ds = -e^{-i\theta\alpha} \frac{(\epsilon^2 + p^2)^{\alpha/2}}{\alpha} \Gamma(1-\alpha).$$

Passing to the limit $\epsilon \to 0_+$ (and hence $\theta \to \pi/2$) yields (1.23). In case $p < 0$ one would have to rotate the contour of integration in (1.24) in the opposite direction.

Next, for $\alpha \in (1,2)$ and $p > 0$ integration by parts gives

$$\int_0^\infty \frac{e^{irp} - 1 - irp}{r^{1+\alpha}}\, dr = \frac{ip}{\alpha} \int_0^\infty (e^{ipr} - 1) \frac{dr}{r^\alpha},$$

and then by (1.23)

$$\int_0^\infty \frac{e^{irp} - 1 - irp}{r^{1+\alpha}}\, dr = \frac{\Gamma(\alpha - 1)}{\alpha} e^{-i\pi\alpha/2} p^\alpha. \tag{1.25}$$

Note that the real parts of both (1.23) and (1.23) are positive. From (1.23), (1.25) it follows that for $\alpha \in (0,2)$, $\alpha \neq 1$,

$$\int_0^\infty \left(e^{irp} - 1 - \frac{irp}{1+r^2} \right) \frac{dr}{r^{1+\alpha}} = i a_\alpha p - \sigma_\alpha e^{-i\pi\alpha/2} p^\alpha \tag{1.26}$$

with

$$\sigma_\alpha = \alpha^{-1}\Gamma(1-\alpha), \qquad a_\alpha = -\int_0^\infty \frac{dr}{(1+r^2)r}, \quad \alpha \in (0,1), \tag{1.27}$$

$$\sigma_\alpha = -\alpha^{-1}\Gamma(\alpha-1), \quad a_\alpha = \int_0^\infty \frac{r^{2-\alpha}\,dr}{1+r^2}, \qquad \alpha \in (1,2). \tag{1.28}$$

The case of $\alpha = 1$ is a bit more involved. In order to deal with it observe that

$$\begin{aligned}\int_0^\infty \frac{e^{irp} - 1 - ip\sin r}{r^2}\, dr &= -\int_0^\infty \frac{1-\cos rp}{r^2}\, dr + i\int_0^\infty \frac{\sin rp - p\sin r}{r^2}\, dr \\ &= -\frac{1}{2}\pi p - ip\log p.\end{aligned}$$

In fact, the real part of this integral is evaluated using a standard fact that $f(r) = (1-\cos r)/(\pi r^2)$ is a probability density (with the characteristic function $\psi(z)$ that equals to $1 - |z|$ for $|z| \le 1$ and vanishes for $|z| \ge 1$), and the imaginary part can be presented in the form

$$\begin{aligned}\lim_{\epsilon\to 0}\left[\int_\epsilon^\infty \frac{\sin pr}{r^2}\, dr - p\int_\epsilon^\infty \frac{\sin r}{r^2}\, dr\right] - p\lim_{\epsilon\to 0}\int_\epsilon^{p\epsilon} \frac{\sin r}{r^2}\, dr \\ = -p\lim_{\epsilon\to 0}\int_1^p \frac{\sin \epsilon y}{\epsilon y^2}\, dy = -p\int_1^p \frac{dy}{y},\end{aligned}$$

which implies the required formula. Therefore, for $\alpha = 1$

$$\int_0^\infty \left(e^{irp} - 1 - \frac{irp}{1+r^2} \right) \frac{dr}{r^{1+\alpha}} = i a_1 p - \frac{1}{2}\pi p - ip\log p \tag{1.29}$$

with

$$a_1 = \int_0^\infty \frac{\sin r - r}{(1 + r^2)r^2}\, dr. \tag{1.30}$$

Formulae (1.26)–(1.30) yield (1.20) and (1.21). □

Exercise 1.4.1. Check that $\tilde{\mu}$ is continuous in (1.22), i.e. that

$$\lim_{\alpha \to 1} \sigma_\alpha \cos\left(\frac{\pi\alpha}{2}\right) = \frac{\pi}{2}. \tag{1.31}$$

Hint: if $\alpha < 1$, then

$$\lim_{\alpha \to 1} \frac{\Gamma(1-\alpha)}{\alpha} \cos\left(\frac{\pi\alpha}{2}\right) = \lim_{\alpha \to 1} \frac{\Gamma(2-\alpha)}{\alpha(1-\alpha)} \cos\left(\frac{\pi\alpha}{2}\right) = \lim_{\alpha \to 1} \frac{1}{1-\alpha} \cos\left(\frac{\pi\alpha}{2}\right).$$

For instance, if $d = 1$, S^0 consists of two points. Denoting their $\tilde{\mu}$-measures by μ_1, μ_{-1} one obtains for $\alpha \neq 1$ that

$$\log \phi_\alpha(y) = i\tilde{b}y - |y|^\alpha \left[(\mu_1 + \mu_{-1}) - i \operatorname{sgn} y (\mu_1 - \mu_{-1}) \tan \frac{\pi\alpha}{2}\right].$$

This can be written also in the form

$$\log \phi_\alpha(y) = i\tilde{b}y - \sigma|y|^\alpha \exp\left\{i\frac{\pi}{2}\gamma \operatorname{sgn} y\right\} \tag{1.32}$$

with some $\sigma > 0$ and a real γ such that $|\gamma| \le \alpha$, if $\alpha \in (0,1)$, and $|\gamma| \le 2 - \alpha$, if $\alpha \in (1,2)$.

If the spectral measure $\tilde{\mu}$ is symmetric, i.e. $\tilde{\mu}(-\Omega) = \tilde{\mu}(\Omega)$ for any $\Omega \subset S^{d-1}$, then $\tilde{b} = b$ and formulas (1.25), (1.26) both give the following simple expression:

$$\log \phi_\alpha(y) = i(b, y) - \int_{S^{d-1}} |(y, s)|^\alpha \tilde{\mu}(ds). \tag{1.33}$$

In particular, if the measure $\tilde{\mu}$ is the Lebesgue measure ds on the sphere, then

$$\log \phi_\alpha(y) = i(b, y) - \sigma|y|^\alpha, \tag{1.34}$$

with the constant

$$\sigma = \int_{S^{d-1}} |\cos\theta|^\alpha ds = 2\pi^{(d-1)/2} \frac{\Gamma((\alpha+1)/2)}{\Gamma((\alpha+d)/2)} \tag{1.35}$$

(where $\theta \in [0, \pi]$ denotes the angle between a point on the sphere and its north pole, directed along y), called the scale of a stable distribution.

Exercise 1.4.2. Check the second equation in (1.35).

One sees readily that the characteristic function $\phi_\alpha(y)$ with $\log \phi_\alpha(y)$ from (1.20) and (1.21) and (1.33) with vanishing $\tilde{b}$ enjoy the property that $\phi_\alpha^n(y) = \phi_\alpha(n^{1/\alpha} y)$. Therefore all stable distributions with index $\alpha \neq 1$ and symmetric distributions with $\alpha = 1$ can be made strictly stable, if centered appropriately.

1.5 Stable laws as the Holtzmark distributions

Possibly the first appearance of non-Gaussian stable laws in physics was due to Holtzmark [134], who showed that the distribution of the gravitation force (acting on any given object), caused by the infinite collection of stars distributed uniformly in $\mathbb{R}^3$ (see below for the precise meaning of this) is given by the $3/2$-stable symmetric distribution in $\mathbb{R}^3$. This distribution is now called the *Holtzmark distribution* and is widely used in astrophysics and plasma physics. We shall sketch a deduction of this distribution in a more general context than usual, showing in particular that, choosing an appropriate power decay of a potential force, any stable law can be obtained in this way, that is as a distribution of the force caused by the infinite collection of points in $\mathbb{R}^d$, distributed uniformly in sectors.

Suppose the force between a particle placed at a point $x \in \mathbb{R}^d$ and a fixed object at the origin is given by

$$F(x) = \gamma x|x|^{-m-1}, \tag{1.36}$$

where γ is a real constant and m is a positive constant. In the classical example of the gravitational or Coulomb forces (the Holtzmark case) $d = 3, m = 2$ and γ depends on the physical parameters of the particles (mass, charge, etc). Suppose now that the position x of a particle is random and is uniformly distributed in the ball B_R of the radius R in $\mathbb{R}^d$. Then the characteristic function of the force between this particle and the origin is

$$\phi_1(p) = |B_R|^{-1} \int_{B_R} e^{i(p,F(x))} dx = 1 + |B_R|^{-1} \int_{B_R} (e^{i(p,F(x))} - 1) dx,$$

where $|B_R|$ denote the volume of B_R. If there are N independent uniformly distributed particles in B_R, then the characteristic function of the force induced by all these particles is clearly

$$\phi_N(p) = \left[1 + |B_R|^{-1} \int_{B_R} (e^{i(p,F(x))} - 1) dx\right]^N.$$

Assume now that the number of particles N is proportional to the volume with a certain fixed density $\lambda > 0$, that is $N = \lambda|B_R|$. We are interested in the limit of the corresponding distribution as $R \to \infty$ (the constant density equation $N = \lambda|B_R|$ makes precise the idea of 'uniform distribution in $\mathbb{R}^d$' mentioned above). The resulting limiting distribution of stars in $\mathbb{R}^d$ is called a *Poisson point process with intensity* λ and will be studied in more detail in Chapter 3.

Thus we are looking for the limit

$$\phi(p) = \lim_{R\to\infty} \left[1 + |B_R|^{-1} \int_{B_R} (e^{i(p,F(x))} - 1) dx\right]^{\lambda|B_R|}.$$